

11th Conference
on Natural Language Processing (KONVENS)

Empirical Methods
in Natural Language Processing

Proceedings of the Conference
on Natural Language Processing 2012

Edited by
Jeremy Jancsary



Scientific series of the ÖGAI
Volume 5

Impressum:

Schriftenreihe der Österreichischen Gesellschaft
für Artificial Intelligende (ÖGAI), Band 5

Herausgeber der Schriftenreihe: Ernst Buchberger
Herausgeber von Band 5: Jeremy Jancsary

Eigentümer: ÖGAI
Verlag: Eigenverlag ÖGAI

Alle: Postfach 177, 1014 Wien, Österreich

© 2012 ÖGAI
Alle Rechte, auch die des auszugsweisen Nachdrucks, vorbehalten.
Die Rechte bezüglich der individuellen Beiträge verbleiben bei den Autoren.

ISBN 3-85027-005-X Online/CD-ROM-Ausgabe

Table of Contents

Main conference

Preface	9
---------------	---

Invited talks

<i>The statistical approach to natural language processing: Achievements and open problems</i> Hermann Ney	19
---	----

<i>Compositionality in (high-dimensional) space</i> Marco Baroni	20
---	----

Oral presentations

<i>Data-driven knowledge extraction for the food domain</i> Michael Wiegand, Benjamin Roth and Dietrich Klakow	21
---	----

<i>Integrating viewpoints into newspaper opinion mining for a media response analysis</i> Thomas Scholz and Stefan Conrad	30
--	----

<i>A supervised POS tagger for written Arabic social networking corpora</i> Rania Al-Sabbagh and Roxana Girju.....	39
---	----

<i>NLP workflow for on-line definition extraction from English and Slovene text corpora</i> Senja Pollak, Anže Vavpetič, Janez Kranjc, Nada Lavrač and Špela Vintar.....	53
---	----

<i>Projecting semantic roles via Tai mappings</i> Hector-Hugo Franco-Peña and Martin Emms.....	61
---	----

<i>Automatic identification of motion verbs in WordNet and FrameNet</i> Parvin Sadat Feizabadi and Sebastian Padó	70
--	----

<i>WordNet-based lexical simplification of a document</i> S. Rebecca Thomas and Sven Anderson.....	80
---	----

<i>Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs</i> Ines Rehbein, Josef Ruppenhofer, Caroline Sporleder and Manfred Pinkal	89
--	----

<i>Semantic analysis in word vector spaces with ICA and feature selection</i> Tiina Lindh-Knuutila, Jaakko Väyrynen and Timo Honkela.....	98
--	----

<i>Aggregating skip bigrams into key phrase-based vector space model for web person disambiguation</i> Jian Xu, Qin Lu and Zhengzhong Liu.....	108
---	-----

<i>Named entity recognition: Exploring features</i> Maksim Tkachenko and Andrey Simanovsky	118
---	-----

<i>Mention detection: First steps in the development of a Basque coreference resolution system</i> Ander Soraluze, Olatz Arregi, Xabier Arregi, Klara Ceberio and Arantza Díaz de Ilarraz ..	128
---	-----

<i>Phonetically aided syntactic parsing of spoken language</i>	
Zeeshan Ahmed, Peter Cahill and Julie Carson-Berndsen.....	137
<i>Linguistic analyses of the LAST MINUTE corpus</i>	
Dietmar Rösner, Manuela Kunze, Mirko Otto, Jörg Frommer	145
<i>S-restricted monotone alignments, exhaustive enumeration, integer compositions, and string transductions</i>	
Steffen Eger.....	155
<i>Use of linguistic features for improving English-Persian SMT</i>	
Zakieh Shakeri, Neda Noormohammadi, Shahram Khadivi and Noushin Riahi.....	165
Poster presentations	
<i>A semantic similarity measure based on lexico-syntactic patterns</i>	
Alexander Panchenko, Olga Morozova and Hubert Naets	174
<i>A multi-level annotation model for fine-grained opinion detection in German blog comments</i>	
Bianka Trevisan, Melanie Neunerdt and Eva-Maria Jakobs.....	179
<i>Robust processing of noisy web-collected data</i>	
Jelke Bloem, Michaela Regneri and Stefan Thater	189
<i>Navigating sense-aligned lexical-semantic resources: The web interface to UBY</i>	
Iryna Gurevych, Michael Matuschek, Tri-Duc Nghiem, Judith Eckle-Kohler, Silvana Hartmann and Christian Meyer	194
<i>Using information retrieval technology for a corpus analysis platform</i>	
Carsten Schnober	199
<i>Three approaches to finding valence compounds</i>	
Yannick Versley, Anne Brock, Verena Henrich and Erhard Hinrichs.....	208
<i>A computational semantic analysis of gradable adjectives</i>	
Mantas Kasperavicius	213
<i>Extending dependency treebanks with good sentences</i>	
Alexander Volokh and Günter Neumann.....	218
<i>Using subcategorization frames to improve French probabilistic parsing</i>	
Anthony Sigogne and Matthieu Constant	223
<i>Statistical denormalization for Arabic text</i>	
Mohammed Moussa, Mohammed Fakhr and Kareem Darwish.....	228
<i>Automatic identification of language varieties: The case of Portuguese</i>	
Marcos Zampieri and Binyam Gebre	233
<i>Extending the STTS for the annotation of spoken language</i>	
Ines Rehbein and Sören Schalowski	238

<i>Comparing variety corpora with vis-à-vis – A prototype system presentation</i>	
Stefanie Anstein	243
<i>Selecting features for domain-independent named entity recognition</i>	
Maksim Tkachenko and Andrey Simanovsky	248
<i>Ambiguity in German connectives: A corpus study</i>	
Angela Schneider and Manfred Stede	254
<i>Corpus-based acquisition of German event-and object-denoting nouns</i>	
Stefan Goritzte, Sebastian Padó	259

PATHOS 2012: First Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis

Preface.....	264
Invited talks	
<i>Emotions and creative language</i>	
Carlo Strappavara.....	268
Contributions	
<i>Unsupervised sentiment analysis with a simple and fast Bayesian model using Part-of-Speech Feature Selection</i>	
Christian Scheible and Hinrich Schütze.....	269
<i>Sentiment analysis for media reputation research</i>	
Samuel Läubli, Mario Schranz, Urs Christen and Manfred Klenner	274
<i>Opinion analysis: The effect of negation on polarity and intensity</i>	
Lei Zhang and Stéphane Ferrari	282
<i>Domain-specific variation of sentiment expressions: A methodology of analysis in academic writing</i>	
Stefania Degaetano, Ekaterina Lapshinova-Koltunski and Elke Teich	291
<i>Creating annotated resources for polarity classification in Czech</i>	
Kateřina Veselovská, Jan Hajíč, Jr. and Jana Šindlerová.....	296
<i>A phrase-based opinion list for the German language</i>	
Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari and Nikolaos Korfiatis	305
<i>Opinion mining in an informative corpus: Building lexicons</i>	
Patrice Enjalbert, Lei Zhang and Stéphane Ferrari	314
<i>What is the meaning of 5*'s? An investigation of the expression and rating of sentiment</i>	
Daniel Hardt and Julie Wulff	319

LThist 2012: First International Workshop on Language Technology for Historical Text(s)

Preface.....	327
Contributions	
<i>Rule-based normalization of historical text – A diachronic study</i>	
Eva Pettersson, Beáta Megyesi and Joakim Nivre	333
<i>Manual and semi-automatic normalization of historical spelling – Case studies from Early New High German</i>	
Marcel Bollmann, Stefanie Dipper, Julia Krasselt and Florian Petran	342
<i>The Sketch Engine as infrastructure for historical corpora</i>	
Adam Kilgarriff, Milos Husak and Robyn Woodrow	351
<i>Evaluating a post-editing approach for handwriting transcription</i>	
Verónica Romero, Joan Andreu Sánchez, Nicolás Serrano and Enrique Vidal.....	357
<i>bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa ... Towards lexical link-up for a corpus of Old Swedish</i>	
Yvonne Adesam, Malin Ahlberg and Gerlof Bouma.....	365
<i>Semantic change and the distribution in lexical sets</i>	
Karin Cavallin	370
<i>Automatic classification of folk narrative genres</i>	
Dong Nguyen, Dolf Trieschnigg, Theo Meder and Mariët Theune	378
<i>The DTA ‘base format’: A TEI-subset for the compilation of interoperable corpora</i>	
Alexander Geyken, Susanne Haaf and Frank Wiegand	383
<i>Building an old Occitan corpus via cross-language transfer</i>	
Olga Scrivner and Sandra Kübler	392
<i>Digitised historical text: Does it have to be mediOCRe?</i>	
Bea Alex, Claire Grover, Ewan Klein and Richard Tobin.....	401
<i>Comparison of named entity extraction tools for raw OCR text</i>	
Kepa Joseba Rodríguez, Michael Bryant, Tobias Blanke and Magdalena Luszczynska	410
<i>From semi-automatic to automatic affix extraction in Middle English corpora: Building a sustainable database for analyzing derivational morphology over time</i>	
Hagen Peukert.....	415
<i>The reference corpus of Late Middle English scientific prose</i>	
Javier Calle-Martín, Laura Esteban-Segura, Teresa Marqués-Aguado and Antonio Miranda-García.....	424

LexSem 2012: Workshop on Recent Developments and Applications of Lexical-Semantic Resources

Preface.....	433
--------------	-----

Contributions

The construction of a catalog of Brazilian Portuguese verbs

Márcia Cançado, Luisa Godoy and Luana Amaral	438
--	-----

Comparing Czech and Russian valency on the material of VALLEX

Natalia Klyueva.....	446
----------------------	-----

Adding a constructicon to the Swedish resource network of Språkbanken

Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg and Sofia Tingsell	452
--	-----

A distributional memory for German

Sebastian Padó and Jason Utt.....	462
-----------------------------------	-----

Towards high-accuracy bilingual sequence acquisition from parallel corpora

Lionel Nicolas, Egon Stemle and Klara Kranebitter.....	471
--	-----

SFLR 2012: Workshop on Standards for Language Resources

Preface.....	480
--------------	-----

Invited talks

Standards for language technology – Relevance and impact

Gerhard Budin.....	484
--------------------	-----

Standards for the technical infrastructure of language resource repositories: Persistent identifiers

Oliver Schonefeld, Andreas Witt	485
---------------------------------------	-----

Standards for the formal representation of linguistic data: An exchange format for feature structures

Rainer Osswald	486
----------------------	-----

Towards standardized descriptions of linguistic features: ISOcat and procedures for using common data categories

Menzo Windhouwer.....	494
-----------------------	-----

Standardizing metadata descriptions of language resources: The Common Metadata Initiative, CMDI

Thorsten Trippel and Andreas Witt.....	495
--	-----

Standardizing lexical-semantic resources – Fleshing out the abstract standard LMF

Judith Eckle-Kohler	496
---------------------------	-----

<i>A standardized general framework for encoding and exchange of corpus annotations: The Linguistic Annotation Framework, LAF</i>	
Kerstin Eckart	506
<i>Standard for morphosyntactic and syntactic corpus annotation: The Morpho-syntactic and the Syntactic Annotation Framework, MAF and SynAE</i>	
Laurent Romary	516
<i>Annotation specification and annotation guidelines: Annotating spatial senses in SemAE-Space</i>	
Tibor Kiss.....	517
<i>Towards standards for corpus query: Work on a Lingua Franca for corpus query</i>	
Elena Frick, Piotr Bánski and Andreas Witt	518
<i>Getting involved into language resource standardization: The map of standards and ways to contribute to ongoing work</i>	
Gottfried Herzog	519

KONVENTS 2012

The 11th Conference on Natural Language Processing

September 19-21, 2012
Vienna, Austria

Preface

The Conference on Natural Language Processing ("Konferenz zur Verarbeitung Natürlicher Sprache", KONVENTS) aims at offering a broad perspective on current research and developments within the interdisciplinary field of natural language processing. It allows researchers from all disciplines relevant to this field of research to present their work.

KONVENTS is held in a two year rotation, organized by the scientific societies DGfS-CL (German Society for Linguistics, Section Computational Linguistics), GSCL (Society for Language Technology and Computational Linguistics) and ÖGAI (Austrian Society for Artificial Intelligence).

This year's KONVENTS, which is already the 11th instalment, is organized by ÖGAI and hosted at the *Juridicum* building of the University of Vienna during September 19-21, 2012.

Contributions on research, development, applications and evaluation, covering all areas of natural language processing, ranging from basic questions to practical implementations of natural language resources, components and systems were welcome. In keeping with the tradition, we furthermore chose a central theme for this year's KONVENTS,

Empirical methods in natural language processing,

and especially encouraged the submission of contributions proposing new methods for learning from substantial amounts of natural language (including speech) data, be they annotated or un-annotated, as well contributions relating to evaluation of such methods.

We are happy to report that we received a substantial number of submissions relevant to the central theme, making for a strong and focussed program. Overall, KONVENTS received 65 submissions, 32 out of which were selected for either oral or poster presentation at the conference. The reviewing process was very thorough, with papers typically receiving 3 reviews by the program committee (in borderline cases up to 7), as well as one additional meta-review by an area chair. We would like to express our sincere gratitude and appreciation to everyone involved in the reviewing process.

Another decided goal for this year's KONVENTS was to attract an increasing number of international submissions. With presenters flying in from Africa, America, Asia, Near East, and all across Europe, this goal was clearly met. Part of this success is certainly owed to the remarkably committed organizers of the four workshops KONVENTS is hosting this year, and we would like to extend our thanks to them. We strongly encourage all conference attendants to visit the workshops.

Finally, we are very proud to have Hermann Ney and Marco Baroni as our invited speakers. Their talks will provide highly interesting perspectives on the central theme of the conference and hopefully initiate vivid discussions.

Jeremy Jancsary, Harald Trost and Ernst Buchberger
Conference Organizers

Vienna, September 2012

Organizers:

Jeremy Jancsary (ÖGAI/Microsoft Research Cambridge, UK)
Harald Trost (ÖGAI/Medizinische Universität Wien, Austria)
Ernst Buchberger (ÖGAI/Medizinische Universität Wien, Austria)

Program Chair:

Jeremy Jancsary (ÖGAI/Microsoft Research Cambridge, UK)

Area Chairs:

Stefanie Dipper (DGfS-CL/Ruhr-Universität-Bochum, Germany)
Alexander Mehler (GSCL/Goethe-Universität Frankfurt, Germany)
Philipp Koehn (University of Edinburgh, UK)
Chris Biemann (TU Darmstadt, Germany)
Ernesto William de Luca (FH Potsdam, Germany)
Gerard de Melo (ICSI, Berkeley, USA)
Michael Pucher (FTW, Vienna, Austria)
Yves Scherrer (Université de Genève, Switzerland)
Brigitte Krenn (OFAI, Vienna, Austria)

Reviewing Board:

Omri Abend (The Hebrew University of Jerusalem, Israel)
Hoenen Armin (Goethe-Universität Frankfurt)
Korinna Bade (Hochschule Anhalt)
Kalina Bontcheva (University of Sheffield)
Fabienne Braune (Universität Stuttgart)
Christian Buck (University of Edinburgh)
Miriam Butt (Universität Konstanz)
Federica Cena (Università degli Studi di Torino)
Oliver Čulo (ICSI, Berkeley, USA)
Christian Chiarcos (ISI/USC, USA)
Berthold Crysmann (CNRS - Laboratoire de Linguistique Formelle)
Phillip DeLeon (New Mexico State University, USA)
Sebastian Egger (FTW)
Ahmed El Kholy (Columbia University, New York)
Daniel Erro (AHOLAB)
Christian Federmann (DFKI Language Technology Lab)
Peter M. Fischer (IDS Mannheim)
Peter Fröhlich (FTW)
Juri Ganitkevitch (Johns Hopkins University)
Stefan Gindl (Modul Universitaet Wien)
Anne Göhring (Universität Zürich)

Elke Greifeneder (Humboldt-Universität zu Berlin)
Peter Grzybek (Universität Graz)
Helmar Gust (Universität Osnabrück)
Eva Hasler (University of Edinburgh)
Leonhard Henning (Searchmetrics, Inc.)
Johannes Hercher (Hasso Plattner Institut)
Inma Hernaez (AHOLAB)
Teresa Herrmann (Universität Karlsruhe)
Gerhard Heyer (Universität Leipzig)
Hans-Christoph Hobohm (FH Potsdam)
Florian Holz (Universität Leipzig)
Matthias Huck (RWTH Aachen)
Angelina Ivanova (University of Oslo)
Emmerich Kelih (Universität Wien)
Ivo Keller (TU Berlin)
Alexandra Klein (OFAI)
Roman Klinger (Fraunhofer SCAI, Germany)
Valia Kordoni (Universität des Saarlandes, Germany)
Udo Kruschwitz (University of Essex)
Sandra Kübler (Indiana University, USA)
Jonas Kuhn (IMS Universität Stuttgart)
Kai-Uwe Kühnberger (Universität Osnabrück)
Vikash Kumar (FTW)
Jérôme Kunegis (Universität Koblenz)
Florian Laws (University of Stuttgart, Germany)
Henning Lobin (Universität Gießen)
Andy Lücking (Goethe-Universität Frankfurt)
Anke Lüdeling (HU Berlin)
Arne Mauser (RWTH Aachen)
Alexander Mehler (Goethe-Universität Frankfurt)
Thomas Meyer (Idiap Research Institute, Martigny)
Sylvia Moosmüller (ARI, Austrian Academy of Sciences)
Preslav Nakov (Qatar Foundation, Qatar)
Friedrich Neubarth (OFAI)
Jan Niehues (Universität Karlsruhe)
Ivelina Nikolova (Bulgarian Academy of Sciences, Sofia, Bulgaria)
Diarmuid Ó Séaghdha (University of Cambridge, UK)
Gerhard Paß (Fraunhofer IAIS)
Johann Petrak (OFAI)
Stefan Petrik (SPSC, Graz University of Technology)
Danuta Ploch (Distributed Artificial Intelligence Laboratory)
Till Plumbaum (Distributed Artificial Intelligence Laboratory)
Simone Ponzetto (University of Rome "La Sapienza", Italy)
Adam Przeźiókowski (Polish Academy of Sciences, Warszawa, Poland)
Uwe Quasthoff (Universität Leipzig, Germany)

Ines Rehbein (Potsdam University)
Martin Riedl (TU Darmstadt, Germany)
Marcel Ruhl (FH Potsdam)
Josef Ruppenhofer (Universität Hildesheim, Germany)
Tanja Samardzic (Université de Genève)
Felix Sasaki (DFKI)
Dietmar Schabus (FTW)
Silke Scheible (Universität Stuttgart)
Thomas Schmidt (IDS Mannheim)
Stephanie Schreitter (OFAI)
Michael Seadle (Humboldt-Universität zu Berlin)
Rico Sennrich (Universität Zürich)
Serge Sharoff (University of Leeds)
Arne Skjærholt (University of Oslo)
Marcin Skowron (OFAI)
Jan Šnajder (University of Zagreb, Croatia)
Caroline Sporleder (Universität des Saarlandes, Germany)
Armando Stellato (Università degli Studi di Roma "Tor Vergata")
Markus Toman (FTW)
Harald Trost (Medizinische Universität Wien)
Yannick Versley (Univ. Tuebingen)
Tim vor der Brück (Goethe-Universität Frankfurt)
Ulli Waltinger (Siemens)
Andreas Witt (IDS Mannheim)
Jörn Wübker (RWTH Aachen)
Feiyu Xu (DFKI, Germany)
Amir Zeldes (HU Berlin)
Torsten Zesch (TU Darmstadt, Germany)
Heike Zinsmeister (Universität Stuttgart)

Invited Speakers:

Hermann Ney (RWTH Aachen University, Aachen/DIGITEO Chair, LIMSI-CNRS, Paris)
Marco Baroni (Università di Trento)

Sponsors:

The Austrian Research Institute for Artificial Intelligence (OFAI)
<http://www.ofai.at>



Conference Program

Wednesday, September 19, 2012

- 08:30–09:15 Registration (at the conference desk)
- 09:15–09:20 Opening address by the president of the organizing society ÖGAI
Ernst Buchberger
- 09:15–09:20 Welcome address by the program chair
Jeremy Jancsary
- 09:30–10:30 First invited talk: *The statistical approach to natural language processing:
Achievements and open problems*
Hermann Ney
- 10:30–11:15 Coffee break and poster sessions
- A semantic similarity measure based on lexico-syntactic patterns*
Alexander Panchenko, Olga Morozova and Hubert Naets
- A multi-level annotation model for fine-grained opinion detection in German blog
comments*
Bianka Trevisan, Melanie Neunerdt and Eva-Maria Jakobs
- Robust processing of noisy web-collected data*
Jelke Bloem, Michaela Regneri and Stefan Thater
- Navigating sense-aligned lexical-semantic resources: The web interface to UBY*
Iryna Gurevych, Michael Matuschek, Tri-Duc Nghiem, Judith Eckle-Kohler, Silvana
Hartmann and Christian Meyer
- Using information retrieval technology for a corpus analysis platform*
Carsten Schnober
- Three approaches to finding valence compounds*
Yannick Versley, Anne Brock, Verena Henrich and Erhard Hinrichs
- A computational semantic analysis of gradable adjectives*
Mantas Kasperavicius
- Extending dependency treebanks with good sentences*
Alexander Volokh and Günter Neumann
- 11:15–12:45 First session
- Data-driven knowledge extraction for the food domain*
Michael Wiegand, Benjamin Roth and Dietrich Klakow

Integrating viewpoints into newspaper opinion mining for a media response analysis
Thomas Scholz and Stefan Conrad

A supervised POS tagger for written Arabic social networking corpora
Rania Al-Sabbagh and Roxana Girju

12:45–14:15 Lunch break (on your own)

14:15–15:15 Second session

NLP workflow for on-line definition extraction from English and Slovene text corpora

Senja Pollak, Anže Vavpetič, Janez Kranjc, Nada Lavrač and Špela Vintar

Projecting semantic roles via Tai mappings

Hector-Hugo Franco-Penya and Martin Emms

15:15–16:00 Coffee break and poster presentations

(See above for a list of the posters)

16:00–17:30 Third session

Automatic identification of motion verbs in WordNet and FrameNet

Parvin Sadat Feizabadi and Sebastian Padó

WordNet-based lexical simplification of a document

S. Rebecca Thomas and Sven Anderson

Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs

Ines Rehbein, Josef Ruppenhofer, Caroline Sporleder and Manfred Pinkal

Thursday, September 20, 2012

- 09:30–10:30 Second invited talk: *Compositionality in (high-dimensional) space*
Marco Baroni
- 10:30–11:15 Coffee break and poster sessions
- Using subcategorization frames to improve French probabilistic parsing*
Anthony Sigogne and Matthieu Constant
- Statistical denormalization for Arabic text*
Mohammed Moussa, Mohammed Fakhr and Kareem Darwish
- Automatic identification of language varieties: The case of Portuguese*
Marcos Zampieri and Binyam Gebre
- Extending the STTS for the annotation of spoken language*
Ines Rehbein and Sören Schalowski
- Comparing variety corpora with vis-à-vis – A prototype system presentation*
Stefanie Anstein
- Selecting features for domain-independent named entity recognition*
Maksim Tkachenko and Andrey Simanovsky
- Ambiguity in German connectives: A corpus study*
Angela Schneider and Manfred Stede
- Corpus-based acquisition of German event-and object-denoting nouns*
Stefan Gorzitze, Sebastian Padó
- 11:15–12:45 Fourth session
- Semantic analysis in word vector spaces with ICA and feature selection*
Tiina Lindh-Knuutila, Jaakko Väyrynen and Timo Honkela
- Aggregating skip bigrams into key phrase-based vector space model for web person disambiguation*
Jian Xu, Qin Lu and Zhengzhong Liu
- Named entity recognition: Exploring features*
Maksim Tkachenko and Andrey Simanovsky
- 12:45–14:15 Lunch break (on your own)
- 14:15–15:15 Fifth session
- Mention detection: First steps in the development of a Basque coreference resolution system*
Ander Soraluze, Olatz Arregi, Xabier Arregi, Klara Ceberio and Arantza Díaz de Ilarrazá

Phonetically aided syntactic parsing of spoken language

Zeeshan Ahmed, Peter Cahill and Julie Carson-Berndsen

15:15–16:00 Coffee break and poster presentations

(See above for a list of the posters)

16:00–17:30 Sixth session

Linguistic analyses of the LAST MINUTE corpus

Dietmar Rösner, Manuela Kunze, Mirko Otto and Jörg Frommer

S-restricted monotone alignments, exhaustive enumeration, integer compositions, and string transductions

Steffen Eger

Use of linguistic features for improving English-Persian SMT

Zakieh Shakeri, Neda Noormohammadi, Shahram Khadivi and Noushin Riahi

17:30–17:45 Closing remarks by the program chair

Jeremy Jancsary

18:00– General meeting of the GSCL

Friday, September 21, 2012

09:00–18:00 Workshops

PATHOS 2012: First Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis

Stefan Gindl, Robert Remus and Michael Wiegand

LThist 2012: First International Workshop on Language Technology for Historical Text(s)

Brigitte Krenn, Karlheinz Mörtz and Thierry Declerck

LexSem 2012: Workshop on Recent Developments and Applications of Lexical-Semantic Resources

Stefan Engelberg and Claudia Kunze

SFLR 2012: Workshop on Standards for Language Resources

Andreas Witt and Ulrich Heid

19:00– Conference dinner

At “Heuriger 10er Marie”, Ottakringer Straße 222-224, A-1160 Vienna

The Statistical Approach to Natural Language Processing: Achievements and Open Problems

Hermann Ney

RWTH Aachen University, Aachen
DIGITEO Chair, LIMSI-CNRS, Paris

Abstract

When IBM research presented a statistical approach to French-English machine translation in 1988, the linguistic community was shocked because this approach was a hit in the face of the then received machine translation theories. Since then we have seen a dramatic progress in statistical methods for speech recognition, machine translation and other tasks in natural language processing. This talk gives an overview of the underlying statistical methods. In particular, the talk will focus on the remarkable fact that, for all these tasks, the statistical approach makes use of the same four principles:

- Bayes decision rule for minimum error rate,
- probabilistic models, e.g. Hidden Markov models or conditional random fields, for handling strings of observations (like acoustic vectors for speech recognition and written words for language translation),
- training criteria and algorithms for estimating the free model parameters from large amounts of data,
- the generation or search process that generates the recognition or translation result.

Most of these methods had originally been designed for speech recognition. However, it has turned out that, with suitable modifications, the same concepts carry over machine translation and other tasks in natural language processing. This lecture will summarize the achievements and the open problems in this area of statistical modelling.

Compositionality in (high-dimensional) space

Marco Baroni
Università di Trento

Abstract

Formal semanticists have developed sophisticated compositional theories of sentential meaning, paying a lot of attention to those grammatical words (determiners, logical connectives, etc.) that constitute the functional scaffolding of sentences. Corpus-based computational linguists, on the other hand, have developed powerful distributional methods to induce representations of the meaning of content-rich words (nouns, verbs, etc.), typically discarding the functional scaffolding as "stop words". Since we do not communicate by logical formulas, nor, Tarzan-style, by flat concatenation of content words, a satisfactory model of the semantics of natural language should strike a balance between the two approaches. In this talk, I will present some recent proposals that try to get the best of both worlds by adapting the classic view of compositionality as function application developed by formal semanticists to distributional models of meaning. I will present preliminary evidence of the effectiveness of these methods in scaling up to the phrasal and sentential domains, and discuss to what extent the representations of phrases and sentences we get out of compositional distributional semantics are related to what formal semanticists are trying to capture.

Data-driven Knowledge Extraction for the Food Domain

Michael Wiegand, Benjamin Roth, Dietrich Klakow

Spoken Language Systems
Saarland University
D-66123 Saarbrücken

{michael.wiegand|benjamin.roth|dietrich.klakow}@lsv.uni-saarland.de

Abstract

In this paper, we examine methods to automatically extract domain-specific knowledge from the food domain from unlabeled natural language text. We employ different extraction methods ranging from surface patterns to co-occurrence measures applied on different parts of a document. We show that the effectiveness of a particular method depends very much on the relation type considered and that there is no single method that works equally well for every relation type. We also examine a combination of extraction methods and also consider relationships between different relation types. The extraction methods are applied both on a domain-specific corpus and the domain-independent factual knowledge base *Wikipedia*. Moreover, we examine an open-domain lexical ontology for suitability.

1 Introduction

There has been only little research on natural language processing in the food domain even though there is a high commercial potential in automatically extracting knowledge involving food items. For example, such knowledge could be beneficial for virtual customer advice in a supermarket. The advisor might suggest products available in the shop that would potentially complement the items a customer has already in their shopping cart. Additionally, food items required for preparing a specific dish or typically consumed at a social occasion could be recommended. The advisor could also suggest an appropriate substitute

for a product a customer would like to purchase if that product is out of stock.

In this paper, we present methods to automatically extract knowledge from the food domain. We apply different relation extraction methods, such as simple manually designed surface patterns or statistical co-occurrence measures, on both a domain-specific corpus and the open-domain factual knowledge base *Wikipedia*. These large corpora are exclusively used as *unlabeled* data. In addition to the corpora, we also assess an open-domain lexical ontology. Moreover, we combine these methods and harness the relationship between different relation types. Since these methods only require a low level of linguistic processing, they have the advantage that they can provide responses in real time. We show that these individual methods have varying strength depending on which particular food relation type is considered.

Our system has to solve the following task: It is given a *partially instantiated relation*, such as *Ingredient-of(FOOD-ITEM=?, pancake)*. The system has to produce a ranked list of possible values that are valid arguments of the unspecified argument position. In the current example, this would correspond to listing ingredients that are necessary in order to prepare *pancakes*, such as *eggs*, *flour*, *sugar* and *milk*. The entities that are to be retrieved are always food items. Moreover, we only consider binary relations. The relation types we examine (such as *Ingredient-of*) are domain specific.

2 Related Work

Previous work on relation extraction focused on domain-independent semantic relations, such as hyponyms (Hearst, 1992; Snow et al., 2006; Pantel and Pennacchiotti, 2006), meronyms (Girju et al., 2003; Pantel and Pennacchiotti, 2006), synonyms (Chklovski and Pantel, 2004), general purpose analogy relations (Turney et al., 2003) and general relations involving persons or organizations (Ji et al., 2010).

There has also been some work on relation extraction in the food domain. The most prominent research addresses ontology or thesaurus alignment (van Hage et al., 2010), a task in which concepts from different sources are related to each other. In this context hyponymy relations (van Hage et al., 2005) and part-whole relations (van Hage et al., 2006) have been explored. In both (van Hage et al., 2005) and (van Hage et al., 2006) the semantic relations are extracted or learned from various types of data. The work which is most closely related to this paper is (Wiegand et al., 2012a) in which extraction methods are examined for the relations that we also address in this paper. However, this paper extends the preliminary study presented in (Wiegand et al., 2012a) in many ways: Apart from providing a more detailed explanation for the performance of the different extraction methods, we also compare them on different types of text data (i.e. domain-specific and domain-independent data). Moreover, we assess in how far existing general-purpose resources (we examine *GermaNet* (Hamp and Feldweg, 1997)) might help for this task. In addition, we propose ways to improve extraction performance by combining different extraction methods and considering inter-relationships between the different relation types.

Many approaches to recognize relations employ some form of patterns. These patterns can be either manually specified (Hearst, 1992), fully automatically learned (Girju et al., 2003) or semi-automatically learned (Chklovski and Pantel, 2004; Pantel and Pennacchiotti, 2006). The levels of representation that are considered in these patterns also vary. For some tasks, elaborate patterns using syntactic information are applied (Hearst, 1992; Girju et al., 2003;

Chklovski and Pantel, 2004; Pantel and Pennacchiotti, 2006). For others, very simple lexical patterns are employed (Turney and Littman, 2003). In particular for web-based approaches, the latter are much easier to cope with as they allow the patterns to be used as ordinary queries for search engines. In addition to the usage of patterns, some statistical co-occurrence measures have also been successfully used to extract certain relations (Turney and Littman, 2003). While patterns are more generally applicable, the usage of co-occurrence measures is only effective if large amounts of data, for instance the web, are used as a dataset.

3 Data and Resources

For our experiments we use a crawl of *chefkoch.de*¹ as a domain-specific dataset. *chefkoch.de* is the largest web portal for food-related issues in the German language. Note that we only consider the *forum* of this website for our experiments. The website also contains some more structured information, such as a *recipe*-section, but this knowledge could only be extracted by writing a rule-based parser processing the idiosyncratic format of those webpages which would – unlike the approaches examined in this paper – not be generally applicable. We obtained the crawl by using *Heritrix* (Mohr et al., 2004). The plain text from the crawled set of web pages is extracted by using *Boilerpipe* (Kohlschutter et al., 2010). The final domain-specific corpus consists of 418,558 webpages (3GB plain text).

In order to use Wikipedia, we downloaded the current dump of the German version of Wikipedia.² This pre-processed corpus contains 385,366 articles (4.5GB plain text).³ All corpora are lemmatized by using *TreeTagger* (Schmid, 1994). In order to have an efficient data access we index the corpora with *Lucene* (McCandless et al., 2010). As a domain-independent lexical database, we use *GermaNet* (Hamp and Feldweg, 1997) which is the German counterpart of *WordNet* (Miller et al., 1990). We use Version 5.3.

¹www.chefkoch.de

²The dump was downloaded in the fourth quarter of 2011.

³Note that we only processed articles from Wikipedia that contain mentions of food items.

4 The Different Relations Types

In this section, we will briefly describe the four relation types we address in this paper. We just provide English translations of our German examples in order to ensure general accessibility.

- *Suits-to(FOOD-ITEM, EVENT)* describes a relation about food items that are typically consumed at some particular cultural or social event. Examples are *<roast goose, Christmas>* or *<popcorn, cinema visit>*.
- *Served-with(FOOD-ITEM, FOOD-ITEM)* describes food items that are typically consumed together. Examples are *<fish fingers, mashed potatoes>*, *<baguette, ratatouille>* or *<wine, cheese>*.
- *Substituted-by(FOOD-ITEM, FOOD-ITEM)* lists pairs of food items that are almost identical to each other in that they are commonly consumed or served in the same situations. Examples are *<butter, margarine>*, *<anchovies, sardines>* or *<Sauvignon Blanc, Chardonnay>*.
- *Ingredient-of(FOOD-ITEM, DISH)* denotes some ingredient of a particular dish. Examples are *<chickpea, falafel>* or *<rice, paella>*.

5 Challenges of this Task

The extraction of relations from the food domain yields some particular challenges. The most striking problem of this task is that the language to be employed to express a specific relation type can be very diverse. For example, the relation type *Ingredient-of* is often expressed in the context of a cooking instruction. Thus, the language that is used may be very specific to the procedure of preparing a particular dish. For instance, Example 1 expresses the relation instance *Ingredient-of(cooking oil, pancake)*. As such, it would be extremely difficult to employ some textual patterns in order to detect this relation as the relevant entities *cooking oil* and *pancake* are not contained within the same sentence. Even if there were a means to acquire such a long-distance pattern, one may doubt that this pattern would capture other relation instances, such as *Ingredient-of(mince meat, lasagna)*, as many of those involve other procedural patterns.

1. Pour some *cooking oil* into the pan. Add 1/6 of the dough and fry the *pancake* from both sides. [Relation: *Ingredient-of(oil, pancake)*]

The inspection of our data using some seed examples displaying prototypical relation instances of our relation types (e.g. *<hot dog, fries>* for *Served-with*) revealed that – despite the language variability – there are some very simple

and general textual patterns, for example the pattern *FOOD-ITEM and FOOD-ITEM* for *Served-with* as illustrated in Example 2. However, many of those simple patterns are ambiguous and can also be observed with other relation types. For instance, the pattern mentioned above could also imply the relation type *Substituted-by* as in Example 3.

2. We both had a *hot dog* and *fries*. [Relation: *Served-with(fries, hot dog)*]
3. I'm looking for a nice fish-recipe for someone who does not like plain fish but who eats *fish fingers* and *fish cake*. [Relation: *Substituted-by(fish fingers, fish cake)*]

Since three of four of our relation types are relation types between two entities of type FOOD-ITEM⁴, there is a high likelihood that two relation types are confused with each other. This would also suggest that the remaining relation type, which is a relation type between entities of types FOOD-ITEM and EVENT, namely *Suits-to*, is easier to cope with. An obvious solution to detect this relation type is just to consider the co-occurrence of two entities with these particular entity types. For a mention of that relation type, such as Example 4, this would work. However, some mechanism must be provided in order to distinguish those meaningful co-occurrences from coincidental ones, such as Example 5.⁵

4. There will be six of us at *Christmas*. I'd like to prepare a *goose*. [Relation: *Suits-to(goose, Christmas)*]
5. Last *Christmas*, I got a moka-pot. Unfortunately, I don't know how to make a proper *espresso*.

6 Method

In the following, we describe the individual extraction methods that are examined in this paper. The first three methods (Sections 6.1-6.3) are taken from (Wiegand et al., 2012a).

6.1 Surface Patterns (PATT)

As surface patterns, manually compiled patterns are exclusively considered. The patterns comprise a set of few generally-applicable and fairly precise patterns. As a help for building such patterns, Wiegand et al. (2012a) recommend to look at mentions of typical relation instances

⁴Note that the entity type DISH in *Ingredient-of* is a subset of FOOD-ITEM.

⁵That is, Example 5 does not express the relation *Suits-to(espresso, Christmas)*.

in text corpora, e.g. *<butter, margarine>* for *Substituted-by* or *<mince meat, meat balls>* for *Ingredient-of*.

As already stated in Section 5, the formulation of such patterns is difficult due to the variety of contexts in which a relation can be expressed. Wiegand et al. (2012a) confirm this by computing lexical cues automatically with the help of statistical co-occurrence measures, such as the *pointwise mutual information*, which have been run on automatically extracted sentences containing mentions of typical relation instances. The output of that process did not reveal any significant additional patterns.

The final patterns exclusively use lexical items immediately before, between or after the argument slots of the relations. Table 1 illustrates some of these patterns. The level of representation used for those patterns (i.e. word level) is very shallow. However, these patterns are precise and can be easily used as a query for a search engine. Other levels of representation, e.g. syntactic information, would be much more difficult to incorporate. Moreover, Wiegand et al. (2012a) report that they could not find many frequently occurring patterns using these representations to find relation instances that could not be extracted by those simple lexical patterns. Additionally, since the domain-specific data to be used comprise informal user generated natural language, the linguistic processing tools, such as syntactic parsers, i.e. tools that are primarily built with the help of formal newswire text corpora, are severely affected by a domain mismatch.

The extraction method PATT comprises the following steps: Recall from the task description in Section 1 that we always look for a list of values for an unspecified argument in a partially instantiated relation (PIR) and that the unspecified argument is always a food item. Given a PIR, such as *Substituted-by(butter, FOOD-ITEM=?)*, we partially instantiate each of the pertaining patterns (Table 1) with the given argument (e.g. *FOOD-ITEM instead of FOOD-ITEM* becomes *FOOD-ITEM instead of butter*) and then check for any possible food item (e.g. *margarine*) whether there exists a match in our corpus (e.g. *margarine instead of butter*). The output of this extraction process is a ranked list of those food items for which

a match could be found with any of those patterns. We rank by the frequency of matches. Food items are obtained using GermaNet. All those lexical items are collected that are contained within the synsets that are hyponyms of *Nahrung* (English: *food*).

6.2 Statistical Co-occurrence (CO-OC)

The downside of the manual surface patterns is that they are rather sparse as they only fire if the exact lexical sequence is found in our corpus. As a less constrained method, one may therefore also consider statistical co-occurrence. The rationale behind this approach is that if a pair of two specific arguments co-occurs significantly often (at a certain distance), such as *roast goose* and *Christmas*, then there is a likely relationship between these two linguistic entities.

By applying a co-occurrence measure one may be able to separate meaningful from coincidental co-occurrences as exemplified in Examples 4 and 5 in Section 5. As a co-occurrence measure, we consider the *normalized Google distance* (*NGD*) (Cilibrasi and Vitanyi, 2007) which is a popular measure for such tasks. The extraction procedure of CO-OC is similar to PATT with the difference that one does not rank food items by the frequency of matches in a set of patterns (all containing the given entity) but the correlation score with the given entity. For instance, given the PIR *Suits-to(FOOD-ITEM=?, Christmas)*, one computes the scores for each food item from our (food) vocabulary and *Christmas* and sorts all these food items according to the correlation scores.

It is believed that this approach is beneficial for relations where the formulation of surface patterns is difficult – this is typically the case when entities involved in such a relation are realized within a larger distance to each other. Thus, CO-OC would tackle one challenge that was presented in Section 5.

6.3 Relation between Title and Body of a Webpage (TITLE)

Rather than computing statistical co-occurrence at a certain distance, one may also consider the co-occurrence of entities between title and body of a webpage. Wiegand et al. (2012a) argue that

Relation Type	#Patterns	Examples
Suits-to	6	FOOD-ITEM at EVENT; FOOD-ITEM on the occasion of EVENT; FOOD-ITEM for EVENT
Served-with	8	FOOD-ITEM and FOOD-ITEM; FOOD-ITEM served with FOOD-ITEM; FOOD-ITEM for FOOD-ITEM
Substituted-by	8	FOOD-ITEM or FOOD-ITEM; FOOD-ITEM (FOOD-ITEM); FOOD-ITEM instead of FOOD-ITEM
Ingredient-of	8	DISH made of FOOD-ITEM; DISH containing FOOD-ITEM

Table 1: Illustration of the manually designed surface patterns.

entities mentioned in the title represent a predominant topic and that a co-occurrence with an entity appearing in the body of a webpage may imply that the entity has a special relevance to that topic and denote some relation. The co-occurrence of two entities in the body is more likely to be coincidental. None of those entities needs to be a predominant topic.

The extraction procedure of this method selects those documents that contain the given argument of a PIR (e.g. *lasagna* in *Ingredient-of(FOOD-ITEM=?, lasagna)*) in the title and ranks food items that co-occur in the document body of those documents according to their frequency.

6.4 Wikipedia Links (LINK)

Since we also evaluate Wikipedia as a corpus for our relation extraction task, we also want to take into account a feature that is specific to this type of resource, namely *Wikipedia links*. According to the guidelines of Wikipedia⁶, links are typically used to connect some article X to another article Y that a reader of article X might be also interested in. Similar to TITLE, we want to examine whether these links have any specific semantics for our domain task.

Using Wikipedia links we extract relations in the following way: The given argument of a PIR is the source article and we rank food items whose articles are linked to from this source article according to their frequency.⁷

6.5 GermaNet - Sibling Synsets in the Hyperonym Graph (GERM)

Finally, we also examine a general-purpose ontology for German, namely GermaNet. This resource organizes different general relations (e.g.

⁶en.wikipedia.org/wiki/Wikipedia:Link

⁷We do not apply any correlation measure for LINK because of the same reasons as we do not apply them for TITLE.

hyperonymy, hyponomy or meronymy) between different synsets being groups of words with a similar meaning. The assignment of a given food item (or event) to a particular synset is simple as these expressions are usually unambiguous in our domain. Since GermaNet is an open-domain ontology it does not specialize for the relations that we consider in this work. Of our relation types, only *Substituted-by* can be modeled with the help of GermaNet.⁸ We found that the *sibling* relationship between different synsets in the hyperonym graph encodes a very similar concept. For instance, *apple*, *pear*, *quince* and *guava* are siblings (their immediate hyperonym is *pome*) and, therefore, they are likely to be substituted by each other. Of course, the degree of similarity also depends on the location of those siblings within that graph. The more specific a synset is (i.e. the deeper it is within the graph), the more similar are its *siblings* to it. We found that the type of similarity that we want to model can only be reliably preserved if the target synset is actually a leaf node. Otherwise, we would also obtain *meat* and *pastries* as an entry for *Substituted-by*. They, too, are siblings (*solid food* is their immediate hyperonym) but these entries are not leaves in the hyperonym graph.

Unlike the other extraction methods there is no straightforward way for this method to provide a ranking of the food items that are extracted. That is why we evaluate them in random order.

7 Experiments

We already stated in Section 1 that the unspecified argument value of a partially instantiated relation (PIR) is always of type FOOD-ITEM. This

⁸Conceptually speaking, a second relationship, namely *Ingredient-of*, could be recognized with the help of the *meronymy* (part-of) relation of GermaNet. Unfortunately, there exist virtually no entries for food items with regard to that relation.

Partially Instantiated Relations (PIRs)	#PIRs
Suits-to(FOOD-ITEM=?, EVENT)	40
Served-with(FOOD-ITEM, FOOD-ITEM=?)	58
Substituted-by(FOOD-ITEM, FOOD-ITEM=?)	67
Ingredient-of(FOOD-ITEM=?, DISH)	49

Table 2: Statistics of partially instantiated relations in gold standard.

is because these PIRs simulate a typical situation for a virtual customer advisor, e.g. such an advisor is more likely to be asked what food items are suitable for a given event, i.e. *Suits-to(FOOD-ITEM=?, EVENT)*, rather than the opposite PIR, i.e. *Suits-to(FOOD-ITEM, EVENT=?)*. The PIRs we use are presented in Table 2.⁹

We use the gold standard from (Wiegand et al., 2012b) for evaluation.¹⁰ For each relation, a certain number of PIRs has been manually annotated (see also Table 2).

Since our automatically generated output are ranked lists of food items, we use *precision at 10 (P@10)* and *mean reciprocal rank (MRR)* as evaluation measures. The two metrics are to some extent complementary. While P@10 evaluates the matches with the gold standard on the 10 most highly ranked items not taking into account on what positions the correct items appear, MRR just focuses on the highest ranked correct item but it also considers the corresponding ranking position.

7.1 Individual Evaluation of the Different Extraction Methods

Table 3 compares the different individual methods on all of our four relation types. (Note that for CO-OC, we consider the best window size for each respective relation type.) For each relation type, the best extraction is achieved with the help of our domain-specific corpus ([chefkoch.de](#)). This proves that the choice of the corpus is at least as important as the choice of the method.

Table 3 also shows that the performance of a particular method varies greatly with respect to

⁹Since the two relation types *Served-with* and *Substituted-by* are reflexive, the argument positions of the PIRs do not matter.

¹⁰Following Wiegand et al. (2012a), we carried out our experiments on an earlier version of that gold standard. Therefore, the statistics regarding PIRs differ between this work and (Wiegand et al., 2012b).

the relation type on which it has been applied. For *Suits-to*, the methods producing some reasonable output are CO-OC and TITLE. For *Served-with*, PATT and CO-OC are effective. For *Substituted-by*, the clear winner is PATT. Not even the lexical resource GermaNet (GERM) can be harnessed in order to have some comparable output. For *Ingredient-of*, TITLE performs best. For that relation type, LINK also provides some reasonable performance. In terms of coverage (P@10 is the more indicative measure for that), we obtain much better results for *Ingredient-of* than for the other relation types. This can be ascribed to the fact that this is obviously the most discussed relation type. Since LINK and TITLE-Wikipedia are very similar in their nature, we manually inspected the output of those methods for some queries in order to find out why LINK performs so much better. We found that LINK is usually a proper subset of what is retrieved by TITLE-Wikipedia. This subset is much more relevant for *Ingredient-of* than the larger list from TITLE-Wikipedia. The fact that *Ingredient-of* is the only relation type which can be properly extracted with the help of Wikipedia does not come as a surprise as the knowledge encoded in that relation type is mostly factual while the other relation types are influenced by social conventions (mostly *Suits-to*) and common taste (*Served-with* and *Substituted-by*). The latter two issues are less present in Wikipedia.

7.2 Interpreting the Results

The results of Table 3 prove that we can partly solve the challenges presented in Section 5. *Suits-to* and *Ingredient-of* can be successfully extracted using methods that bypass the modeling of the difficult surface realizations. TITLE is very effective for *Ingredient-of*, i.e. it produces a fairly unambiguous output. Thus, for this relation type, we have found a method that does not confuse this relation type with the other relation types exclusively involving entities of type FOOD-ITEM (i.e. *Served-with* and *Substituted-by*).

Table 4 takes a closer look at CO-OC in that it compares different window sizes. (Note that we only consider MRR as this measure is more sensitive to changes in ranking quality than P@10.) This comparison may shed some light into why

Method	Resource	Suits-to		Served-with		Substituted-by		Ingredient-of	
		P@10	MRR	P@10	MRR	P@10	MRR	P@10	MRR
GERM	GermaNet	NA	NA	NA	NA	0.191	0.322	NA	NA
PATT	Wikipedia	0.000	0.000	0.048	0.131	0.024	0.177	0.000	0.000
CO-OC	Wikipedia	0.138	0.417	0.086	0.152	0.076	0.315	0.114	0.215
TITLE	Wikipedia	0.095	0.186	0.076	0.173	0.051	0.160	0.267	0.186
LINK	Wikipedia	0.000	0.000	0.083	0.155	0.058	0.214	0.400	0.646
PATT	chefkoch.de	0.023	0.133	0.343	0.617	0.303	0.764	0.076	0.331
CO-OC	chefkoch.de	0.340	0.656	0.310	0.584	0.172	0.553	0.335	0.581
TITLE	chefkoch.de	0.300	0.645	0.171	0.233	0.049	0.184	0.776	0.733

Table 3: Comparison of the different individual methods (for CO-OC the best window size is considered).

Window	Suits-to	Served-with	Substituted-by	Ingredient-of
2	0.371	0.584	0.553	0.372
5	0.511	0.545	0.537	0.496
10	0.579	0.544	0.527	0.536
20	0.644	0.532	0.534	0.581
50	0.656	0.469	0.512	0.558
sentence	0.525	0.550	0.515	0.544
document	0.618	0.431	0.500	0.377

Table 4: MRR of *CO-OC* using different window sizes.

a particular method only works for certain relations. Small window sizes are very effective for *Served-with* and *Substituted-by*. This means that the entities involved in those relations tend to appear close to each other. This is a pre-requisite that our short-distance patterns (PATT) fire. For the other relation types, in particular *Suits-to*, the entities involved can be fairly far apart from each other (i.e. 50 words in between). For such relation types short-distance patterns are not effective.

Table 4 also shows that more natural boundaries for the entities involved in a relation, i.e. the sentence and the entire document, are less effective than choosing a fixed window size.

7.3 Combination of Extraction Methods

Table 5 compares the performance of the best individual method for each relation type with some combination. The combination always uses the best performing individual method (for each respective relation type) and the method which in combination with the best gives the largest improvement (this is usually the second best method). We experimented with standard merging methods of rankings, such as linear interpolation or multiplication of the inverted ranks. However, they did not result in a notable improvement. Presumably, this is due to the fact that the output of several methods – this is usually the method

that is combined with the best individual method – cannot be suitably represented as a ranking as the entries are more or less equipollent. This is most evident for GERM (as discussed in Section 6.5) but it also applies for LINK. For the latter, we may create a ranking based on frequency but since most links are only observed once or twice, this criterion is hardly discriminant. We therefore came up with another combination procedure that reflects this property. We mainly preserve the ranking produced by the best individual method but boost entries that also occur in the output of the other method considered since these entries should be regarded most reliable. We empirically increase the rank of those entries by n ranking positions. In order to avoid overfitting we just consider three configurations for n : 5, 10 and 20. Table 5 shows that, indeed, some improvement can be achieved by this combination scheme.

7.4 Relationship between Relation Types

Finally, we also examine whether one can improve performance of one relation type by considering some relationship towards another relation type. Recall that *Served-with*, *Substituted-by* and *Ingredient-of* are all relation types between two food items. Therefore, there is a chance that those three relation types get confused. We found

	Suits-to		Served-with		Substituted-by		Ingredient-of	
	P@10	MRR	P@10	MRR	P@10	MRR	P@10	MRR
best individual	0.340	0.656	0.343	0.617	0.303	0.764	0.776	0.733
combination methods ranking increase	0.365[†]	0.722*	0.378[‡]	0.648*	0.310	0.794	0.773	0.835[‡]
	CO-OC _{ch} +TITLE _{ch} 10 ranks		PATT _{ch} +CO-OC _{ch} 20 ranks		PATT _{ch} +GERM 5 ranks		TITLE _{ch} +LINK 5 ranks	

Table 5: Comparison of the best individual method and the best combination of (two) methods. _{ch} indicates that this method has been applied on *chefkoch.de*; significantly better than *best individual* *: at $p < 0.1$; [†]: at $p < 0.05$; [‡]: at $p < 0.01$ (paired t-test).

Suits-to(?, picnic)	Served-with(broccoli, ?)	Substituted-by(beef roulades, ?)	Ingredient-of(?, falafel)
sandwiches*	broad noodles	goulash*	chickpea*
fingerfood	potatoes*	roast*	cooking oil*
noodle salad*	salt potatoes*	roast beef*	garlic*
meat balls*	croquettes	braised meat*	water
potato salad*	sweet corn	marinated beef*	coriander*
melons*	spaetzle	rolled pork*	onions*
fruit salad*	noodle casserole	roast pork*	parsley*
small sausages	fillet of pork*	cutlet	flour*
sparkling wine	mushrooms*	ragout	cumin*
baguette*	rice*	rabbit	salt*

Table 7: The 10 most highly ranked food items for some automatically extracted relations; *: denotes match with the gold standard.

Method	P@10	MRR
Served-with _{ind}	0.343	0.617
Served-with _{comb}	0.378	0.648
Served-with _{ind} + \neg Substituted-by _{ind}	0.393	0.698
Served-with _{comb} + \neg Substituted-by _{comb}	0.431[†]	0.754*

Table 6: Filtering *Served-with* with the output of *Substituted-by*; *ind*: best individual method from Table 5; *comb*: combination method from Table 5; significantly better than *Served-with_{ind}* + \neg Substituted-by_{ind} *: at $p < 0.05$; [†]: at $p < 0.01$ (paired t-test).

that *Served-with* is most affected by this as it gets mostly confused with *Substituted-by*. This comes as no surprise as Table 4 showed that these relation types are very similar with respect to the distance in which their participating entities appear to each other. We try to improve the extraction of *Served-with* by deleting those entries that have also been retrieved for *Substituted-by* (we denote this by \neg Substituted-by). Table 6 shows that this filtering method largely increases the performance of *Served-with*.

Table 7 illustrates some automatically generated output using the best configuration for each relation type. Even though not all retrieved entries match with our gold standard, most of them are (at least) plausible candidates. Note that for our gold standard we aimed for high precision rather

than completeness.

8 Conclusion

In this paper, we examined methods for automatically extract knowledge for the food domain from unlabeled text. We have shown that different relation types require different extraction methods. We compared different resources and found that a domain-specific corpus consisting of web forum entries provides better coverage of the relations we are interested in than the open-domain data we examined. Further improvement can be achieved by combining different methods that may also rely on different resources and using interrelationships between different relation types. Since our methods only require a low level of linguistic processing, they may serve for applications that have to provide responses in real time.

More information about this work including a demo can be found at: www.lsv.uni-saarland.de/personal/Pages/michael/relFood.html

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (Software-Cluster) under grant no. “01IC10S01”.

References

- Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–40, Barcelona, Spain.
- Rudi Cilibarsi and Paul Vitanyi. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370 – 383.
- Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 80–87, Edmonton, Canada.
- Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.
- Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 539–545, Nantes, France.
- Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC KBP 2010 Knowledge Base Population Track. In *Proceedings of the Text Analytics Conference (TAC)*, Gaithersburg, MD, USA.
- Christian Kohlschutter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 441–450, New York City, NY, USA.
- Michael McCandless, Erik Hatcher, and Otis Gospodnetić. 2010. *Lucene in Action*. Manning Publications, 2nd edition.
- George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.
- Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An Introduction to Heritrix, an open source archival quality web crawler. In *Proceedings of the International Web Archiving Workshop (IWAQ)*.
- Patrick Pantel and Marco Pennacchiotti. 2006. Expresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 113–120, Sydney, Australia.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom.
- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 801–808, Sydney, Australia.
- Peter Turney and Michael Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In *Proceedings of ACM Transactions on Information Systems (TOIS)*, pages 315–346.
- Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 482–489, Borovets, Bulgaria.
- Willem Robert van Hage, Sophia Katreko, and Guus Schreiber. 2005. A Method to Combine Linguistic Ontology-Mapping Techniques. In *Proceedings of International Semantic Web Conference (ISWC)*, pages 732 – 744, Galway, Ireland. Springer.
- Willem Robert van Hage, Hap Kolb, and Guus Schreiber. 2006. A Method for Learning Part-Whole Relations. In *Proceedings of International Semantic Web Conference (ISWC)*, pages 723 – 735, Athens, GA, USA. Springer.
- Willem Robert van Hage, Margherita Sini, Lori Finch, Hap Kolb, and Guus Schreiber. 2010. The OAEI food task: an analysis of a thesaurus alignment task. *Applied Ontology*, 5(1):1 – 28.
- Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012a. Web-based Relation Extraction for the Food Domain. In *Proceeding of the International Conference on Applications of Natural Language Processing to Information Systems (NLDB)*, pages 222–227, Groningen, the Netherlands. Springer.
- Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, and Dietrich Klakow. 2012b. A Gold Standard for Relation Extraction in the Food Domain. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 507–514, Istanbul, Turkey.

Integrating Viewpoints into Newspaper Opinion Mining for a Media Response Analysis

Thomas Scholz

Heinrich-Heine University
Institute of Computer Science
D-40225 Düsseldorf, Germany
scholz@cs.uni-duesseldorf.de

Stefan Conrad

Heinrich-Heine University
Institute of Computer Science
D-40225 Düsseldorf, Germany
conrad@cs.uni-duesseldorf.de

Abstract

Opinion Mining in newspaper articles is very interesting for media monitoring, because the results can represent the foundation for many analyses such as a Media Response Analysis. But for these kinds of analyses we need to know the viewpoint for an extracted sentiment. In this paper we apply various commonly used approaches for sentiment analysis and expand research by analysing the specific point of view and viewpoint features. The evaluation on real world datasets of German Media Response Analyses in contrast to a state-of-the-art method illustrates that our approach provide the assignment of a point of view for a given statement. Furthermore, we show how the features can increase the performance of existing solutions and how the viewpoints influence the polarity of sentiment.

1 Introduction

With the growth of online portals of newspapers and news services, Opinion Mining in news has become a central issue for media monitoring. Companies, organisations such as political parties, associations or maybe even distinguished public figures would like to know the polarity of sentiment (we also call it the tonality) in texts which concern them. How does the media talk about the new product of company XY? Is the tonality in the news changing after the speech of the chairman/chairwoman of the party? These questions are answered through a Media Response Analysis (MRA) (Michaelson and Griffin, 2005; Watson and Noble, 2007).

1.1 Media Response Analysis

In order to create a report about the media attention, an initiator of a MRA has first to define key words of interest like the name of companies (subsidiary companies, competitors), products or important persons (chairmen/chairwomen, press agents). Thereby, all texts containing these words are automatically collected by a crawler system. In a MRA, several media analysts read all texts, extract relevant statements from the texts and determine a tonality and the belonging viewpoint for every statement. The following statement is positive for US president Barrack Obama and the Democratic Party, e.g.:

- President Obama made the tough, politically unpopular decision to rescue an industry in crisis, a decision that saved more than 1.4 million American jobs. (Code: positive, Democrats)

To put this into practice today, a big human effort is needed. At the same time, the tasks are getting more and more difficult. The number of potentially relevant articles is increasing and the clients want to obtain their results in real time, because they have to be able to react quickly to dramatic changes in tonality. As a consequence, more machine-aided methods are needed for the field of Opinion Mining in newspaper articles.

While much research has been carried out on the detection on the sentiment (Pang and Lee, 2008), we want to concentrate on the detection of the viewpoint of a statement.

1.2 Problem Definitions

In a MRA, the statements have a tonality and an assignment of the viewpoint (this could be the client's organisation or a competitor).

Definition I: Viewpoint Tonality. *Given a statement s which consists of the words w_i with $i \in \{1, \dots, s_n\}$. The t_1 which determines a tonality y and a point of view assignment g for the statement s .*

$$t_1 : s = (w_1, \dots, w_{s_n}) \mapsto (y, g); \quad (1)$$

$$y \in \{pos, neg\}, g \in \{g_1, \dots, g_m\}$$

The m different views are known before the analysis. The following statement has a positive sentiment for the German chancellor Merkel and her political party CDU, e.g.:

- That itself illustrates how important it was that chancellor Merkel has implemented reforms in Europe. (Code: positive, CDU)

This is the way state-of-the-art companies in media monitoring code their MRA results. A harder task (even for humans and therefore not the common procedure) is the determination of the tonality from a certain point of view.

Definition II: View Modified Tonality. *Given a statement s and a point of view g . The function t_2 which determines the tonality y_g from this point of view.*

$$t_2 : (s, g) \mapsto y_g \in \{pos, neg\} \quad (2)$$

In this example, the tonality y_g is modified by a given point of view g . For example, a statement that is considered positive from a certain viewpoint A, can be negative for another viewpoint B.

- The government quarrels, the SPD acts: The time has come to establish legal rules for a quota of women in commercial enterprises. (Code A: positive, SPD; Code B: negative, CDU)

The SPD is the biggest opposing party in Germany and political competitor of the CDU. While the point of view should be extracted from the statement itself for the **Viewpoint Tonality**, this task needs more external or world knowledge (the SPD is a competitor of the CDU, e.g.).

In this paper, we examine these problems. We present a new ontology-based approach for the determination of viewpoints. In addition, we explain viewpoint features which improve current methods in sentiment analysis for statements which are modified through a viewpoint. We evaluate our approach against a state-of-the-art method on two German corpora: A corpus of a real MRA about the finance industry and the public available pressrelations dataset of the election campaigns which consists of press releases of the CDU and SPD (Scholz et al., 2012). Furthermore, we want to analyse the influence of the viewpoint to the tonality.

This paper contains the following: In section 2 we analyse the related work, before we give a short overview on the basic polarity determination in section 3. In section 4 we present our viewpoint assignment algorithm and viewpoint features. After that we evaluate our methods on real world datasets of a MRA in comparison with DASA (Qiu et al., 2010), before we give a conclusion and a brief outlook on our future work in the last section.

2 Related Work

Recent research in sentiment analysis is focused on Opinion Mining in customer reviews (Pang and Lee, 2008).

One major aspect of Opinion Mining in product reviews is the collection of sentiment bearing words (Harb et al., 2008; Kaji and Kitsuregawa, 2007) or the construction of complex patterns which represent not only the sentiment, but also extract the relationship between the sentiment and the features of the product (Kobayashi et al., 2004). The point of view aspect plays less important role, because a customer expresses only one view (his/her own view) and different viewpoints mostly occur by comparisons of different products (Liu, 2010). As shown in the examples, the viewpoints are almost essential in the newsarea domain and some statements do not have any sentiment without a viewpoint.

Many techniques rely on a sentiment dictionary such as SentiWordNet (Baccianella et al., 2010). In this work, we apply the SentiWS dictionary (Remus et al., 2010). The sources of SentiWS are an English lexicon (General Inquirer),

which is translated into German, a large collection of product reviews, and a special German dictionary. They compute and refine a sentiment score for their words based on the three sources by applying the Pointwise Mutual Information method (Church and Hanks, 1989) for the score. Hereby, they obtain 1686 positive and 1818 negative German words in lemma.

In social media, approaches (Harb et al., 2008; Pak and Paroubek, 2010) analyse tweets from Twitter or blog-entries to extract negative and positive opinions. Pak and Paroubek (2010) are looking for emoticons (smilies) to determine the tonality of the tweets. Likewise, they collect some statistics about frequent POS-tags: Personal pronouns appear more often in a subjective tweet and verbs in past tense are a small hint for a negative tonality. This is unfortunately inappropriate in our case, because news articles do not contain emoticons.

However, far too little attention has been paid to Opinion Mining in newspaper articles and especially the integration of viewpoints. Most of the approaches on this topic only work with single words/phrases (Wilson et al., 2009) or reported speech objects (Balahur et al., 2009; Balahur et al., 2010; Park et al., 2011), because quotations are often more subjective. The opinion holder (the speaker) can be deduced in the cases and sometimes even the object of the opinion (e.g. another person or an organisation which appears in the reported speech). But this technique can only capture opinions which are part of a quotation in the news. We have examined 4000 statements of our evaluation dataset, of which less than 22% of the statements contain quotation marks and only less than 5% have a proportion of quoted text bigger than 50% of the whole statement. As a result, this technique cannot analyse 78% to 95% of the statements. Another approach (Devitt and Ahmad, 2007) works with news articles about a company takeover. It computes graph-based features which require a sentiment dictionary as well as a common dictionary to create the graph. The nodes are concepts of words and the edges represent relations between the words, but this approach does not handle different viewpoints.

Park et al. (2011) extract groups for a certain topic. The groups have contrasting views about

this topic. To extract these groups, they identify the speaker of a reported speech object who agree or disagree with other speakers or organisations of the same group and the opposing group, respectively. Unfortunately, this approach does not fit in with the requirements of a MRA, because the determination of tonality and not the different opinion groups are interesting for such an analysis. The different groups are commonly known from the beginning, anyway. Thomas et al. (Thomas et al., 2006) headed in the same direction by using a graph-based approach with same speaker links and different speaker agreement links in congressional debates.

Qiu et al. (2010) propose an approach called DASA for an online advertising strategy. They analyse web forums for an ad selection based on the consumer attitudes. Although their intention is something different, the basic problem is the same: the extraction of a viewpoint. They use syntactic parsing and a rule base approach to extract topic words which are associated with negative sentiment to propose products from rivals. We also expand this approach to non negative sentiments to create topic words and use this approach for our comparison.

3 Determination of Sentiment Polarity

First, we determine four basic tonality features (**Basic Tonality Features** α) based on the four word categories adverbs, adjectives, nouns, and verbs, which are the most important word classes for tonality (Bollegrala et al., 2011; Remus et al., 2010).

For the weighting of the polarity (our tonality score TS), we use existing methods such as chi-square (Kaji and Kitsuregawa, 2007), the Pointwise Mutual Information method (Church and Hanks, 1989; Kaji and Kitsuregawa, 2007; Turney, 2002) and the German sentiment lexicon SentiWS (Remus et al., 2010). The chi-square value (Kaji and Kitsuregawa, 2007) is given by the statistical measure in which the null hypothesis is assumed which says that one word appears in positive statements with the same probability as in negative statements.

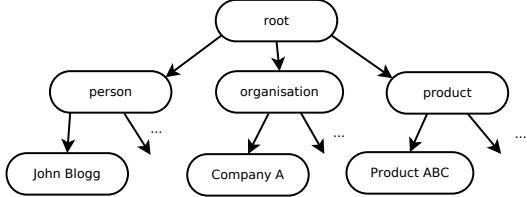


Figure 1: A sample ontology.

$$TS_{\chi^2}(w) = \begin{cases} \chi^2(w) & \text{if } P(w|neg) < P(w|pos) \\ -\chi^2(w) & \text{otherwise} \end{cases} \quad (3)$$

The Pointwise Mutual Information (PMI) based tonality (Kaji and Kitsuregawa, 2007) uses the strength of the association between the word w and positive and negative statements, respectively.

$$TS_{PMI}(w) = \log_2 \frac{P(w|pos)}{P(w|neg)} \quad (4)$$

By using these methods, we construct dictionaries which contain words of the four categories weighted by the sentiment score TS .

We use the SentiWS (Remus et al., 2010) as another method. It contains a tonality value for 3504 words in lemma.

We compute four tonality features for one statement (**Basic Tonality Features α**). Every feature is the average of the tonality score in one category: The first feature is the average of the scores of all the statement's adjectives, the second of all nouns, and so on.

4 Viewpoint of Statements

For the viewpoint of a MRA it is very important to know which entities (persons, organisations, products) play a role in a given statement. To recognize viewpoints based on entities we propose an ontology based approach.

4.1 Ontology based Approach

Ontologies are very helpful in structuring knowledge. For our ontology based approach of viewpoint determination we organise our different entities in a hierarchical structure (shown in Figure 1). The entities are persons, organisations, and products. Persons can have an important role in an

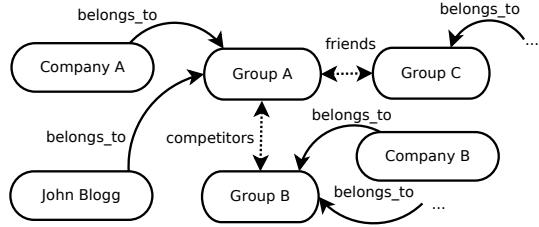


Figure 2: Sample ontology relationships.

organisation, e.g. a special function such as press agent or chairman. Organisations can be companies, political parties and so on. Products can be something the client's companies or the competitors are producing and/or selling.

All these entities have a group attribute. In other words, they belong to one group. Also, each group stands in relationship to every other group. These relationships can be neutral, friendly, or a competitive. Hence, every entity has one of these relationships to each of the other entities. Figure 2 shows an example.

It is not very time consuming to extract this information because media analysts keep this information in so-called code books. These pieces of information are used during the human analysis process (e.g. as tooltips) because it is very hard for humans to remember all the names and relationships between the entities.

4.2 Viewpoint Features

The approach considers how often the article mentions friendly entities and how often the competitors are mentioned. As a result, two features represent the persons and two features the organisations/products (**Viewpoint Features β**).

$$f_{\beta_1}(s) = \frac{pf(s, g)}{p(s)} \quad f_{\beta_2}(s) = \frac{of(s, g)}{o(s)} \quad (5)$$

In equation 5, $p(s)$ and $o(s)$ are the number of persons and organisations, respectively, in the statement s . $pf(s, g)$ and $of(s, g)$ are the friendly persons and friendly organisations/products, respectively, for group g . Friendly persons could be members of the initiator of the MRA (e.g. the managing director of the analysis customer) or a member of a cooperation organisation.

$$f_{\beta_3}(s) = \frac{pc(s, g)}{p(s)} \quad f_{\beta_4}(s) = \frac{oc(s, g)}{o(s)} \quad (6)$$

$pc(s, g)$ and $oc(s, g)$ are persons and organisations/products, respectively, of group g 's competitors. Friends and competitors are deduced by the relationships in the ontology.

Furthermore, the influenced tonality is stored by **Viewpoint Tonality Features** γ .

$$f_{\gamma_1}(s) = \sum_{w \in F_w} TS(w) \quad (7)$$

$$f_{\gamma_2}(s) = \sum_{w \in R_w} TS(w) \quad (8)$$

F_w and R_w are the sets of words which belongs to friend and competitor entities, respectively, and are determined in this way: Our method creates a scope around an entity. All words in the scope have a smaller distance to all other entities (in number of words between them). In the sentence “Merkel is acclaimed for her government’s work, so the SPD is still performing relatively poorly in polls.” the method would associate “acclaimed” with “Merkel” and “poorly” with “SPD”.

4.3 Determination of the Assignment

To assign a viewpoint for the statements, our algorithm determines the probability of one entity belonging to one group (we use this algorithm also for the friend and competitor assignment). As a consequence, a statement belongs to one or more specific groups, if the probability is maximal under the assumption that all entities within this statement belong to this group (if the two probabilities are equal and maximal, the statement belongs to two groups and so on).

$$g = \arg \max_{g_i \in \{g_1, \dots, g_m\}} P(s|g_i) \quad (9)$$

$$P(s|g_i) = \sum_{e \in E_s} P(e|g_i) \quad (10)$$

The probability of one statement belonging to one specific group is the sum of all single probabilities with which entity e belongs to group g_i (E_s are all entities in statement s).

Algorithm 1: Entity Similarity

```

Data: statement entity  $e$ , group entity  $e_g$ 
Result: similarity  $\sigma$ 
1 if  $e$  and  $e_g$  are the same type then
2    $l_1 \leftarrow \text{getListOfTokens}(e);$ 
3    $l_2 \leftarrow \text{getListOfTokens}(e_g);$ 
4    $m = \max(\text{getSize}(l_1), \text{getSize}(l_2));$ 
5   foreach token  $t_1 \in l_1$  do
6     foreach token  $t_2 \in l_2$  do
7       if  $t_1$  equals  $t_2$  then
8          $m = m - 1;$ 
9       end
10    end
11  end
12   $\sigma = 0.9^m;$ 
13 end
14 else
15    $\sigma = 0.0;$ 
16 end
```

Figure 3: Pseudocode of Entity Similarity.

For this purpose, an entity of statement s is compared to all entities E_{g_i} which belong to group g_i in the ontology.

$$P(e|g_i) = \max_{e_g \in E_{g_i}} (\text{sim}(e, e_g)) \quad (11)$$

The similarity function $\text{sim}(e, e_g)$ compares the name e of the entity in the text and the name of member e_g of group g . If they are the same, the value is 1.0. If e consists of the same tokens and only one token is different, the value is 0.9 (see the pseudocode in Figure 3). A token is one part of a name, e.g. the surname of a person. So, a name of a person could, for example, consist of two tokens: the first name and the surname. The method *equals* in Algorithm 1 checks, if two tokens are have the same string representation (e.g. both names start with “John”).

This is useful, when persons are only mentioned with their surname or when a product’s or organisation’s name is not named in its entirety (company XY → company XY Inc., product ABC → product ABC international).

This also means that two persons, who share the same first name, could have a similarity of at least 0.9. This might sound unpleasant, but we have noticed that the persons are almost always mentioned by their full names in the complete articles, so we include a orthomatching module and a pronomial coreferencer (cf. section 5.2).

Group	Organis.	Persons	Products	all
Finance				
A	6	178	6	190
B	2	58	1	61
C	4	53	2	59
D	5	79	1	85
E	2	16	0	18
Politics				
CDU	9	237	0	246
SPD	9	149	0	158

Table 1: Size of the evaluation ontologies.

5 Evaluation

5.1 Experiment Design

We use two datasets for our experiments: The first test corpus called '**Finance**' (with 4,000 statements) represents a real MRA about a financial service provider and its four competitors. The corpus was created by up to ten media analysts (professional domain experts in the field of media response analyses). Their quality assurance ensures an inter-annotator agreement of at least 75% to 80% using Fleiss' kappa. For our second corpus which we call '**Politics**', we use the press-relations dataset¹ (Scholz et al., 2012) which consists of 1,521 statements from press releases and has two viewpoints (the two competitive parties: CDU and SPD). Here the agreement is 88.08% (Cohen's kappa, two annotators).

All statements (4,000 or 1,521, respectively) are annotated with a tonality and a viewpoint for this tonality (one of the five companies or the two parties, respectively). Small examples from the **Politics** dataset are mentioned in the first section.

For the **Finance** corpus, we create a RDF ontology by extracting all entities of the code book from customer group A (see Table 1; all companies' names are made anonymous for reasons of data protection). Group A has four competitors (group B to E) and the groups D and E have a friendly relationship, because D has taken over E in the first few days of the MRA. All other relationships are competitive. For the **Politics** corpus, we create an ontology in which CDU and SPD are competitors. We add all party members

of the seventeenth German Bundestag² (the German parliament) and add some synonyms of the party and concepts such as 'government' or 'opposition' as organisations (see Table 1).

For evaluation of the **Viewpoint Features** β and **Viewpoint Tonality Features** γ , we use 30 % of the statements to construct sentiment dictionaries which are weighted by the methods explained in section 3 and the remaining statements as training and test set (20% training and 80% test; we use this small training and big test set to guarantee that this approach will also work in practice). In addition, we use SentiWS (Remus et al., 2010) with 1686 positive and 1818 negative words as another baseline. We change the tonality according to viewpoint: the tonality is changed to the negative, if a statement is exclusively positive for a competitor and negative statements for a competitor become positive. For the classification, we use a SVM (Rapidminer³ standard implementation with default parameters) which performed better than other machine learning techniques (k-means, Naive Bayes) in this evaluation task. The evaluation shows the results in different combination of the features. So, $\alpha+\beta$ is the combination of set α and the feature set β and so on and all means the selection of all features.

5.2 Text Preprocessing

For better results, we analyse not only the statements, but also the whole text from which a statement is taken. The basic framework for our approach is GATE (General Architecture for Text Engineering)⁴, which provides a Named Entity Recognition (NER) for several languages. We tag POS with the TreeTagger (Schmid, 1995) whose results have been incorporated in the NER. We add the ontology entities and design new JAPE Rules⁵ to improve our NER: In order to guarantee that these entities are found in most case, the rules treat the ontology entities with the highest priority. Furthermore, we readjust the orthomatching module and edit some rules to avoid failures due to some of the names of organisations. In addition,

²collected from <http://www.bundestag.de>

³Rapid-I: <http://rapid-i.com/>

⁴GATE: <http://gate.ac.uk/>

⁵Developing Language Processing Components with GATE Version 6 (a User Guide): <http://gate.ac.uk/>

¹<http://www.pressrelations.de/research/>

Method	$ s $	c	nc	w
DASA				
Viewpoint A	994	0.5613	0.2374	0.2012
Viewpoint B	1173	0.5303	0.2293	0.2404
Viewpoint C	812	0.5123	0.3067	0.1810
Viewpoint D	682	0.5967	0.2038	0.1994
Viewpoint E	339	0.6018	0.2655	0.1327
All	4000	0.5518	0.2458	0.2025
our approach				
Viewpoint A	994	0.7606	0.0986	0.1408
Viewpoint B	1173	0.7894	0.0648	0.1458
Viewpoint C	812	0.7906	0.0370	0.1724
Viewpoint D	682	0.8372	0.0440	0.1188
Viewpoint E	339	0.8348	0.0590	0.1062
All	4000	0.7945	0.0635	0.1420

Table 2: Results of the group assignment on **Finance** in comparison with DASA (Qiu et al., 2010).

tion, we implement a German pronomial coreferencer. Thus, we can resolve persons, organisations, and products in statements, even if they are only mentioned by he/she/it in the statement, but mentioned by name in the rest of the article.

5.3 Setup for DASA

For our comparison, we implement the DASA algorithm (Qiu et al., 2010). For the sentiment analysis step, we use the polarity annotation of the datasets to identify the opinion words with the same polarity, because we are more interested in the perspective component than in the sentiment analysis component. We calculate the dependencies between opinion and topic words with the Stanford Parser for German (Rafferty and Manning, 2008) and apply the rules R1 to R7 in descending order as described in Qiu (2010). Furthermore, we also expand their approach to non negative sentiments. For their task of selecting ads it is very reasonable to search for negative words and then to recommend rivals’ products, but we also need a viewpoint for non negative statements. We use the same RDF ontologies to create the input information for the DASA method, so that DASA knows the rivals and can use the entities as topic words for the assignment.

5.4 Experiment Results

Table 2 and 3 show the results of the viewpoints assignment. $|s|$ is the number of statements, c are the correctly assigned statements, w are the ones

Method	$ s $	c	nc	w
DASA				
CDU	992	0.5927	0.2258	0.1815
SPD	529	0.5350	0.2155	0.2495
All	1521	0.5726	0.2222	0.2051
our approach				
CDU	992	0.8700	0.0766	0.0534
SPD	529	0.6759	0.0964	0.2287
All	1521	0.8021	0.0835	0.1144

Table 3: Results of the group assignment on **Politics** in comparison with DASA (Qiu et al., 2010).

Method	α	$\alpha+\beta$	$\alpha+\gamma$	all
SentiWS	0.5066	0.5678	0.5200	0.5888
PMI	0.6094	0.6450	0.5909	0.6406
Chi-square	0.6275	0.6388	0.6272	0.6281

Table 4: Results of view based modified tonality on **Finance**.

incorrectly assigned, and nc could not be classified (the probability of all viewpoints is zero or DASA does not find a viewpoint, respectively). The average performance for statements are correctly classified in more than 79% or 80% of cases, less than 15% or 12% are classified incorrectly and over 6% or 8% are not classified at all. This is an improvement about 24% or 22% against the DASA algorithm.

We use the accuracy evaluation metric for the evaluation of the **View Modified Tonality**.

$$ACC = \frac{c}{n} \quad (12)$$

In equation 12, c is the number of correctly predicted statements and n is the number of all statements in the test set. In contrast to the view assignment, we only use the non neutral statements from the **Politics** corpus in this task (1,029 statements). But we use all 4,000 statements from **Finance**, because all statements are subjective (positive or negative).

Table 4 and 5 show the statement results of which the tonality is modified by a given point of view. The improvement expands from over 1% (chi-square, $\alpha+\beta$ combination on **Finance**) to over 8% (SentiWS, all features on both datasets). Almost all methods achieved the best results, if all features are combined.

In contrast, Table 6 shows the results using the **Basic Tonality Features** α , if the tonality is not

Method	α	$\alpha+\beta$	$\alpha+\gamma$	all
SentiWS	0.5342	0.5259	0.5259	0.6177
PMI	0.6294	0.6077	0.6260	0.6427
Chi-square	0.6694	0.6494	0.6678	0.6761

Table 5: Results of view based modified tonality on **Politics**.

Method	not mod. Finance	not mod. Politics
SentiWS	0.6455	0.5526
PMI	0.6393	0.6528
Chi-square	0.6848	0.6878

Table 6: Results of both datasets which are not modified through a viewpoint.

modified through a given view. The results are of course higher, because this task of a not modified tonality is simpler and the results do not provide any information about the viewpoint. But the combination of all viewpoint features approach these results (the PMI method or SentiWS with all features achieved even a better result on **Finance** or **Politics**, respectively).

5.5 Error Analysis

If we examine the statement set **Finance** for the group assignment task, the reason for the worst performance of the customer group is the type of the collected statements. The customer is given more attention during a MRA and so articles are collected which do not directly contain the known entities, but they include general messages: “Markets are suffering from the consequences of the economical crisis.” This is also the reason why this group has the highest rate of not classified statements.

One reason for the better performance of our approach against the DASA algorithm is the entity similarity and assignment of the most likely viewpoint which decreases the number wrong and especially not classified statements.

Both methods (DASA and our approach) perform only slightly better on **Politics**, although **Politics** has only two different viewpoints, because this area is more characterized by comparative statements. The press releases of CDU are talking much about the SPD and vice versa, what it makes more difficult to extract the correct viewpoint of a statement.

The results of the view modified tonality are not on such a high level. But this is a very hard task even for humans and not the state-of-art method to code statements in a real media analysis. Besides, the improvement of the viewpoint features approach the existing solutions of the not modified results which do not provide any information about the viewpoint.

6 Conclusion and Future Work

In conclusion, our ontology based approach provides a tonality based on a specific viewpoint. The group assignment algorithm and the viewpoint features allow the coding of statements into certain groups and tonality mutation based on viewpoint. Results of the evaluation suggest that both options (viewpoint assignment and viewpoint features) are possible for view based approach. The viewpoint features improve the accuracy (the results are not far away from the result, when the tonality is not modified through a view), while the assignment algorithm provide concrete viewpoint determination. This information is valuable and a two-process solution of calculating tonality and viewpoint fits the way of state-of-the-art companies coding their MRA.

But the evaluation also shows that the information itself about the viewpoint can increase the calculation of the tonality. So, our future work will cover the design of more techniques to improve the classification even further by using this approach for the viewpoint assignment first.

Acknowledgments.

This work is funded by the German Federal Ministry of Economics and Technology under the ZIM-program (Grant No. KF2846501ED1). The authors want to thank our cooperation partner pressrelations, especially Leszek Blacha for the creation of the dataset and Gudrun Karbe for helpful comments during the writing process.

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*.

- Alexandra Balahur, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen, and Mijail Kabadjoy. 2009. Opinion mining on newspaper quotations. In *Proc. of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, pages 523–526.
- Alexandra Balahur, Ralf Steinberger, Mijail Kabadjoy, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*.
- Danushka Bollegala, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 132–141.
- Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proc. of the 27th annual meeting on Association for Computational Linguistics, ACL '89*, pages 76–83.
- Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.
- Ali Harb, Michel Plantié, Gerard Dray, Mathieu Roche, François Trouset, and Pascal Poncelet. 2008. Web opinion mining: how to extract opinions from blogs? In *Proc. of the 5th international conference on Soft computing as transdisciplinary science and technology, CSTST '08*, pages 211–217.
- Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
- Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proc. of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 584–589.
- Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.
- David Michaelson and Toni L. Griffin. 2005. A new model for media content analysis. Technical report, The Institute for Public Relations.
- Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of the 7th conference on International Language Resources and Evaluation (LREC'10)*.
- Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.
- Sounil Park, KyungSoon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 340–349.
- Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. 2010. Dasa: Dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9):6182 – 6191.
- Anna N. Rafferty and Christopher D. Manning. 2008. Parsing three german treebanks: lexicalized and unlexicalized baselines. In *Proc. of the Workshop on Parsing German, PaGe '08*, pages 40–46.
- R. Remus, U. Quasthoff, and G. Heyer. 2010. SentiWS – a publicly available german-language resource for sentiment analysis. In *Proc. of the 7th International Language Resources and Evaluation (LREC'10)*.
- Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proc. of the ACL SIGDAT-Workshop*, pages 47–50.
- Thomas Scholz, Stefan Conrad, and Lutz Hillekamps. 2012. Opinion mining on a german corpus of a media response analysis. In *Proc. of the 17th international conference on Natural language processing and information systems, TSD'12*.
- Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.
- Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 417–424.
- T. Watson and P. Noble, 2007. *Evaluating public relations: a best practice guide to public relations planning, research & evaluation*, chapter 6, pages 107–138. PR in practice series. Kogan Page.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

A Supervised POS Tagger for Written Arabic Social Networking Corpora

Rania Al-Sabbagh

Department of Linguistics

University of Illinois at Urbana-Champaign

USA

+1-217-848-0020

alsabba1@illinois.edu

Abstract

This paper presents an implementation of Brill's Transformation-Based Part-of-Speech (POS) tagging algorithm trained on a manually-annotated Twitter-based Egyptian Arabic corpus of 423,691 tokens and 70,163 types. Unlike standard POS morpho-syntactic annotation schemes which label each word based on its word-level morpho-syntactic features, we use a function-based annotation scheme in which words are labeled based on their grammatical functions rather than their morpho-syntactic structures given that these two do not necessarily map. While a standard morpho-syntactic scheme makes comparisons with other work easier, the function-based scheme is assumed to be more efficient for building higher-up tools such as base-phrase chunkers, dependency parsers and for NLP applications like subjectivity and sentiment analysis. The function-based scheme also gives new insights about linguistic structural realizations specific to Egyptian Arabic which is currently an under-resourced language.

1 Introduction

Part-of-Speech (POS) tagging is an enabling technology required for higher-up Natural Language Processing (NLP) tools such as chunkers and parsers – syntactic, semantic and discourse; all of which are used for such applications as subjectivity and sentiment analysis, text summarization and machine translation among others. Labeling words for their grammatical categories (i.e. POS tagging) is a non-trivial process given the inherent ambiguity of natural languages at various linguistic levels.

Genre-specific features can also pose extra challenges to POS tagging. The interactive conversational nature of the microblogging service Twitter introduces highly-dialectal input in which

Roxana Girju

Department of Linguistics

University of Illinois at Urbana-Champaign

USA

+1-217-244-3104

girju@illinois.edu

new words are coined on frequent basis. This implies that using non-corpus-based approaches, using POS taggers designed for Modern Standard Arabic (MSA) or leveraging Egyptian Arabic (EA) taggers from MSA ones are unlikely to perform well. These implications are empirically proved in prior work. Habash and Rambow (2006) achieve a coverage rate of only 60% for Levantine Arabic using MSA morphological analyzer. Abo Baker et al. (2008) and Salloum and Habash (2011) build linguistically inaccurate morphological analyzers trying to extend MSA tools to Arabic dialects. Duh and Kirchhoff (2005) build a minimally supervised POS tagger for EA of only 70.88% accuracy by using an MSA morphological analyzer and adding the Levantine Arabic TreeBank to their EA training corpus to benefit from the cross-dialect overlap in Arabic.

We, therefore, start our experiments for building a POS tagger for EA tweets with a supervised Transformation-Based Learning (TBL) approach trained and tested on EA tweets only. One advantage of this approach is its non-stochastic mechanism that is unlikely to be affected by frequent new word coinages and the sparsity they introduce to the training corpus.

We propose a function-based POS annotation scheme for this paper. Instead of standard morpho-syntactic annotation schemes which use word-level morpho-syntactic information for POS tagging, our scheme labels words based on their grammatical functions instead of their morpho-syntactic structures. Direct mapping between the word grammatical function and its morpho-syntactic structure is not always granted. This annotation scheme requires a new tagset that adapt tags from standard tagsets like Arabic TreeBank (ATB) and also uses new tags. The main advantage of our function-based scheme and tagset is to enhance developing such NLP tools as chunkers, dependency and discourse parsers and such applications as

subjectivity and sentiment analysis systems which is one main application in mind while building our POS tagger. Evidence on enhancing subjectivity and sentiment analysis systems and dependency parsers as well as comparing our scheme and tagset to more standardized ones are both kept for future work. Our POS tagger is still a contribution to the repository of Dialectal Arabic (DA) NLP tools which is to-date limited.

The rest of this paper is organized as follows: section 2 briefly discusses related work to POS tagging, focusing particularly on dialectal Arabic. Section 3 describes our POS tagset and adaptations made from standardized tagsets. Section 4 explains our function-based tokenization and tagging scheme. Section 5 describes the corpus and preprocessing procedures. Section 6 explains the annotation process and inter-annotator agreement rates. Section 7 discusses Brill's implementation of transformation-based learning to POS tagging and one application of it on MSA. Section 8 elaborates on our evaluation results and error analysis. Finally, section 8 outlooks major conclusions and future plans.

2 Related Work

Of the most recent NLP tools built for EA is Habash et al. (2012). Extending the Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002) and following the POS guidelines by the Linguistic Data Consortium (LDC) for Egyptian Arabic (Maamouri et al., 2012a as cited in Habash et al., 2012), they build the large-scale morphological analyzer – CALIMA. It relies on tabular representation of complex prefixes, complex suffixes, stems and compatibility across them (prefix-stem, prefix-suffix and stem-suffix). Tested against a manually-annotated EA corpus of 3,300 words (Maamouri et al., 2012b as cited in Habash et al., 2012), CALIMA achieves a coverage rate of 92.1% where coverage is defined as the percentage of the test words whose correct analysis in context appears among the analyses returned by the analyzer. It also provides among its results a correct answer for POS tags 84% of the time.

With the goal of utilizing MSA morphological tools to create an EA training corpus and using data from several varieties of Arabic in combination, Duh and Kirchhoff (2005) build a minimally supervised EA tagger without the need to develop dialect-specific tools or resources. For data, they use the CallHome Egyptian Arabic corpus from LDC, the LDC Levantine Arabic corpus and the Penn Arabic

Treebank corpus parts 1 to 3. For the morphological analyzer, they use Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2002) designed for MSA. Their approach bootstraps the EA tagger in an unsupervised way using POS information from BAMA and subsequently improves it by integrating additional data from other dialects given the assumption that Arabic dialects do overlap. Tested against Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002), their best accuracy rate is 70.88%. Adding word-level features such as affixes and constrained lexicon first raises accuracy from 62.76% to 69.83% and then adding Levantine data to the training corpus raises accuracy to 70.88%.

MAGEAD (Habash and Rambow, 2006) is a morphological analyzer and generator for the Arabic language family – MSA and dialect. It uses the root-pattern-features representation for online analysis and generation of Arabic words. Tested against the Levantine Arabic Treebank, MAGEAD achieves a context-type recall rate of 95.4% and a context-token recall rate of 94.2%. Context-token/type recall is defined as the number of times MAGEAD gets the contextually correct analysis for that word token/type.

Diab et al. (2010) build a large annotated corpus for multiple Arabic dialects, of which EA is a part. The corpus contains texts from blogs covering the domains of social issues, religion and politics linguistically analyzed at different levels. In addition to morphological analyses, Diab et al. (2010) give information about POS tags, the degree of dialectness of each word and sentence boundaries. Much of the work is being done manually or is the output of MAGEAD – after being tuned for DA. Performance rates for each task are not, however, reported.

3 The POS Tagset

There is a large number of Arabic POS tagsets including: BUCKWALTER (Buckwalter, 2002) used in the Penn Arabic TreeBank (ATB), Khoja tagset (Khoja, 2001), PADT tagset (Hajič et al., 2004), Reduced Tagset (RTS) (Diab, 2007) and CATiB POS tagset (Habash and Roth, 2009). Each of these tagsets represents a different level of granularity: at one end of the continuum is the most fine-grained tagset of Buckwalter with over 500 tags and at the opposite end is the most coarse-grained tagset of CATiB with only six tags. For a full review of these tagsets refer to Habash (2010).

Our tagset mixes and matches tags across fine-grained tagsets – to achieve the following goals:

- Give a fine-grained level of accurate linguistic description for the word POS and its semantic features of gender, number, person, aspect, voice, tense and mode.
 - Tag words that can be used – as in future work
 - as classification features for base-phrase chunkers and parsers. These words include function words such as interrogatives, complementizers, conditionals and the like.
 - Tag parts of speech that can be used – as in future work – as subjectivity and sentiment classification features like modals and negation among others.

In appendix (A), we compare and contrast our tagset with that of ATB and RTS to facilitate comparing results with other taggers – if any – that are using different tagsets. Our tagset is a subset of ATB and is a superset of RTS.

We add new tags to label Twitter-specific information and EA-specific grammatical categories like fixed expressions, existentials and aspectual progressives. Twitter-specific information requires tags for mentions(MNT), hashtags (HSH), emoticons (EMO), URLs (URL) and speech effects (LNG; for LeNGthened words) (e.g. اooooوووی Awwwwwy (very) and خنیق خنیقیк xnyyyyyq (boring)).

Approximately, 1% of our corpus is given our new tag EXP – for fixed expression. We define fixed expressions according to the following criteria:

- They can be either unigram or multiword expressions;
 - Multiword fixed expressions are frozen in the sense that their individual words are not substitutable for synonyms. However, some of those expressions might have shorter versions;
 - Their meaning is not compositional and are rarely – if not never – used literally;
 - Their grammatical behavior does not match that of nouns, verbs, adjectives or adverbs. In other words, they cannot be head nouns or verbs in noun or verb phrases, respectively. They cannot modify nouns like adjectives or modify verbs like adverbs.
 - They are used for pragmatic purposes to show, for example, shock as in يالهوي *yAlhwiy* (Oh my goodness!), surprise as in ياطولي *yAHlwiy* (how interesting!) and frustration as in الصير *AlSbr mn Endk yArb* (lit: patience is from you, Lord; gloss: Oh, Lord! Grant me patience) among other emotions.

The unigram fixed expression يَالْهُوَي yAlhwy (Oh my goodness) is diachronically composed of the vocative particle يَ yA (oh), the noun لَهُوَ lhw (goodness) and the possessive pronoun مِنْ y (my). Yet, it cannot be decomposed into its parts and none of its parts can be substituted for a synonym. It functions only to show shock, anger, frustration, emotions and the like. Meanwhile, its syntactic behavior does not fit in the paradigms of verbs, nouns, adjectives or adverbs.

The same thing applies to the multiword fixed expression الصبر من عندك يارب *AlSbr mn Endk yArb* (lit: patience is from you, Lord; gloss: Oh, Lord! Grant me patience). It is very rarely used literally as a prayer. As one expression, it does not grammatically behave like nouns, verbs, adjectives or adverbs. It is typically used as an expression of frustration or anger. Yet, it shows some degree of structural flexibility given that a shorter form exists الصبر يارب *AlSbr yArb*.

Two other tags that we use although they do not have equivalents in previous Arabic tagsets are: EX and PG for existentials and aspectual progressives, respectively. Unlike MSA, existentials in EA are not expressed by the deictic *hnAk* (there) or the imperfect verb *ywjd* (exist). They are expressed by the preposition *fy* or *فه* *fyh*. Should these prepositions be used as existentials, they can syntactically map to a complete sentence such as *في موافقة fyh mwAfqp* (there is an agreement). Thus adding the EX tag to our tagset serves the purpose of facilitating phrase-boundary identification in later annotation layers for chunkers and parsers.

Unlike MSA, EA has an aspectual progressive verb prefix $\rightarrow b$ found in examples like بيكتب *byktb* (he's writing), بيقول *byfkr* (he's thinking) and بتقول *btqwl* (she's saying). The aspectual progressive prefix is split off in tokenization and is tagged as PG.

Another tag that we add is MD to tag modals and modal adjuncts – in both verbal and nominal forms. For example, both the modal verb *ymkn* (may) in يمکن *ymkn* *yjtmEw* (they may meet) and the modal adjective *Drwry* (must) in ضروري *Drwry* *nrwH* (we must go) are both tagged as MD. MD is used to tag all modality types – epistemic, deontic and evidential.

Some tagsets like RTS give simple, comparative and superlative adjectives one tag – JJ. ATB labels only simple and comparative adjectives, given that superlative adjectives are not morphologically marked. In our tagset, simple, comparative and

superlative adjectives are tagged as JJ, JJR and JJS, respectively.

We collapse some tag subsets in ATB into one. Instead of distinguishing connective particles from coordinating conjunctions, both are given the CC tag because both coordinate base and complex phrases. Thus و w (and), او >w (or), و بعدين wbEdyn (and then) and وكمان wkmAn (and also) are all tagged as CC.

Another collapsed tag subset from ATB is the interrogative subset. Instead of three different tags for interrogative adverbs, particles and pronouns, one tag is used, namely INT for interrogative. For instance, the interrogative adverbs ازاي <zAy (how) and فین fyn (where) as well as the interrogative pronouns ايه Ayh (what) and مين myn (who) are all tagged as INT. The Arabic interrogative particles – هل hl and أ > used in MSA to form yes/no questions are not used in EA. When encountered in MSA tweets, they are also labeled as INT.

Relative adverbs and pronouns are also collapsed into one tag RL. EA has one relative pronoun – اللي Ally (who, which, that). When MSA relative pronouns or adverbs are encountered, they are tagged as RL.

Our verb tag subset does not define the voice feature (active vs. passive) which is given a separate tag – P for passive and the absence of such a tag indicates an active voice. Our noun tag subset does not indicate the number feature – singular, dual or plural – of the noun since these are given their separate tag subset. Therefore, NN is a common noun and NNP is a proper noun whether singular, dual, plural or a collective noun.

Based on our 49 base tagset, each content word and some function words are given complex tag vectors of the form <person>_<number>_<gender>_<voice>_<grammatical category>. Currently, our corpus has a total of 4,272 unique vectors. Some examples are in table (1).

Input	Tokenized	POS Tagged
يقول byqwl (he's saying)	b- yqwl	b/PG yqwl/3_SG_M_VBP
شافهم \$Afhm (he saw them)	\$Af -hm	\$Af/3_SG_M_VBD hm/3_PL_OBJP
يتكتب bttktb (it's being written)	b-ttktb	b/PG ttktb/3_SG_F_P_VBP
الحكومة AlHkwmp (the government)	Al-Hkwmp	Al/DT Hkwmp/SG_F_NN

كويسين kwysyn (good; plural)	kwysyn	kwysyn/PL_JJ
هي hy (she)	Hy	hy/3_SG_F_SBJP
ليكى lyky (for you)	ly-ky	ly/IN ky/2_SG_F_OBJP

Table 1: Tokenization and POS tagging examples

4 Function-Based Tokenization and POS Tagging

Almost all tokenization and POS tagging approaches for Arabic – MSA or dialectal Arabic – rely on word-level morpho-syntactic structures for tokenization and POS tagging. In this paper, we present a function-based tokenization and POS tagging scheme in which words are tokenized and POS tagged based on their grammatical functions rather than their morpho-syntactic structures given that these two do not necessarily map. For example, المسيرة زمانها خلصت zmAnhA in Almsyrap zmAnhA xlSt (the march must have finished) is labeled as MD (i.e. modal) because it functions as a modal (*must have*) despite being morpho-structurally a noun زمان zmAn (time; era) and a possessive pronoun ها hA (her; hers). The same word in another context – as in مصر بتتدى زمانها الجديد mSr btbtidy zmAnhA Aljdyd (Egypt is starting its new era) – is tokenized as zmAn-hA and tagged as zmAn/SG_M_NN hA/3_SG_F_PP\$.

Using this function-based scheme for our tokenization and POS tagging provides a gold standard corpus for training and testing lexico-syntactic disambiguators, base-phrase chunkers and parsers. We leave it for future work to compare the performance of those tasks when trained on our function-based scheme and when trained on morpho-syntactic schemes.

The grammatical categories affected by our function-based scheme are: existentials, prepositional phrases, active participles, modals, superlative adjectives, multiword connective particles and fixed expressions. These are the grammatical categories which are typically ambiguous in terms of the mapping between their morpho-syntactic structures and their grammatical functions.

فيه مؤتمر فيه fyh (there is/are) in فيه صحفى الساعه 9 fyh m&tmr SHfy AlsAEp 9 (there is a press conference at 9 o'clock) consists morphologically from the preposition في fy (in) and the enclitic pronoun هـ h (him). However, in this context, this morphological structure is irrelevant

because the enclitic pronoun is an impersonal pronoun without a referent. The entire word functions as an existential and is thus tagged as *fyh/EX* without tokenizing the final pronoun *h*. The same word in *في حنجمع HntjmE fyh* (we'll gather in it) is treated differently because it literally means *in it*; thus it is first tokenized as *fy-h* and then tagged as *fy/IN* and *h/3_SG_M_OBJP*.

Prepositional Phrases (PPs) are not always literally used in EA. For example, بسرعة *bsrEp* - which morphologically consists of the preposition *بـ b* (with) and the noun سرعة *srEp* (speed) – functions as an adverb in انزلوا بسرعة ع الميدان *Anzlw bsrEp E AlmydAn* (come quickly to the square). Therefore, in this sentence, it is tokenized as one word and tagged as *bsrEp/RB*. In مهتم بسرعة حل الأزمة *mhtm bsrEp Hl Al>zmp* (he's concerned with a quick crisis solution), the same PP is literally used and thus it is first tokenized as *b-srEp* and then tagged as *b/IN srEp/SG_F_NN*. The same procedure is used with more complex PPs like بـ *b\$kl mZbwT* (in a perfect way). In عايزين نحلها بـ *EAyzyn nHlhA b\$kl mZbwT* (we want to sort it out in a perfect way), the PP functions as an adverb modifying the verb نحلها *nHlhA* (sort it out). Therefore, in this context it is tagged as one word – *b\$kl_mZbwT/RB*.

Active participles in ATB are tagged as nouns or adjectives in POS annotation level and then a verbal noun (VN) tag is added in the treebank annotation level. In EA, active participles are not only used as nouns and adjectives but also as imperfect verbs. Thus they are tagged according to their grammatical function in context as NN, JJ or VBP. In عايزين نفهم *EAyzyn nfhm* (we want to understand), the active participle عايزين *EAyzyn* is used as a verb meaning *we want*; thus it is tagged as *Eyzyn/1_PL_VBP*. The same applies to فاهماهم *fAhmAhm* in هي مش فاهماهم *hy m\$ fAhmAhm* (she does not understand them) in which *fAhmAhm* functions as a verb meaning *understand* attached to the object pronoun *them*. Thus it is tokenized as *fAhmA-hm* and tagged as *fAhmA/3_SG_F_VBP* and *hm/3_PL_OBJP*.

When active participles are used as nouns or adjectives, they are tagged as such. In هو شاهد اثبات *hw \$Ahd AvbAt* (he's an prosecution witness), the active participle شاهد *\$Ahd* (witness) is tagged as *\$Ahd/SG_M_NN* being used as a noun. In المياه كافية *Almyh kAfyh* (the water is enough), the active participle كافية *kAfyh* is used as an adjective and is thus tagged as *kAfyh/SG_F_JJ*.

Modals – including verbal, nominal and adjunct modals – are also tokenized and POS tagged

functionally. In يمكن نيجي *ymkn nyjy* (we may come), the modal verb يمكن *ymkn* (may) is tagged as *ymkn/MD*. The same modal function can be realized using an adjective form ممكن *mmkn* (may) as in ممکن مایکونش مفهوم *mmkn mAykwn\$ mfhwm* (I may not be understood); which is thus tagged as *mmkn/MD*. The same adjective in a different context like كل شيء ممكن *kl \$y' mmkn* (everything is possible) is not used modally and is tagged as *mmkn/SG_M_JJ*.

Modal adjuncts are typically multiword, yet should they be modals, they are tokenized and tagged as one unit. In في احتمال انه مابينعش *fy AHtmAl Anh mAynfE\$* (there is a possibility that it won't work), the modal adjunct في احتمال *fy AHtmAl* is tokenized and tagged as one word – *fy_AHtmAl/MD*. The same word احتمال *AHtmAl* can be tagged as a noun in a different context like احتمال فوزه ضعيف *AHtmAl fwzh DEyf* (his possibility/chance of winning is weak) – *AHtmAl/SG_M_NN*.

Multiword connective particles like وكمان *wkmAn* (and also) and وبعدين *wbEdyn* (and then) are also tokenized and tagged as one word. Each of these particles morphologically consists of the coordinating conjunction *و w* (and) and a connective particle. These are tokenized and tagged as *wkmAn/CC* and *wbEdyn/CC*.

Multiword fixed expressions – that match our definition of multiword fixed expressions in section 3 – function as one whole unit to serve a certain pragmatic meaning. These are tokenized and tagged, thus, as one unit. The multiword fixed expression لا مواجهة *IA m&Axzp* (lit: no offense; gloss: excuse me) consists of the negative particle لا *IA* (no) and the noun مواجهة *m&Axzp* (offense); yet being a fixed frozen expression it is tagged as *IA_m&Axzp/EXP*. Similarly, متوا بغيظكم *mwtwA bgyZkm* (lit: get lost with your anger; gloss: go to hell!) is not decomposed into a verb, a preposition, a noun and a possessive pronoun but is tokenized and tagged as *mwtwA_bgyZkm/ EXP*. Tagging the pragmatic values of these fixed expressions – both unigram and multiword – is kept for future work.

5 Corpus Description

Our corpus is 22,834 tweets complied over the period from May 2011 to December 2011. It is a subset of the microblog portion of YADAC (Al-Sabbagh and Girju, 2012). The corpus contains 423,691 tokens and 70,163 types preprocessed according to the procedures in the following lines.

Only tweets scored above the degree of dialectness threshold set by Al-Sabbagh and Girju (2012) are selected. This guarantees a highly dialectal corpus; yet MSA is still likely to be found. Arabic tweets written in Latin Script (AiLS) are already excluded as well as Foreign tweets written in Arabic Script (FiAS). Two normalization steps are used for spelling variation and speech effects.

To reduce the effect of spelling variation – given the lack of standard spelling conventions in EA and most Arabic dialects – we used a normalization rule-based module based on the conventions set by the Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2011)¹. To augment the performance of the spelling normalization module and deal with cases which CODA does not currently handle, we use the vowel-based spelling variation model and the 138-entry lexicon of unpredictable spelling variations both built by Al-Sabbagh and Girju (2012).

6 Gold-Standard Annotation

Two annotators of intermediate-level linguistic training (i.e. undergraduate linguistics students) – who are native EA speakers – are used to annotate the corpus over a period of 4 months. Two other annotators – graduate linguistics students – are then used to review the annotations for consistency and correctness over a period of one month. The first two annotators achieve an inter-annotator Kappa coefficient rate of 97.3% for tokenization and 88.5% for POS tagging. The review annotators achieve a rate of 99.6% for tokenization and 98.2% for POS tagging. Main differences between the two groups of

¹ Nizar Habash, Mona Diab and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic (CODA). *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May 2012, Istanbul, Turkey, 711-718

annotators are about the consistency in applying our function-based annotation scheme.

The first annotation phase – tokenization – is a light stemming process to split off the following clitics and affixes:

- definite article الـ *Al* (the)
 - prepositions بـ *b* (with) and لـ *l* (for)
 - connective conjunctions وـ *w* (and) and فـ *f* (then);
 - vocative particle ياـ *yA* (oh, hey)
 - object pronouns نـي *ny* (me), تـ *nA* (us), كـ *k* (you; singular), كـم *km* (you; plural), هـ *h* (him), هـا *hA* (her) and هـم *hm* (them)
 - possessive pronouns يـ *y* (my), تـ *nA* (our), كـ *k* (your; singular), كـم *km* (your; plural), هـ *h* (his), هـا *hA* (her) and هـم *hm* (their)
 - aspectual progressive prefix بـ *b*
 - future tense prefixes هـ *h* and هـ *H* (will)

Gender and number affixes are not split off, however, given that they affect semantic feature tagging: جمیل *jmyl* (beautiful; masculine) is tagged as *jmyl/SG_M_JJ*; whereas جمیلۃ *jmylp* (beautiful; feminine) is tagged as *jmylp/SG_F_JJ*.

Light stemming also involves reversing morphotactic changes resulted from clitic/affix attachment. Attaching a possessive pronoun to a noun ending in *ta' marbuta* – تاءً مربوطة – changes it into *ta' maftouha* – تاءً مفتوحة – as in قضيّة *qdDyp* (issue) and قضيّتي *qdDty* (my issue). Similarly, attaching an object pronoun to a plural verb ending in و *wA* – the verb plural marker – leads to removing the *alef* as in شافونـي *\$Afwny* (they saw) and شافـوا *\$AfwnA* (they saw me). When splitting off clitics and affixes, these morphotactic changes are reversed.

The second annotation phase – POS tagging – involves tagging words according to our function-based scheme and lexicon lookups. For fixed expressions that match our definition, we have a lexicon of 539 expressions that are 0.5% of our corpus. We also have a lexicon of unambiguous function words that contains 2,193 words. A lexicon from (Elghamry et al., 2008) is used for tagging the semantic features of gender and number. Words not found in the lexicon are manually labeled.

7 Transformation-Based Learning

Brill (1994) introduces Transformation-Based Learning (TBL) to POS tagging as an error-driven, corpus-based approach to induce tagging rules out of a gold-standard training corpus. It captures linguistic information in a small number of simple non-

stochastic rules as opposed to large numbers of lexical and contextual probabilities.

AlGahtani et al. (2009) apply TBL to MSA tagging with a main modification of applying the algorithm to lexeme-affix combinations. Affixes are used as cues for POS tags, while affix-free words are looked up in a lexicon. For Out-Of-Vocabulary (OOV) words, they first use the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2002) with a disambiguation module to pick the contextually correct word analysis; and second, words found in neither the lexicon nor BAMA output are tagged as proper nouns. Trained on 90% of ATB and tested on 10%, the algorithm achieves an accuracy rate of 96.5% which is comparable to the state-of-the-art results achieved for MSA using other algorithms like Hidden Markov Models (AlShamsi and Guessoum, 2006), Support Vector Machines (Diab, 2009) and memory-based learning approaches (Marsi and Bosch, 2005). That is why AlGahtani et al. (2009) argue that TBL is simple, non-complex, language-independent and of comparable results to other POS tagging approaches.

In this paper, we use Brill's TBL implementation for POS tagging and a tokenizer built on the same algorithm.

8 Evaluation and Discussion

We perform 10-fold cross validation and use standard precision, recall and F-measure as evaluation matrices. Our output is evaluated in three modes: tokenization (TOK) only, POS tagging without semantic feature labeling (POS) and POS tagging with tokenization and semantic features labeling (ALL). According to the results in table (2), the tokenization module achieves comparable results to the stat-of-the-art systems built for MSA. The performance of POS module and the semantic-feature module – that decreases performance dramatically – still need improvements.

	Precision	Recall	F-Measure
TOK	95%	94%	94.5%
POS	86.5%	88.8%	87.6%
ALL	81%	83.6%	82.3%

Table 2: Precision, recall and F-measure rates for the TOK, POS and ALL modules

About 6% of our corpus is three-letter words like أكل >*kl*. These words are highly ambiguous as they can have multiple readings based on the short vowel pattern with which they are produced. Given that EA

text – like most MSA text – does not use diacritics to mark the short vowel patterns in text, these word are highly ambiguous. Our example >*kl* can read as >*kl* (food), >*kal* (he ate), >*akul* (I eat). The word مصر *mSr* can be *maSr* (Egypt) or *muSir* (persistent). Ambiguity increases when the word has more than one reading of the same grammatical category – nominal or verbal. If a word is ambiguous as a verb or a noun, the different contextual distribution of the verbs and nouns can resolve the ambiguity. Yet, if the word is ambiguous as noun or adjective – both are nominal categories – contextual distribution might not be as efficient for disambiguation. The same applies when the word is ambiguous between an imperfect verb and a perfect verb. Therefore, the adjective *muSir* (persistent) is always tagged as noun for *maSr* (Egypt) in our output. Similarly, >*akul* (I eat) is erroneously tagged as the perfect verb >*kal* (he ate).

The same thing applies to two-letter words. The word حـ *gd* can be a noun meaning *grandfather*, a noun meaning *seriousness* or an adverb meaning *seriously*. Similarly, حـ *Hb* can be a noun meaning *love* or *grains*, or a perfect verb meaning *he loved*.

Intra-grammatical-category ambiguity is also evident in longer words like المـصـري *AlmSry* (the Egyptian) which has two possible nominal tags – noun vs. adjective. Being both nominal, nouns and adjective occur in similar contexts and also share a considerable number of clitics and affixes, which might not be useful POS features in this case.

Ambiguous function words with lexical meanings also lead to output errors. For example, طـبـ *Tyb* is both an interjection meaning *then* as in *then what?* and an adjective meaning *kind*. In الشـعـبـ الـمـصـرـي طـبـ *Al\$Eb AlmSry Tyb*, *Tyb* can mean both depending on how the sentence is read. With a rising final intonation, it means *then (the Egyptian people, then?)*. With a falling final intonation, it means *kind (the Egyptian people is kind)*. There is no way to represent intonation in written text and a wider context across multiple tweets is required to decide whether this tweet is a part of a conversation: if it is, then both intonations are possible and a deeper analysis of the conversation is required to know which intonation is intended; if not, then the falling intonation and thus the *kind* meaning is more likely.

The word بـقـى *bqY* can be a perfect verb meaning *remained* as in *hw dh Ally bqY* (this is what remained) or a discourse particle meaning *so* as in بـقـى هـ دـهـ الـلـاـيـ بـقـى *bqY hw dh AlIhl AlEbgry* (so is this the genius solution?). Similarly, خـلاـصـ *xLAS*

can be a fixed unigram expression meaning "that's it" as in *xlAS m\$ nAfE* (that's it. It's not working) or a noun meaning *salvation* خلاص الناس دى علی ایده *AlnAs dy ELY Aydh* (the salvation of those people is through him).

Typos contribute less than 0.5% of errors. This might indicate that the corpus is not as noisy as it might have been assumed.

Our function-based scheme might have introduced ambiguities at this level of annotation because for example instead of tagging all instances of *bsrEp* as *b/IN* and *srEp/SG_F_NN*, the algorithm has to learn when each instance is used as an adverb and when it is used as a PP. the same thing applies to all grammatical categories affected by our scheme. This, however, does not mean that these are *new* ambiguity types caused by our scheme; eventually these ambiguities will come up in other higher annotation layers.

Results in table (2) show that tagging the semantic features of gender, number, person, aspect, tense and mode decrease performance by about 5%. EA normalizes the morphological distinctions of many of these features and only through long dependencies – which are beyond our tagger – these features can be recovered. For example, كتبت *ktbt* can refer to a perfect verb in 1st person (I wrote), 2nd (you wrote) or 3rd feminine person (she wrote) based on how it is pronounced. With the absence of disambiguating diacritics in written EA, verbs of this class are highly ambiguous in terms of person.

The morphological distinction between duals and plurals is waived in the morphology of EA nouns, verbs, adjectives and pronouns. A plural form of any of these grammatical categories can refer to either duals or plurals. Long dependencies and sometimes metalinguistic information are required to recover the number feature. Alkuhlani and Habash (2012) conduct a series of experiments regarding recovering such latent semantic features; some of which are tried for EA in our future work.

9 Conclusion and Future Work

This paper presented a transformation-based POS tokenizer and tagger for Egyptian Arabic tweets. It proposed a function-based scheme in which words are tokenized and tagged based on their function rather than their morpho-syntactic structure. Among the grammatical categories in which morpho-syntactic structures and grammatical functions do not always map are existentials, prepositional

phrases, active participles, modals, superlative adjectives, multiword connective particles and fixed expressions. The function-based algorithm is expected to enhance performance for higher-order NLP tools such as chunkers and parsers. Despite the promising results, which introduce a new NLP tool to the repository of the resource-poor EA language, much improvement is required.

Short-term improvement plans include: (1) using a different algorithm known for high performance on text processing tasks like Support Vector Machine and defining both tokenization and POS tagging as classification problems; (2) comparing the function-based scheme to ours to know how much ambiguity is resolved or introduced by our function-based scheme; and (3) comparing the two scheme in terms of their performance and enhancement for higher-order NLP tools.

Long-term improvement plans include: (1) building working on word sense disambiguation modules to improve performance on highly ambiguous words and (2) building modules to accommodate for such features as intonation that are unrecoverable from text, yet can affect performance.

References

- Eric Brill. 1994. Some Advances in Rule-Based Part of Speech Tagging. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Washington, 722-727
- Erwin Marsi and Antal van den Bosch. 2005 Memory-based Morphological Analysis Generation and Part-Of-Speech Tagging of Arabic. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* 1-8. Ann Arbor: Association for Computational Linguistics
- Fatma AlShamsi and Ahmed Guessoum. 2006. A Hidden Markov Model-Based POS Tagger for Arabic. *JADT 2006: 8th International Conference of Text Statistical Analysis*.
- Hanaa Kilany H. Gadalla A. Arram, Yacoub, A. El-Habashi and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22
- Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.
- Jan Hajíč, Otakar Smrž, Peter Zemánek, Jan Šnáidauf, Emanuel Beška. 2004. Prague Arabic Dependency Treebank: Development in Data and Tools. *Proceedings of NEMLAR-2004*, 110–117.

- Kevin Duh, and Katrin Kirchhoff. 2005. POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, June 2005, 55-62
- Khaled Elghamry, Rania Al-Sabbagh and Nagwa Elzeiny. 2008. Cue-Based Bootstrapping of Arabic Semantic Features. *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data*, March 2008, Lyon, France, 85-95
- Mona Diab. 2007. Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic, June 2007, page 89-96
- Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, Tagging and Base Phrase Chunking. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 22-23 April 2009, Cairo, Egypt, 285-288
- Mona Diab, Hacioglu, K., Jurafsky, D. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base-Phrase Chunks. In: S. Dumais, D.M., Roukos, S. (Eds.) *HLT-NAACL 2004: Short Papers*, pp. 149–152. Association for Computational Linguistics, Boston (2004)
- Mona Diab, Nizar Habash, Owen Rambow, Mohammed Al-Tantawy and Yassine Benajiba. 2010. COLABA: Arabic Dialect Annotation and Processing. *LREC Workshop on Semitic Language Processing*, Malta, May 2010
- Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool Publishers.
- Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. *Proceedings of COLING-ACL*, Sydney, Australia, 2006
- Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, 4 August 2009, 221-224
- Nizar Habash, Mona Diab and Owen Rambow. 2011. Conventional Orthography for Dialectal Arabic (CODA) Version 0.1 – July 2011. A Technical Report, Center for Computational Learning Systems – CCLS, Columbia University.
- Nizar Habash, Ramy Eskander and Abdelatti Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON 2012)*, pages 1-9, Montreal, Canada, June 7, 2012
- Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet Another Dialectal Arabic Corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May 2012, Istanbul, Turkey, 2882-2889
- Sarah Alkuhlani and Nizar Habash. 2012. Identifying Broken Plurals, Irregular Gender, and Rationality in Arabic Text. *EACL 2012*, 675-685
- Shabib AlGahtani, William Black and John McNaught. 2009. Arabic Part-Of-Speech Tagging Using Transformation-Based Learning. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009
- Shereen Khoja. 2001. APT: Arabic Part-of-Speech Tagger. *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania
- Tim Buckwalter. 2002. Arabic Morphological Analyzer (AraMorph). Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 & ISBN 1-58563-257-0
- Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic English Statistical Machine Translation. *Proceedings of the First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties*, Edinburgh, Scotland, 10–21

Appendix A: A Detailed Description and Comparison of Our Tagset with ATB and RTS Tagsets

POS	ATB	RTS	Ours	Tag	Comments
Abbreviation	✓	✓	✓	ABR	Examples include titles like <i>د./d/</i> (Dr.), <i>م./A/</i> (Mr.) and <i>ج.ر.ق /jm.E/</i> (Arabic Republic of Egypt). It is noteworthy that RTS includes abbreviations with the NN tag that is also used for singular common nouns.
Accusative	✓	—	—	—	No case marking neither for EA – which does not consider it – nor for MSA tweets – when found – is given.

Adjective	✓	✓	✓	JJ	simple adjectives like كويں /kwys/ (good), شغل /\$gA// (fine)
Adverb	✓	✓	✓	RB	
Case	✓	—	—	—	Case marking is not a feature of EA. Even for MSA tweets in our corpus that can result from users quoting press news and the like, case marking is ignored.
Command Verb	✓	✓	✓	VB	
Cardinal Number	✓	✓	✓	CD	
Comparative Adjective	✓	—	✓	JJR	
Connective Particle	✓	—	✓	CC	Both connective particles and coordinating conjunctions are collapsed into the CC tag. Both coordinate base and complex phrases. Examples of this tag include: و /w/ (and), وبعدین /wbEdyn/ (and then) and وكمان /wkmAn/ (and also).
Coordinating Conjunction	✓	✓			
Definite	✓	—	—	—	Definite particle is tokenized and tagged as DT. The information about whether a noun or an adjective is definite is thus structurally defined: in the tokenized corpus if a noun/adjective is preceded by the definite article, it is definite; otherwise it is not.
Demonstrative Pronoun	✓	✓	✓	DM	Demonstratives are phrase boundary markers and thus they are distinguished from determiners which are not.
Determiner	✓	✓	✓	DT	
Dialect	✓	—	—	—	Although the corpus contains MSA tweets coming mostly from users copying from press agencies, the dialect-standard distinction is not marked in the corpus.
Direct Object	✓	—	—	—	Object nouns are not tagged, but object pronouns are. They are split-off during tokenization and tagged as OBP. Distinction between direct and indirect pronoun objects is not marked in our tagset given that it is structurally – rather than – morphologically defined: indirect object pronouns are preceded by a preposition but direct ones are not.
Dual	✓	—	✓	DU	The number features – singular, dual and plural – are given separate tags that can be combined with any relevant content word tag such as nouns, verbs, adjectives and personal pronouns.
Emphatic Particle	✓	—	—	—	
Existential	—	—	✓	EX	Neither ATB nor RTS label existentials probably because they are expressed in MSA via the imperfect verb يوجد /ywjd/ or the demonstrative هناك /hnAk/ both meaning <i>there is/are</i> . In EA, there are two possible forms of existentials: قـ /fj/ and فـهـ /fjh/ (lit: in; gloss: there is/are). Given that both existentials are ambiguous with the preposition <i>in</i> which has the same

					forms, and that existentials can be phrase boundary anchors, we use the EX tag to label them.
Feminine	✓	—	✓	F	
Focus Particle	✓	—	—	—	
Foreign Word	✓	✓	✓	FW	Tweets in Arabic script that contain one or more foreign words have these words labeled as FW.
Foreign Script	✓	—	—	—	Tweets entirely in foreign script whether they contain words from a foreign language – like English words – or contain Arabic words are filtered out in corpus preprocessing.
Future	✓	—	—	FT	Future tense is marked in EA by the verb prefixes <i>ــا /h/</i> or <i>ــه /H/</i> (will). These two are split out in tokenization and are tagged as FT for future. The same tag is used then for the MSA separate future particle <i>سوف /swf/</i> . Although it does not mark phrase boundaries, it is important for verb tense identification.
Future Particle	✓	—	—	—	
Genitive	✓	—	—	—	
Imperfect Verb	✓	✓	✓	VBP	The tense of an imperfect verb can be present, future or progressive. This fine-grained tense classification is not represented by the tagset; yet this information is structurally predictable given that the split-off affixes indicating each tense are POS tagged.
Indefinite	✓	—	—	—	
Indicative	✓	—	—	—	
Interjection	✓	✓	✓	UH	
Interrogative Adverb	✓	✓	✓	INT	Interrogative adverbs like <i>فین /fin/</i> (how) and <i>ازای /zay/</i> (where) and interrogative pronouns like <i>ایه /Ayh/</i> (what) and <i>مین /myn/</i> (who) are all collapsed into the tag INT. Interrogative particles like <i>هل /h/</i> and <i>ی /y/</i> used in MSA to form yes/no questions are not used in EA, yet if they exist in MSA tweets, they are also labeled as INT.
Interrogative Particle	✓	—			
Interrogative Pronoun	✓	—			
Jussive Particle	✓	—	—	—	
Masculine	✓	—	✓	M	The gender features – masculine and feminine – are given separate tags that can be combined with any relevant content word tag such as nouns, verbs, adjectives and personal pronouns.
Mood	✓	—	—	—	
Negative Particle	✓	—	✓	NG	Negative particles come in two forms: a circumfix <i>mA...\$</i> as in <i>ما راحوش /mAraHw\$/</i> (they didn't go), and a number of free morphemes including <i>مش /m\$/</i> and <i>ي /y/</i> both meaning <i>no</i> among others. Bound and free negative

					particles are both labeled as NG; although the free-morpheme form does not mark phrase boundaries, it is an important marker for sentiment analysis – one goal for building this tagger.
Noun	✓	—	✓	NN	NN is used for common nouns whether singular, dual, plural or collective nouns. The gender and number features are given their own tags.
Noun Quantifiers	✓	—	✓	QNT	Quantifiers like <i>/\$wyp/</i> (a little), <i>/ktyr/</i> (a lot) among others are given their own tag – QNT.
Noun Suffix	✓	—	—	—	This is the same as the possessive pronoun – given the PP\$ tag. This is the only suffix that is split-off; whereas gender and number suffixes are not.
Nominative	✓	—	—	—	
Ordinal Number	✓	—	✓	OD	
Passive	✓	✓	✓	P	The passive feature is not reflected in the verb tagset for simplicity, yet a P tag combined with a verb tag indicates a passive voice and the absence of the P tag indicates an active voice – the default.
Particle	✓	✓	✓	PRT	All particles that (1) do not mark phrase boundaries and (2) do not mark verb tense or negation are collapsed in one tag – PRT.
Partial Word	✓	—	✓	PW	Given that each tweet is limited to 140 characters, users who go over the limit produce incomplete erroneous words. These count about 0.3% of our corpus.
Perfect Verb	✓	—	✓	VBD	The VBD tag in our tagset does not define voice (active vs. passive); it only defines the perfect aspect of the verb. Voice is given a separate tag – P.
Person	✓	—	✓	I 2 3	1 st , 2 nd and 3 rd person
Plural	✓	—	✓	PL	
Possessive Pronoun	✓	✓	✓	PP\$	
Preposition	✓	✓	✓	IN	
Progressive	—	—	✓	PG	One verb prefix that is specific to EA – in comparison to MSA – is the progressive prefix <i>→ /b/</i> as in <i>/byqwl/</i> (he's saying) and <i>/bnEAfr/</i> (we're struggling). The progressive prefix is split-off in tokenization and is tagged as PG.
Pronoun	✓	✓	✓	OBJP SBJP	Pronouns are split based on their grammatical functions into: object pronouns – typically tokenized from the verb endings – and subject pronouns – which are the free-morpheme subject pronouns here. Possessive pronouns are also given their own tag – PP\$.
Proper Noun	✓	✓	✓	NNP	NNP refers to proper nouns whether they are singular, dual or plural. Number features are tagged with a separate set of tags.

Pseudo Verb	✓	—	—	—	
Punctuation	✓	✓	✓	PNC	
Relative Adverb	✓	—	✓	RL	There is only one relative pronoun in EA – ﴿الّي /Allī/ (who, which, that). Even when relative pronouns/adverbs from MSA are encounter they are given the RL tag.
Relative Pronoun	✓	✓			
Response Conditional Particle	✓	—	✓	CN	Conditional particles are phrase boundary markers.
Restrictive Particle	✓	—	—	—	
Singular	✓	—	✓	SG	
Subjunctive	✓	—	—	—	
Subordinate Conjunction	✓	✓	✓	SC	Like CN, SC are phrase boundary markers.
Suffix	✓	—	—	—	Noun suffixes are the possessive pronouns only, given that our light stemming approach does not split off gender and number suffixes. Possessive pronouns are given the PP\$ tag and gender and number features are represented by their own tags.
Superlative Adjective	—	—	✓	JJS	Although superlative adjectives are not morphologically marked, they have implications for sentiment analysis and thus they are tagged as JJS.
Transcription Error	✓	—	—	—	
Typo	✓	—	—	—	
Verb	✓	—	✓	—	
Verbal Noun	✓	—	—	—	Verbal nouns or deverbal nouns are tagged as common nouns
Verbal Adjective	✓	—	—	—	Verbal adjectives are tagged as adjectives.
Vocative Particle	✓	—	✓	VC	
Not found words	—	✓	✓	NF	
Fixed Expressions	—	—	✓	EXP	
Modals	—	—	✓	MD	Fire-grained distinctions between different modality types – epistemic, evidential, deontic and volitive – are not marked in this tag. All linguistic modality types and their verbal or nominal realizations are labeled as MD.

Twitter Mentions	—	—	✓	MNT	
Twitter Hashtags	—	—	✓	HSH	
Twitter Emoticons	—	—	✓	EMO	
Twitter URLs	—	—	✓	URL	
Lengthened Words	—	—	✓	LNG	

NLP Workflow for On-line Definition Extraction from English and Slovene Text Corpora

Senja Pollak^{1,2}

Anže Vavpetič^{1,3}

Janez Kranjc^{1,3}

¹Jožef Stefan Institute

³International Postgraduate School

Jožef Stefan

Jamova 39, 1000 Ljubljana, Slovenia

senja.pollak@ijs.si

Nada Lavrač^{1,3,4}

Špela Vintar²

²Faculty of Arts, Aškerčeva 2, 1000

Ljubljana, Slovenia

⁴University of Nova Gorica, Vipavska 13,

5000 Nova Gorica, Slovenia

Abstract

Definition extraction is an emerging field of NLP research. This paper presents an innovative information extraction workflow aimed to extract definition candidates from domain-specific corpora, using morphosyntactic patterns, automatic terminology recognition and semantic tagging with wordnet senses. The workflow, implemented in a novel service-oriented workflow environment CloudFlows, was applied to the task of definition extraction from two corpora of academic papers in the domain of Computational Linguistics, one in Slovene and another in English. The definition extraction workflow is available online, therefore it can be reused for definition extraction from other corpora and is easily adaptable to other languages provided that the needed language specific workflow components were accessible as public services on the web.

1 Introduction

Extracting domain-specific knowledge from texts is a challenging research task, addressed by numerous researchers in the areas of natural language processing (NLP), information extraction and text mining. Definitions of specialized concepts/terms are an important source of knowledge and an invaluable part of dictionaries, thesauri, ontologies and lexica, therefore many approaches for their extraction have been proposed by NLP researchers. For instance, Navigli and Velardi (2010), Borg et al. (2010) and Westerhout (2010) have reported very good results with nearly fully

automated systems applied to English or Dutch texts. While most of the approaches follow the Aristotelian view of what constitutes a definition (*X is a Y which ...*), the concept of definition itself is rarely discussed in detail or given enough attention in the results interpretation. A popular way to circumvent the fuzziness of the “definition of definitions” is to label all non-ideal candidates as defining or knowledge-rich contexts to be validated by the user. In line with this philosophy, the definition extraction approach proposed in this work can be tuned in a way to ensure higher recall at a cost of lower precision.

Our work is mainly focused on Slovene, a Slavic language with a very complex morphology and less fixed word order, hence the approaches developed for English and other Germanic languages, based on very large - often web-crawled - text corpora, may not be easy to adapt. In general, definition extraction systems for Slavic languages perform much worse than comparable English systems (e.g., Przepiórkowski et al. (2007), Degórski et al. (2008a, 2008b), Kobyliński and Przepiórkowski (2008)). One of the reasons is that many Slavic languages, including Slovene, lack appropriate preprocessing tools, such as parsers and chunkers, needed for the implementation of well-performing definition extraction methods. Another obstacle is the fact that very large domain corpora are rarely readily available.

The main challenge addressed in this paper and the main motivation for this research is to develop a definition extraction methodology and a tool for extracting a set of candidate definition sentences from Slovene text corpora. This work follows our

work reported in Fišer et al. (2010), in which we have reported on the methodology and the experiments with definition extraction from a Slovene popular science corpus (consisting mostly from textbook texts). In addition to definition candidate extraction we used a classifier trained on Wikipedia definitions to help distinguishing between good and bad definition candidates. When analyzing the results we observed that the main reason for the mismatch between the classifier's accuracy on Wikipedia definitions versus those extracted from textbooks was the fact that, in authentic running texts of various specialized genres, definitions run an entire gamut of different forms, only rarely fully complying with the classical Aristotelian *per genus et differentiam* formula.

While this work inherits the basic methodology from Fišer et al. (2010), again focusing on Slovene definition extraction, incorporating the same basic methods of extracting definition candidates, this paper extends our previous work in many ways. First, the modules initially developed for Slovene have now been extended to enable the extraction of definition candidates also from English corpora. Second, the modules have been refined and implemented as web services, enabling their inspection and reuse by other NLP researchers. Next, the modules have been composed into an innovative definition extraction workflow. Moreover, this completely reimplemented approach has been evaluated on an different corpus, both regarding its genre and size. The corpus is from a very specific domain, which is a much more realistic scenario when developing specialized terminological dictionaries.

The developed workflow was applied to definition extraction from two corpora of academic papers in the area of Computational Linguistics, one in Slovene and another in English. The developed workflow has been implemented in our recently developed service-oriented workflow construction and management environment CrowdFlows¹ (Kranjc et al., 2012). The definition extraction workflow is available on-line², therefore it can be reused for definition extraction from other corpora and is easily adaptable to other lan-

guages provided that the needed language specific workflow components were accessible as public services on the web.

The paper is structured as follows. Section 2 summarizes the main definition extraction methods incorporated into the NLP definition extraction workflow, followed by the actual definition extraction workflow description in Section 3. Experimental evaluation of the workflow on the Slovene and English Computational Linguistics corpora is presented in Section 4. Section 5 concludes with a discussion, conclusions and plans for further work.

2 Summary of main definition extraction methods

Like in Fišer et al. (2010), we employ three basic methods to extract definition candidates from text. The approach postulates that a sentence is a definition candidate if one of the following conditions is satisfied:

- It conforms to a predefined lexico-syntactic pattern (e.g., NP [nominative] is_a NP [nominative]),
- It contains at least two domain-specific terms identified through automatic term recognition,
- It contains a wordnet term and its hypernym.

The first approach is the traditional pattern-based approach. In Fišer et al. (2010), we use a single, relatively non-restrictive *is_a* pattern, which yields useful candidates if applied to structured texts such as textbooks or encyclopaediae. However, if used on less structured authentic specialized texts, such as scientific papers or books used in the experiments described in this paper, a larger range of patterns yields better results. For the described experiments, we used eleven different patterns for Slovene and four different patterns for English.

The second approach is primarily tailored to extract knowledge-rich contexts as it focuses on sentences that contain at least n domain-specific single or multi-word terms. The term recognition module³ identifies potentially relevant termi-

¹<http://clowdflows.org>

²<http://clowdflows.org/workflow/76/>

³The term extraction methodology is described in detail in Vintar (2010).

nological phrases on the basis of predefined morphosyntactic patterns (Noun + Noun; Adjective + Noun, etc.). These noun phrases are then filtered according to a weighting measure that compares normalized relative frequencies of single words in a domain-specific corpus with those in a general reference corpus. As a reference corpus we used FidaPlus⁴ for Slovene and BNC⁵ for English. The largest coverage is achieved under the condition that the sentence contains at least two domain terms (term pair). Additional conditions are that the first term should be a multi-word term at the beginning of a sentence, and that there is a verb between a term pair (a detailed comparison of results obtained with different settings is beyond the scope of this paper).

The third approach exploits the *per genus et differentiam* characteristic of definitions and therefore seeks for sentences where a wordnet term occurs together with its direct hypernym. For Slovene, we use the recently developed sloWNet (Fišer and Sagot, 2008) which is considerably smaller than the Princeton WordNet (PWN) (Fellbaum, 1998) and suffers from low coverage of terms specific to our domain.

3 NLP workflow for on-line definition extraction

This section describes the NLP workflow, implemented in the CrowdFlows workflow construction and execution environment. We first present the underlying principles of workflow composition and execution. We then present a technical description of the CrowdFlows environment, followed by a detailed presentation of the individual steps of the definition extraction workflow.

3.1 Basics of workflow composition and execution

Data mining environments, which allow for workflow composition and execution, implemented using a visual programming paradigm, include Weka (Witten et al., 2011), Orange (Demšar et al., 2004), KNIME (Berthold et al., 2007) and Rapid-

⁴FidaPlus is a 619-million word reference corpus of Slovene (<http://www.fidaplus.net>).

⁵Mike Scotts wordlist from the BNC World corpus (<http://www.lexically.net/downloads/version4/downloading%20BNC.htm>) was used.

Miner (Mierswa et al., 2006). The most important common feature is the implementation of a workflow canvas where workflows can be constructed using simple drag, drop and connect operations on the available components. This feature makes the platforms suitable also for non-experts due to the representation of complex procedures as sequences of simple processing steps (workflow components named *widgets*).

In order to allow distributed processing, a service-oriented architecture has been employed in platforms such as Orange4WS (Podpečan et al., 2012) and Taverna (Hull et al., 2006). Utilization of web services as processing components enables parallelization, remote execution, and high availability by default. A service-oriented architecture supports not only distributed processing but also distributed workflow development.

Sharing of workflows has previously been implemented at the myExperiment website of Taverna (Hull et al., 2006). It allows users to publicly upload their workflows so that they become available to a wider audience and a link may be published in a research paper. However, the users who wish to view or execute these workflows are still required to install specific software in which the workflows were designed.

The CrowdFlows platform (Kranjc et al., 2012) implements the described features with a distinct advantage. CrowdFlows requires no installation and can be run on any device with an internet connection, using any modern web browser. CrowdFlows is implemented as a cloud-based application that takes the processing load from the client's machine and moves it to remote servers where experiments can be run with or without user supervision.

3.2 The CrowdFlows environment illustrated by a simplified NLP workflow

CrowdFlows consists of the workflow editor (the graphical user interface, as shown in Figure 1) and the server-side application, which handles the execution of the workflows and hosts a number of publicly available workflows.

The workflow editor consists of a workflow canvas and a widget repository, where widgets represent embedded chunks of software code, representing downloadable stand-alone applica-

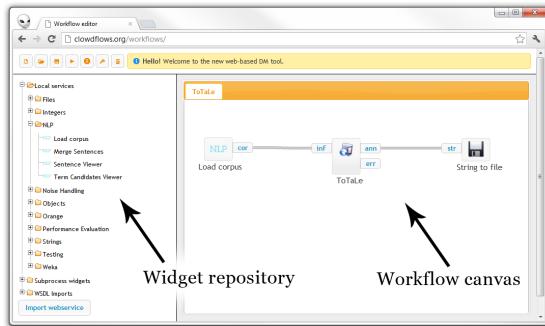


Figure 1: A screenshot of the workflow editor in the Google Chrome browser.

tions which look and act like traditional applications but are implemented using web technologies and can therefore be easily embedded into third party software. All our NLP processing modules were implemented as such widgets, and their repository is shown in the menu at the left-hand side of the CloudFlows canvas in the widget repository. The repository also includes a wide range of default widgets. The widgets are separated into categories for easier browsing and selection.

By using CloudFlows we were able to make our workflow public, so that anyone can execute it. The workflow is simply exposed by a unique address which can be accessed from any modern web browser. Whenever the user opens a public workflow, a copy of the workflow appears in her private workflow repository in CloudFlows. The user may execute the workflow and view its results or expand it by adding or removing widgets.

3.3 A detailed description of the definition extraction workflow and its components

The entire definition extraction workflow implemented in CloudFlows is shown in Figure 2.

The widgets implementing the existing software components include:

- ToTaLe tokenization, morphosyntactic annotation and lemmatization tool (Erjavec et al., 2010) for Slovene and English⁶.
- LUIZ term recognition tool (Vintar, 2010) for Slovene and English, with a new imple-

⁶In future versions of the workflow, we plan to replace the ToTaLe web service with ToTrTaLe which handles also ancient Slovene and produces XML output.

mentation of scoring and ranking of term candidates.

The core definition extraction widgets include:

- Pattern-based definition extractor,
- Term recognition-based definition extractor,
- WordNet- and sloWNet-based definition extractor.

Numerous other new auxiliary text processing and file manipulation widgets were developed and incorporated to enable a seamless workflow execution. These include:

- Load corpus widget, which allows the user to conveniently upload her corpus in various formats (PDF, txt, doc, docx) either as single files or several files together in one flat ZIP file,
- Term candidate viewer widget, which formats and displays the terms (and their scores) returned by the term extractor widget (a subset of the extracted term candidates is illustrated in Figure 3),
- Sentence merger widget, which allows the user to join (through intersection or union) the results of several definition extraction methods,
- Definition candidate viewer widget, which, similarly to the term candidate viewer widget, formats and displays the candidate definition sentences returned by the corresponding methods (Figure 4 illustrates the widget's output, listing the extracted definition candidates to be inspected by the user).

4 Experimental evaluation on the Language Technologies corpus

This section describes the corpus, the experimental results achieved and the quantitative and qualitative evaluation of results.

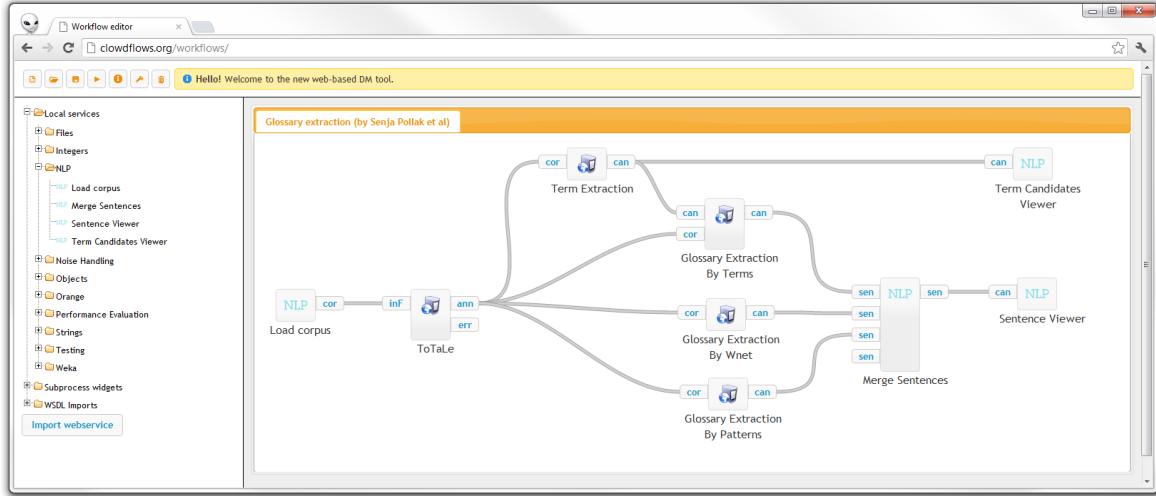


Figure 2: A screenshot of the entire definition extraction workflow.

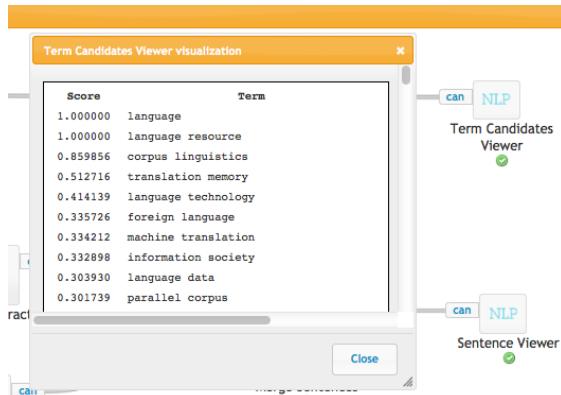


Figure 3: Viewing of selected term candidates.

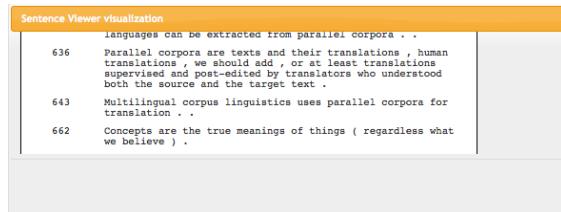


Figure 4: Viewing of selected definition candidates.

4.1 The corpus

For highly specialized domains and for languages other than English, the web may not provide an ideal corpus, especially not for the purpose of terminology extraction where a certain level of representativeness and domain coverage is crucial. Our corpus consists of papers published in the proceedings of the biennial Language Technolo-

gies conference (Jezikovne tehnologije) that has been organized in Slovenia since 1998. The articles are in Slovene or English. To improve vocabulary coverage we added other text types from the same domain, including Bachelors, Masters and PhD theses, as well as several book chapters and Wikipedia articles.

The total size of the corpus is 44,750 sentences (903,621 tokens) for Slovene and 43,019 sentences (929,445 tokens) for English.

4.2 Experimental results

In this section we evaluate the term extraction (see Subsection 4.2.1) and the glossary extraction method (in Subsection 4.2.2). More attention is paid to the latter, where not only quantitative results are provided, but we also analyze and discuss the results from the linguistic perspective.

4.2.1 Term extraction results

We evaluated top 200 (single- or multi-word) domain terms for each language (see Table 1). Each term was assigned a score of 1-5, where 1 means that the extracted candidate is not a term (e.g., *table*) and 5 that it is a fully lexicalized domain-specific term designating a specialized concept (e.g., *machine translation*). The scores between 2 and 4 are used to mark varying levels of domain-specificity on the one hand (e.g., *evaluation* is a term, but not specific for this domain; score 3), and of phraseological stability on the other (e.g., *translation production* is a termi-

nological collocation, not fully lexicalized, compositional in meaning; score 3).

Precision	English terms	Slovene terms
Yes (2-5)	0.775	0.845
Yes (5)	0.48	0.55

Table 1: Precision of term extraction method.

The second part of term evaluation involved the assessment of recall. The domain expert annotated a random text sample of the Slovene and English corpus with all terminological expressions (approximately 65 for each language), and the samples were then compared to the lists of terms extracted by the LUIZ system. Table 2 shows the results for both samples using either all term candidates or just the top 10,000/5,000.

	Number of terms	Recall
Slovene	38,523	0.694
	10,000	0.527
	5,000	0.444
English	25,007	0.779
	10,000	0.644
	5,000	0.491

Table 2: Recall of terminological candidates extracted from the Slovene and English Language Technologies corpus.

4.2.2 Definition extraction results

The results of definition extraction methods on the Language Technologies corpus are presented in Table 3, showing the number of candidates extracted with each individual method, as well as the number of candidates obtained with the intersection of at least two methods (*Intersect*) and those extracted by at least one of the three methods (*Union*). The latter shows that by using all the methods of the NLP definition extraction workflow we extracted 4,424 definition candidates for English and 6,638 for Slovene.

The reason for extracting a larger candidate set for Slovene compared to English is that the Slovene corpus is larger in the number of sentences, that the pattern-based approach is more elaborate (containing 11 patterns compared to only 4 patterns for English), and that the number of extracted Slovene terms is larger.

Number of candidates	English def. candidates	Slovene def. candidates
Patterns	474	1,176
Terms	866	1,539
Wordnet	3,278	4,415
Union	4,424	6,638
Intersect	192	472

Table 3: Definition candidates extracted from the Language Technologies corpus.

Precision	English def. candidates	Slovene def. candidates
Patterns	0.44	0.26
Terms	0.08	0.15
Wordnet	0.13	0.05
Union	0.09	0.11
Intersect	0.33	0.25

Table 4: Precision of definition extraction methods.

From a set of extracted definition candidates, obtained as outputs of each of the methods, 100 sentences were randomly selected and used for the evaluation of the precision of our workflow (see Table 4).

Precision is better for English than for Slovene. Concerning the patterns, the reason can be in less fixed word order in Slovene, while for the wordnet-based method we observed that the selected wordnet pairs were too general and that many domain specific terms were not found in sloWNNet.

To evaluate the recall of our methods, we randomly selected 1,000 sentences for each language. In the Slovene data set there were 21 definitions out of which 10 were extracted by at least one of our methods (0.4762 recall). The English 1,000 sentences random corpus contained 25 definitions, out of which 15 were extracted (0.6 recall). We plan to perform further evaluation to get the results on a larger test set.

To gain a better insight into the types of definition candidates, we reassessed each method and analyzed their output. It is clear from these results that simple patterns still procure best results, while the union of different methods yields a lot of potentially interesting candidates, but much more noise.

When analyzing the evaluation sets we observed that a definition in real text is often not easy to define and evaluate. A lot of sentences in running text can be considered as borderline cases, often without the hypernym and defining the term either through its extension or its purpose; see the examples (i) and (ii) that have not been identified by any method, but can be considered as definitions.

- (i) *Z aktivno kamero lahko torej “s pogledom sledimo” obrazu govorca, kadar se ta premika. [With an active camera we can “eyetrack” the face of the speaker when he is moving.]*
- (ii) *Osembitni kodni nabor ISO 8859-2 je na mestih s kodami od 0 do 127 identičen standardu ISO 646, na preostalih 128 mestih pa kodira vse potrebne značke za pisanje v albanščini, češčini, [...] in slovenščini. [The 8-bit ISO 8859-2 code-page is identical to ISO 646 at codes ranging from 0 to 127, while it uses the other 128 codes to encode the characters used for writing Albanian, Czech [...] and Slovene.]*

The analysis of candidate sentences shows that the notion of definition, especially when we attempt to formalize it, needs to be reconsidered. Different definition types found in the evaluation set include: *formal definitions with genus and differentia structure* (X is Y), whereby the definendum does not necessarily occur at the beginning of the sentence; definitions with genus and differentia structure, where the verb is other than the verb “biti” [to be]; sentences where a term is not defined through its hypernym, but through a *sibling concept and the differentia*; *informal definitions*, subordinated in a sentence and introduced with a relative pronoun; *extensional definitions*, i.e. definition which instead of specifying the hypernym lists all the possible realizations of a concept (X includes Y, Z and Q); *defining by purpose* (hypernym is omitted); *definition as textual formula* used for mathematical concepts.

5 Discussion, conclusions and further work

One of the contributions of this paper is the improvement and an in-depth experimental assessment of the individual methods constituting our NLP definition extraction workflow. The other main contribution is the implementation of the

definition extraction workflow, which has been made publicly available within a novel service-oriented workflow composition and management platform CloudFlows. The contributions and plans for further work are discussed in more detail below.

Based on the qualitative analysis of our methods, we identified a number of definition types not traditionally covered by definition extraction systems. Based on these findings we started to improve our methodology in several ways. Even if compared to previous experiments the patterns were already extended from strict *is_a* pattern to a larger set of patterns, the approach could be further extended to cover all the alternatives listed above (e.g., extensional definitions). The pattern-based method had the highest precision, but should be extended with other methods to ensure better coverage (e.g., by the candidates at the intersection of wordnet- and term-based methods). Concerning the term-based extraction method we are conducting further experiments, based on the threshold for the termhood parameter setting and the additional restriction that the terms identified should be in the nominative case (for Slovene). Finally, the wordnet method was improved by limiting sloWNet nouns in Slovene to nominative case only and using a different setting of the window parameter. Regarding the evaluation of recall, the experiments on a larger test set are being performed.

Concerning the new NLP workflow implementation of our definition extraction modules based on morphosyntactic patterns, automatic terminology recognition and semantic tagging with WordNet/sloWNet senses, its on-line availability and modularity are a great advantage compared to the existing NLP software, including other terminology and definition extraction tools. The workflow implementation within the novel CloudFlows workflow composition and execution engine enables workflow reuse for definition extraction from other corpora, experiment reproducibility, as well as the ease of workflow refinement by the incorporation of new NLP modules implemented as web services and workflow extensions to other languages.

In future work we plan to refine the definition extraction components to improve the precision

and the recall and to develop new workflow components for on-line natural language processing.

Acknowledgments

We are grateful to Darja Fišer for past joint work on definition extraction from Slovene texts. The presented work was partially supported by the Slovene Research Agency and the FP7 European Commission project “Large scale information extraction and integration infrastructure for supporting financial decision making” (FIRST, grant agreement 257928).

References

- Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Ktter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel. 2007. KNIME: The Konstanz Information Miner. In *GfKl. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 319-326.
- Claudia Borg, Mike Rosner and Gordon J. Pace. 2010. Automatic grammar rule extraction and ranking for definitions. In *Proceedings of the Seventh International Conference on Language Resources and Evaluations*, LREC 2010, Valletta, Malta, 2577-2584.
- Łukasz Degórski, Michał Marcinczuk and Adam Przeźiórkowski. 2008a. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco, 837-841.
- Łukasz Degórski, Łukasz Kobyliński and Adam Przeźiórkowski. 2008b. Definition extraction: Improving balanced random forests. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 353-357.
- Janez Demšar, Blaž Zupan, Gregor Leban and Tomaž Curk. 2004. Orange: From experimental machine learning to interactive data mining. In *Proceedings of ECML/PKDD-2004*, LNCS Volume 3202, 537-539.
- Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In *Proceedings of the 7th International Conference on Language Resources and Evaluations*, LREC 2010, Valletta, Malta, 1806-1809.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. Online version: <http://wordnet.princeton.edu>
- Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue* (LNCS 2546). Berlin; Heidelberg: Springer, 61-68.
- Darja Fišer, Senja Pollak and Špela Vintar. 2010. Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC’10), 2932-2936.
- Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Matthew R. Pocock, Peter Li and Thomas M. Oinn. 2006. Taverna: A tool for building and running workflows of services. *Nucleic Acids Research* (Web-Server-Issue) 34: 729-732.
- Łukasz Kobyliński and Adam Przeźiórkowski. 2008. Definition extraction with balanced random forests. In *Proceedings of GoTAL’2008*, 237-247.
- Janez Kranjc, Vid Podpečan and Nada Lavrač. 2012. CloudFlows: A cloud-based scientific workflow platform. In *Proceedings of ECML/PKDD-2012*. Springer LNCS (in press).
- Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz and Timm Euler. 2006. YALE: rapid prototyping for complex data mining tasks. In *Proceedings of KDD-2006*, ACM, 935-940.
- Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010). Uppsala, Sweden, 1318-1327.
- Vid Podpečan, Monika Zemenova and Nada Lavrač. 2012. Orange4WS environment for service-oriented data mining. *The Computer Journal*, 55(1): 82-98.
- Adam Przeźiórkowski, Łukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň and Beata Wójtowicz. 2007. Towards the automatic extraction of definitions in Slavic. In *Proceedings of the Balto-Slavonic NLP workshop at ACL 2007*, Prague, 43-50.
- Špela Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16(2): 141-158.
- Ian H. Witten, Eibe Frank and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Third Edition. Morgan Kaufmann.
- Eline Westerhout. 2010. *Definition Extraction for Glossary Creation: A study on extracting definitions for semi-automatic glossary creation in Dutch*. Netherlands Graduate School of Linguistics / Landelijke - LOT Dissertation Series.

Projecting Semantic Roles via Tai Mappings

Hector-Hugo Franco-Peña
Trinity College Dublin
Dublin, Ireland.
francoph@scss.tcd.ie

Martin Emms
Trinity College Dublin
Dublin, Ireland.
mtemms@scss.tcd.ie

Abstract

This work takes the paradigm of projecting annotations within labelled data into unlabelled data, via a mapping, and applies it to Semantic Role Labelling. The projections are amongst dependency trees and the mappings are the Tai-mappings that underlie the well known tree edit-distance algorithm. The system was evaluated in seven different languages. A number of variants are explored relating to the amount of information attended to in aligning nodes, whether the scoring is distance-based or similarity-based, and the relative ease with which nodes can be ignored. We find that all of these have statistically significant impacts on the outcomes, mostly in language-independent ways, but sometimes language dependently.

1 Introduction

There are a number of pattern recognition scenarios that have the characteristics that one has some kind of *structured* test data (sequence, tree, graph, grid) within which some annotation is missing, and one wants to infer the missing annotation by exploiting fully annotated training data. A possible approach is to seek to define *alignments* between training and test cases, and to use these to *project* annotation from the training to test cases. This has been successfully used in computational biology, for example, to project annotation via sequence alignments (Marchler-Bauer et al., 2002) and graph alignments (Kolar et al., 2008). Semantic Role Labeling (SRL) can be seen as a further instance of this pattern recognition scenario

and we will describe in this paper an SRL system which works by projecting annotations over an alignment. This paper extends results reported by Franco-Peña (2010). In the remainder of this section we give a brief overview of SRL. Section 2 then describes our system, followed in sections 3 and 4 by discussion of its evaluation.

For a number of languages, to an existing syntactic treebank, a layer of *semantic role* information has been added: the evolution of the Penn Treebank into PropBank is an example (Palmer et al., 2005). Role-label inventories and annotation principles vary widely, but our system has been applied to data annotated along the same lines as exemplified by PropBank. The example below illustrates a PropBank role-labelling.¹

[Revenue]_{A1} edged [up]_{A5} [3.4 %]_{A2} [to \$904 million]_{A4} [from \$874 million]_{A3} [in last year's third quarter]_{TMP}

A lexical item (such as *edge*), is given a frameset of enumerated core argument roles (A0 ... A5). In the example, A3 is the start point of the movement, and a minimal PropBank commitment is that A3 and the other enumerated role identifiers are used consistently across different tokens of *edge*. Across different lexical items, commitments concerning continuity in the use of the enumerated arguments are harder to state – see the conclusions in section 5. There are also named roles (such as TMP above), whose use across different items is intended to be consistent.

¹To save space, this shows simply a labelling of subsequences, omitting syntactic information. Figure 2 shows annotation added to a syntactic structure.

Arising from the CoNLL-2009 SRL evaluation shared task (Hajič et al., 2009), for seven languages (**Catalan**, **Chinese**, **Czech**, **English**, **German**, **Japanese**, **Spanish**), there is data consisting of a role-annotation layer added to syntactic information. The syntactic information is expressed as dependency trees. In some cases this is derived from a primary constituent-structure representation (eg. English), and in other cases it is the 'native' representation (eg. Czech). For each language, tree nodes have four kinds of syntactic information: FORM: a word form; LEMMA: lemma of the word form; POS: part-of-speech tag (tag sets are language specific); DEPREL: the dependency relation to its head word (the relation-sets are language specific). Additionally in each tree, \mathcal{T} , a number of nodes are identified as predicate nodes. Each predicate node $p^{\mathcal{T}}$, is linked to a set of argument nodes, $\text{args}(p^{\mathcal{T}})$. For each $a_i^{\mathcal{T}} \in \text{args}(p^{\mathcal{T}})$, the link $(p^{\mathcal{T}}, a_i^{\mathcal{T}})$ is labelled with a role. The role-labeling follows the PropBank approach.

2 An Alignment-based SRL system

Sub-tree extraction A preliminary to the role-labelling process itself is to extract a sub-tree that is relevant to a given predicate and its arguments. Where p is a particular predicate node of a tree \mathcal{T} , let $\text{sub_tree}(\mathcal{T}, p)$ stand for the relevant sub-tree of \mathcal{T} . There is considerable latitude in how to define this, and we define it in a simple way, via the least upper bound, lub , of p and $\text{args}(p)$: $\text{sub_tree}(\mathcal{T}, p)$ includes all nodes on paths down from lub to p and $\text{args}(p)$. This is the same subtree notion as used by Moschitti et al. (2008). Figure 1 illustrates. Henceforth all trees will be assumed to have been extracted in this way.

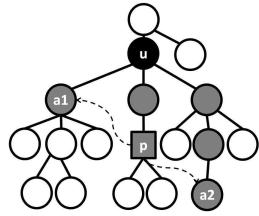


Figure 1: Sub-tree extraction: $a_1, a_2 = \text{args}(p)$, $u = \text{lub}(a_1, a_2, p)$.

Alignment and Projection Let an *alignment* of trees S and T be a 1-to-1, partial mapping $\alpha :$

$S \mapsto T$, from the nodes of S into the nodes of T . If S is role-labelled, such an alignment projects a role for test tree argument a_i^T if it is aligned to a training tree argument, a_j^S , and the predicate nodes p^S and p^T are aligned: the role of a_j^S is projected to a_i^T . Such a role-projecting tree will be termed '*usable for* T '. Fig. 2 shows an example alignment between subtrees, with the aligned sub-trees shown in the context of the trees S and T from which they come. Argument nodes \mathcal{T}_4 , \mathcal{T}_6 and \mathcal{T}_7 would receive projected labels A_1 , A_2 , and A_3 from \mathcal{S}_7 , \mathcal{S}_{12} and \mathcal{S}_{13} . The first two are correct, whilst \mathcal{T}_7 's annotation should be A_4 .

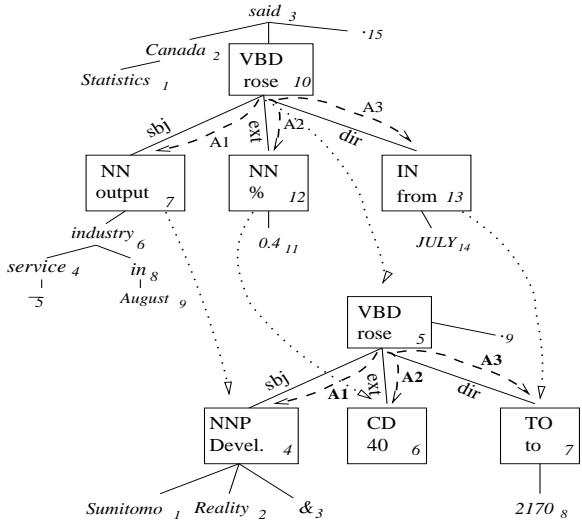


Figure 2: An example alignment.

Algorithm outline Let $[\Delta]_{d_i}^i$ be an equivalence class, contains all training samples at distance d_i to T . The training set can be thought of as a sequence of equivalence classes $[\Delta]_{d_1}^1, [\Delta]_{d_2}^2, \dots$, for increasing distance.² The algorithm works with a PANEL of nearest neighbours, which is always a prefix of this sequence of equivalence classes. Where T is defined by predicate p and arguments $a_1 \dots a_n$, the algorithm to predict a label for each a_i is

1. make sorting of training trees and set PANEL of nearest neighbours to be first equivalence class $[\Delta]_{d_1}^1$
2. (i) make a set of predictions from the usable members of PANEL (ii) if there is a most fre-

²Or alternatively a sequence of similarity equivalence classes, at decreasing similarities to T .

quent prediction, return it (iii) if there is a tie, or no usable members, add next equivalence class to PANEL if possible and go to (i), else return 'unclassified'

Tai mappings In this work, alignments are restricted to be so-called *Tai* mappings (Tai, 1979): amongst all possible 1-to-1, partial mappings from S into T , $\alpha : S \mapsto T$, these are mappings which respect *left-to-right order* and *ancestry*.³ Then to select a *preferred* alignment, a score is assigned to it. The definitions relevant to this which are given below follow closely those of Emms and Franco-Peña (2012).

Because a mapping α is partial on S and *into* T , there is a set $\mathcal{D} \subseteq S$ ('deletions'), of those $i \in S$ which are not mapped *to* anything and a set $\mathcal{I} \subseteq T$ ('insertions'), of those $j \in T$ which are not mapped *from* anything, and the alignment scorings make reference to the sets α , \mathcal{D} and \mathcal{I} .

We consider both a 'distance' scoring, $\Delta(\alpha : S \mapsto T)$, whose *minimum* value is used to select the alignment, and a 'similarity' scoring $\Theta(\alpha : S \mapsto T)$, whose *maximum* value is used. The 'distance' scoring, $\Delta(\alpha : S \mapsto T)$, is given by

$$\sum_{(i,j) \in \alpha} C^\Delta(i^\gamma, j^\gamma) + \sum_{i \in \mathcal{D}} C^\Delta(i^\gamma, \lambda) + \sum_{j \in \mathcal{I}} C^\Delta(\lambda, j^\gamma)$$

Here $(.)^\gamma$ gives the label of a node, and a function C^Δ is used to give label-dependent costs to the members of α , \mathcal{D} and \mathcal{I} . This is the major parameter of the distance scoring and the various setting for it which were considered are detailed below; at a general level, to accord minimally with intuition, it should always be the case that $C^\Delta(x, y) \geq 0$, and $C^\Delta(x, y) \geq C^\Delta(x, x)$ for non-identical x and y . A 'similarity' scoring, $\Theta(\alpha : S \mapsto T)$, is given by

$$\sum_{(i,j) \in \alpha} C^\Theta(i^\gamma, j^\gamma) - \sum_{i \in \mathcal{D}} C^\Theta(i^\gamma, \lambda) - \sum_{j \in \mathcal{I}} C^\Theta(\lambda, j^\gamma)$$

where C^Θ is a function defining costs for the members of α , \mathcal{D} and \mathcal{I} . To accord minimally

³More precisely, where $anc(x, y)$ and $left(x, y)$ denote the 'ancestor of' and 'to the left of' relations, $\forall (i, j) \in \alpha, \forall (i', j') \in \alpha$, the mapping must satisfy (i) $left(i, i')$ iff $left(j, j')$ and (ii) $anc(i, i')$ iff $anc(j, j')$

with intuition, $C^\Theta(x, y) \leq C^\Theta(x, x)$. Additionally, in this work, we also assume that $C^\Theta(x, \lambda) = C^\Theta(\lambda, x) = 0$.

Besides ranking alternative alignments between fixed S and T , the minimum distance alignment score defines a 'distance', $\Delta(S, T)$, for the pair (S, T) , and maximum similarity alignment scores define a 'similarity', $\Theta(S, T)$, and these are used to rank alternative neighbours for a tree to be labelled. The algorithm to calculate $\Delta(S, T)$ and $\Theta(S, T)$ follows very closely that of Zhang and Shasha (1989): although originally proposed in the context of 'distance' and minimisation, it is straightforwardly adaptable to the context of 'similarity' and maximisation.

Cost settings On this data-set the label is in general a 4-tuple (p, d, l, f) of part-of-speech, dependency-relation, lemma, and word form. Four settings for the swap costs, $C^\Delta(x, y)$, are considered: **B**('binary'), **T**('ternary'), **H**('hamming') and **FT**('frame ternary'), based on the matches/mis-matches on these features. For any given feature a , let a^δ represent match/mismatch on a , with 1 for mis-match and 0 for match. The different swap settings are then defined as below

	$C^\Delta(x, y)$	values
B	$p^\delta \times d^\delta$	0, 1
T	$\frac{1}{2}[p^\delta + d^\delta]$	0, $\frac{1}{2}$, 1
H	$\frac{1}{4}[p^\delta + d^\delta + l^\delta + f^\delta]$	0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, 1
FT	$\frac{1}{2}[p^\delta + d^\delta] + fr^\delta$	0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2

For FT, fr refers to a synthesised attribute: for predicate nodes, fr is the frame identifier, and otherwise $fr = __$. The effect is that $fr^\delta = 1$ if one node is a predicate and the other is not, or if both are predicates but from different frames. Besides these swap settings, the deletion cost $C^\Delta(x, \lambda) = C^\Delta(\lambda, x)$ was varied between 1 and 0.5. From each swap setting $C^\Delta(x, y)$, two 'similarity' settings were derived by $C^\Theta(x, y) = \delta - C^\Delta(x, y)$, for $\delta = 1$ or 2.

Some aspects of these choices are based on results concerning distance and similarity which we previously established (Emms and Franco-Peña, 2012), where we showed that the conversion

$$\begin{aligned} C^\Theta(x, y) &= 2\kappa - C^\Delta(x, y) \\ C^\Theta(x, \lambda) &= C^\Delta(x, \lambda) - \kappa \\ C^\Theta(\lambda, y) &= C^\Delta(\lambda, y) - \kappa \end{aligned}$$

converts C^Δ to C^Θ such that the same ordering is induced over the alternative alignments for any given pair (S, T) . Call such settings **A-duals**. For a given choice from B/T/H/FT, the four above-mentioned settings for C^Δ and C^Θ thus stand in the following A-duality relationships:

distance	A-dual similarity
(a) $C^\Delta(x, \lambda) = 1$	(c) $C^\Theta(x, y) = 2 - C^\Delta(x, y)$
(b) $C^\Delta(x, \lambda) = 0.5$	(d) $C^\Theta(x, y) = 1 - C^\Delta(x, y)$

For such dual settings, the labelling potential of a given training sample is necessarily *identical* under the two settings. However, this A-duality property is distinct, potentially, from **N-duality**, which is the property that the two settings induce the same ranking of candidate neighbours $\{S_1 \dots S_N\}$ to a given T . Because the role-labelling system is driven both by neighbour and alignment ordering, it is an empirical question whether or not A-dual settings will deliver the same outcomes.

We will refer to settings (a) and (b) as 'dist(del=1)' and 'dist(del=0.5)', and settings (c) and (d) as 'sim(2-swap)' and 'sim(1-swap)'.

3 Experiment Procedure

The seven languages of the CoNLL-2009 SRL task were used (Hajič et al., 2009), with the same division into training sets and evaluation sets.

Due to our computational limitations the English training data set was limited to the first 20,000 sentences and the Czech training data set was limited to the first 10,000 sentences.

A simple labeling accuracy score is reported. When a two-way contrast is considered, the significance of a difference in outcomes under the two settings was determined by the McNemar test (McNemar, 1947), at levels of significance $p < 0.05$ and $p < 0.001$, as do Fürstenau and Lapata (2011) in their role-labelling work.

4 Results

4.1 Contrasting Swap Settings

Figure 3 shows a graph with the accuracy for the seven languages on the evaluation data set, using the dist(del=1) setting.⁴ The languages are sorted by the accuracy of the H setting. The baseline

⁴The full table of values, including the 3 out-of-domain cases, appears as the first column in Table 4.

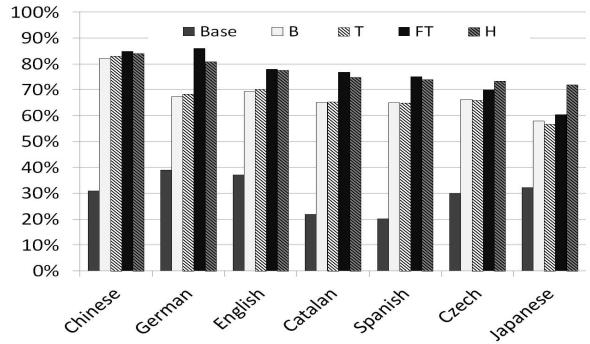


Figure 3: Accuracy of Tree edit distance across the seven languages.

(first bar) corresponds to the accuracy that would be attained by always choosing the label that is most frequent for that language.

Looking first at the H setting, the performance across the languages ranges from 73%(Czech) to 84%(Chinese), which in all cases substantially out-performs the majority-class base line (which is in the range 20(Spanish)–40%(German)). The other cost-settings also clearly out-perform this base line. It also seems that the variation in accuracies across the languages is not relatable to the variation in this majority-class base line.

The performance under the B and T settings is perhaps surprising, when one recalls that these two settings pay no attention at all to lexical differences between nodes, and refer only to the dependency relations and the part of speech. Nonetheless, even in this setting of tolerance to node exchange, it seems that in concert with the structure-respecting requirements of Tai mappings, surprisingly high accuracies can be obtained.

Chinese has the highest overall accuracy, with even the very simplest settings reaching relatively high accuracy. It is interesting to note that amongst the highest performing other systems which have used the same data, performance on the Chinese data-set has tended to be *below* that of other languages, with Che et al. (2009) and Dai et al. (2009) reporting over 81 F1 in all languages except Chinese where the F1 reported was 76.38 and 76.23. It suggests that tree distance and labelling by alignment methods may be especially suitable for the Chinese data set.

Japanese reports the worst results, especially

for the B/T/FT settings. This is probably due to the fact that, on inspection, the Japanese data gives 96.1% of syntactic dependencies the same dependency relation, practically canceling the contribution of the DEPREL feature.

The outcomes for the Spanish and Catalan data sets are very similar to each other. This is not unexpected as for the most part one is a translation of the other and they were annotated in the same way with the same set of labels.

Table 1 summarises the outcomes of pair-wise comparisons of the swap settings. Under '1st > 2nd', $!l, *m, **n$ appears if there were l data-sets on which the 1st setting out-performed the 2nd setting, for m of these the outcomes were significantly different ($p = 0.05$), and for n of these the difference holds at a stricter significance level ($p = 0.001$).

Settings	1st > 2nd	2nd > 1st	avge 2nd-1st
B-T	!4, *2, **1	!6, *3, **3	0.341%
B-FT	!0, *0, **0	!10, *10, **10	7.45%
B-H	!0, *0, **0	!10, *10, **9	7.779%
T-FT	!0, *0, **0	!10, *10, **10	7.109%
T-H	!0, *0, **0	!10, *9, **9	7.438%
FT-H	!7, *6, **4	!3, *3, **3	0.329%

Table 1: Comparing swap-settings, for $\text{dist}(\text{del}=1)$, on the 7 evaluation data-sets, and 3 out-of-domain data-sets.

The T and B settings turn out to give rather similar outcomes. Table 1 shows that T's margin over B averages out to 0.341%. There are 3 strictly significant cases where T out-performs B, and 1 cases in the other direction.

In its turn the H setting always out-performs the T setting, 9 times out of 10 at the strictest significance level, with the average margin being 7.438%. Thus penalising lexical difference seems always to improve performance, and nearly always substantially, though for the Chinese and out-of-domain English data sets, the margin for H over T falls to less than 1%.

The outcomes with FT are more language dependent. FT out-performs H more often ($7!6*$) than H out-performs FT ($3!3*$) and English is the one data-set on which the two are not significantly different. Japanese shows the highest margin in favour of H (11.5%) whilst German shows the highest margin in favour of FT (5.2%). The poor

relative performance of FT for Japanese is again probably a function of the uninformative nature of its dependency annotation.

For all languages FT out-performs T, at the strictest significance level, with the margin averaging out to 7.1%. For German the margin is especially large at 17.8%.

4.2 Contrasting Representations

Table 2 compares the results obtained with trees to results obtained with a *linearised* version, using just a node sequence corresponding to the sequences of words that spans the predicate and argument nodes. Encoding these as linear trees, the tree-matching in this case reduces to the standard Levenshtein matching.

Settings	Tree > Lev.	Lev > Tree	avge Lev-Tree
B	*8, **6	*0, **0	-2.83%
T	*10, **10	*0, **0	-4.574%
FT	*10, **10	*0, **0	-7.758%
H	*10, **10	*0, **0	-8.113%

Table 2: Comparing Tree and Levenshtein outcomes.

As is evident, for all languages and all swap-settings, the alignment on the linear representation gives substantially poorer results than the alignment on the trees. For T/FT/H in each single experiment the tree version produce better score than the linear version, and in B it was never detected a statistical advantage of the linear version over the tree version. This indicates that the Tai-mapping constraints on the tree-representation definitely leverage information that is beneficial to this labeling task.

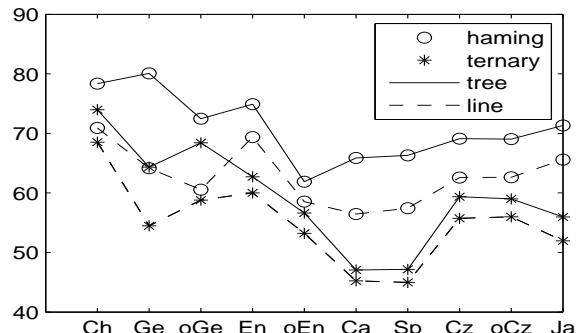


Figure 4: Tree vs Levenshtein distance (T & H).

Setting	max(dist)	max(sim)	difference
Chinese	B **	82.82%	83.46%!
	T =	83.98%!	83.86%
	FT *	85.88%!	85.53%
	H **	85.06%	85.88%!
German	B *	67.41%!	65.74%
	T *	68.25%!	65.74%
	FT **	87.71%!	85.01%
	H =	81.38%!	80.82%
o-German	B *	69.51%!	67.59%
	T *	69.1%!	67.09%
	FT *	75.46%!	73.03%
	H =	72.7%	73.53%!
English	B **	69.34%	70.21%!
	T =	70.79%!	70.48%
	FT =	78.75%!	78.49%
	H *	78.72%	79.43%!
o-English	B =	63.08%	63.67%!
	T =	66.26%!	65.14%
	FT =	69.83%!	69.51%
	H =	66.68%	67.52%!
Catalan	B **	65.7%!	64.11%
	T **	65.7%!	63.88%
	FT **	79.35%!	77.75%
	H =	75.8%!	75.58%
Spanish	B **	65.5%!	64.23%
	T **	65.7%!	63.98%
	FT **	77.25%!	76.2%
	H =	75.16%	75.64%!
Czech	B =	66.28%!	66.21%
	T **	66.69%!	66%
	FT *	70.05%!	69.53%
	H **	74.05%	74.77%!
o-Czech	B *	64.42%	64.92%!
	T =	65.08%!	64.73%
	FT **	69.24%!	68.17%
	H *	72.56%	73.36%!
Japanese	B =	57.93%!	57.31%
	T **	57.97%!	55.86%
	FT **	61.52%!	58.94%
	H **	75.34%!	71.32%
total	B	*4, **2	*3, **2
	T	*6, **4	*0, **0
	FT	*8, **5	*0, **0
	H	*1, **1	*4, **2
	all	*19, **12	*7, **4

Table 3: max(dist) vs max(sim): max(dist) is best of dist(del=1) and dist(del=0.5), max(sim) is best of sim(swap=2) and sim(swap=1).

4.3 Contrasting Distance and Similarity

Table 3 compares the best results of distance and similarity. In the case of distance, the best value is between dist(del=1) and dist(del=0.5), and in the case of similarity, the best value is between sim(2-swap) and sim(1-swap). Figure 5 plots the this contrast for the T and FT swap settings.

For the T overall the tree distance performs better than tree similarity. For T, this occurs 4 times

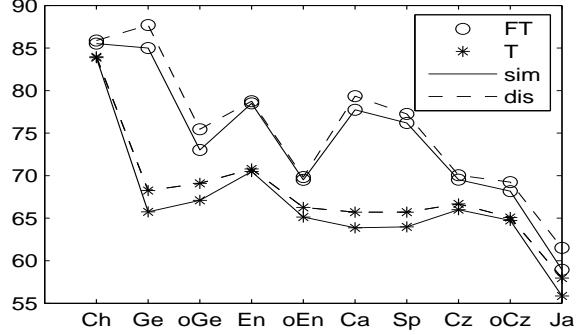


Figure 5: Distance outcomes compared to Similarity outcomes (T and FT).

at the strictest significance level, 6 times at the less strict level, there is no language where similarity outperforms distance, and the average margin is 1.28%. For B, this trend is less clear, with just 2 cases where distance out-performed similarity at the strictest significance level. The margins are small and average out at 0.45% in favour of the distance setting.

For FT, this occurs 5 times at the strictest significance level, 8 times at the less strict level, there is no language where similarity outperforms distance, and the average margin is 1.29%. German, Catalan, Spanish and Japanese show the largest margin, whilst Chinese, English and Czech the smallest.

For the H setting, overall the margin between distance and similarity is small, 0.04%. Similarity out-performs distance 2 times at the strictest significance level and 4 times at the less strict level. Japanese is unusual, being the only case where the comparison is statistically significantly in favour of *distance*, by a margin of 4.03%.

Recall from section 2 that the distance and similarity settings can be paired off as alignment-duals, namely 'dist(del=1)' with 'sim(2-swap)' and 'dist(del=0.5)' with 'sim(1-swap)'. For such dual settings the labelling potential of a given training sample is necessarily *identical* under the two settings. We noted that this does not theoretically guarantee identical system outcomes: A-duality does not imply N-duality, that is identical *neighbour-ordering*. The experiments actually show that it is indeed not the case for these data-sets that the A-dual settings are also N-dual settings. Space precludes giving details of this,

but it is implied by Table 3 and Fig. 5: for accuracies a_1 and a_2 attained by the distance alternatives, if N-duality held, the dual similarity settings would also attain accuracies a_1 and a_2 , and we would observe no differences between maximum distance outcomes and maximum similarity outcomes.

4.4 Contrasting two Distance settings

Cost setting	del=1	del=.5	difference
Chinese	B **	82.14%	82.82%!
	T **	83%	83.98%!
	FT **	84.78%	85.88%!
	H **	83.93%	85.06%!
German	B *	67.41%!	65.83%
	T =	68.25%!	67.41%
	FT *	86.03%	87.71%!
	H =	80.82%	81.38%!
o-German	B *	69.51%!	67.76%
	T =	69.1%!	68.84%
	FT =	74.62%	75.46%!
	H =	72.53%	72.7%!
English	B =	69.34%!	69.18%
	T **	70.12%	70.79%!
	FT **	77.98%	78.75%!
	H **	77.57%	78.72%!
o-English	B =	63.08%!	63.01%
	T =	65.77%	66.26%!
	FT *	69.83%!	68.39%
	H *	66.68%!	64.9%
Catalan	B **	65.12%	65.7%!
	T *	65.23%	65.7%!
	FT **	76.77%	79.35%!
	H **	74.76%	75.8%!
Spanish	B **	64.97%	65.5%!
	T **	64.89%	65.7%!
	FT **	75.04%	77.25%!
	H **	73.95%	75.16%!
Czech	B **	66.28%!	65.47%
	T **	65.84%	66.69%!
	FT =	69.99%	70.05%!
	H **	73.28%	74.05%!
o-Czech	B **	64.42%!	63.21%
	T =	64.73%	65.08%!
	FT **	69.24%!	68.19%
	H =	72.56%!	72.23%
Japanese	B =	57.93%!	57.4%
	T *	56.69%	57.97%!
	FT *	60.43%	61.52%!
	H **	71.92%	75.34%!
total	B	*4, **2	*3, **3
	T	*0, **0	*6, **4
	FT	*2, **1	*6, **4
	H	*1, **0	*6, **6

Table 4: Tree distance with del=1 vs del=0.5.

Table 4 contrasts the two tree-distance settings⁵, one where the deletion cost is 1 and an-

⁵The higher of the two values is used in table 3 for

other where the deletion cost is 0.5.

The main observation is that there is a tendency for the del=0.5 setting to out-perform del=1. The differences are small, usually less than one per cent. Figure 6 plots the outcomes for T and FT, and Figure 7 shows H outcomes.

For B, 3 times del=0.5 was strictly statistically better than del=1, but the converse was also the case 2 times, with margin averaging out at 0.44% in favour of del=1. For T, del=0.5 often out-performed del=1, and significantly so (*6, **4), whilst del=1 never significantly out-performed del=0.5. Averaged over all the data-sets, there is a small margin for del=0.5: 0.48%.

For FT, again del=0.5 often out-performed del=1, and significantly so (*6, **4), with the margins being a little larger than the case for T. For two of the out-of-domain data-sets (o-Cz and o-En), the relationship reverses, with del=1 statistically significantly out-performing the del=0.5 setting. Averaging there is a margin in favour of del=0.5 of 0.78%, with largest margin shown by Spanish (2.21%) and Catalan (2.57%). See Figure 6.

For H, the effect of switching from del=0.5 to del=1 follows a very similar pattern to that found for FT and where del=0.5 exceeds del=1, the effect is a little more pronounced than it was for FT, with 6 cases strictly statistically significant. Again for o-Cz, the relationship is reversed, and for the other two out-of-domain data-sets either del=1 is significantly better or statistically indistinguishable. Averaged across all the data-sets the margin is 0.73% in favour of del=0.5 for H.

Concerning the out-of-domain data-sets, one could speculate that the reason why the del=0.5 setting does not improve over the del=1 setting is that this makes swaps relatively more costly, and that the out-of-domain data-sets require a greater tolerance to swaps.

5 Conclusions and future work

In an approach to projecting annotations over Tai-mappings, we have explored the effects of variations of possible parameters. We have considered a number of settings concerning the costing of swaps, referring to greater and lesser amounts of

max(dist).

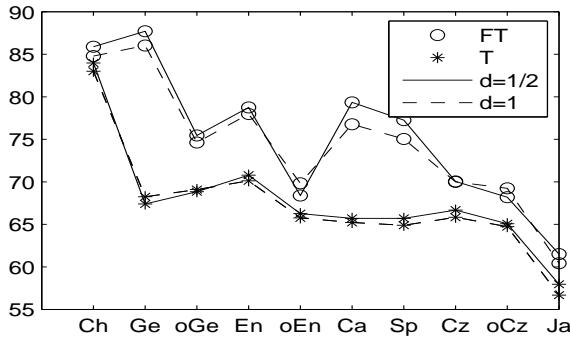


Figure 6: Comparing $\text{dist}(\text{del}=1)$ to $\text{dist}(\text{del}=0.5)$ (T and FT).

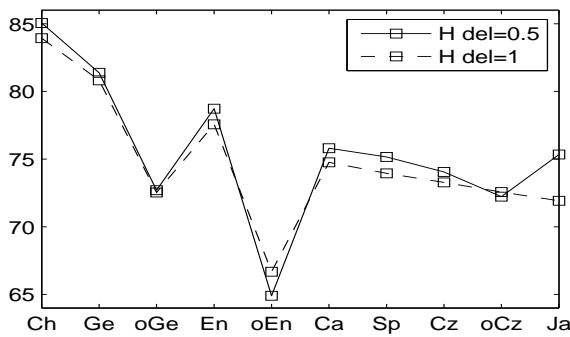


Figure 7: Comparing $\text{dist}(\text{del}=1)$ to $\text{dist}(\text{del}=0.5)$ (H).

the available information. With the H setting, using syntactic and lexical information, the performance across the languages ranges from 74% to 85%, and the performance under the B and T settings, which attend to no lexical information was also substantial. The linearised representation was shown to be clearly out-performed by the tree representation. Distance- and similarity-based alignment scoring were shown to give different outcomes, with distance overall out-performing similarity on the B, T and FT settings, but with this no longer the case for the H setting.

There were also some language-specific findings, amongst them that Japanese results with B/T/FT settings were noticeably poor, almost certainly due to that data-set's lack of variation in dependency labels, and that whilst mostly lowering the cost of deletion improved performance, this was not the case for the out-of-domain data-sets.

The non-negligible performance under the B and T settings, which attend to no lexical information is perhaps worth further scrutiny. If across

different lexical items, PropBank aimed for no continuity of use of enumerated core arguments, intuition would suggest that in these B and T settings, performance should be very low. Concerning the A0 and A1 roles, PropBank commits to some consistency concerning possession of prototypically agent and patient properties, and for other enumerated roles, to consistency within certain verb groups. One direction for future research would be to investigate what aspects of performance are due to consistent use across lexical items, by automatically introducing inconsistency by permutations amongst the identifiers for a particular item.

There has been little comparable work using an alignment approach to SRL an exception being Fürstenau and Lapata (2011), though there are significant differences: they work with FrameNet (Fillmore et al., 2004), and use an alignment not constrained by ancestry or linear order. Also, rather than taking each unlabelled item, T , in turn, and using its nearest labelled neighbours, $\{S_1, \dots, S_n\}$, their aim is to take each labelled exemplar, S , from a frame *lexicon*, in turn, and use it to project annotation to its nearest neighbours $\{T_1, \dots, T_n\}$ in an unlabelled corpus. For all of these reasons, we cannot at the moment make any meaningful quantitative comparisons with their work. Nonetheless, it seems reasonable to expect the contrasts we have described concerning cost settings and distance versus similarity to apply to the kind of data-set expansion scenario they discuss and investigating whether this is so is a potential avenue for further research. Conversely it would be interesting to see how our findings are effected if we were to replace our notion of alignments which must be Tai mappings, with the notion of alignment from their work.

Acknowledgements

All calculations were performed on the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing. This cluster was funded through grants from Science Foundation Ireland.

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

References

- Wanxiang Che, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin, and Ting Liu. 2009. Multi-lingual Dependency-based Syntactic and Semantic Parsing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 49–54, Boulder, Colorado, June. Association for Computational Linguistics.
- Qifeng Dai, Enhong Chen, and Liu Shi. 2009. An Iterative Approach for Joint Dependency Parsing and Semantic Role Labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 19–24, Boulder, Colorado, June. Association for Computational Linguistics.
- Martin Emms and Hector-Hugo Franco-Penya. 2012. On order equivalences between distance and similarity measures on sequences and trees. In *Proceedings of ICPRAM 2012 International Conference on Pattern Recognition Application and Methods*.
- C.J. Fillmore, J.Ruppenhofer, and C.F. Baker, 2004. *Frontiers in Linguistics*, chapter Framenet and representing the link between semantic and syntactic relations.
- Hector-Hugo Franco-Penya. 2010. Edit Tree Distance alignments for Semantic Role Labelling. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 79–84, Uppsala, Sweden. Association for Computational Linguistics.
- Hagen Fürstenau and Mirella Lapata. 2011. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(June 2011):135—171.
- Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Adam Meyers, Joakim Nivre, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*.
- Michal Kolar, Michael Lassig, and Johannes Berg. 2008. From protein interactions to functional annotation: graph alignment in Herpes. *BMC Systems Biology*, 2(1).
- Aron Marchler-Bauer, Anna R Panchenko, Benjamin A Shoemaker, Paul A Thiessen, Lewis Y Geer, and Stephen H Bryant. 2002. Cdd: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, 30(1):281–3.
- Q McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.
- Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Kuo-Chung Tai. 1979. The tree-to-tree correction problem. *Journal of the ACM (JACM)*, 26(3):433.
- Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing*, 18:1245–1262.

Automatic Identification of Motion Verbs in WordNet and FrameNet

Parvin Sadat Feizabadi and Sebastian Padó

Institut für Computerlinguistik

Universität Heidelberg

Im Neuenheimer Feld 325

69120 Heidelberg

feizabadi, pado@cl.uni-heidelberg.de

Abstract

This paper discusses the automatic identification of motion verbs. The context is the recovery of unrealized location roles from discourse context or “locational inference”, a special case of missing argument recovery. We first report on a small corpus study on verb classes for which location roles are particularly relevant. This includes motion, orientation and position verbs. Then, we discuss the automatic recognition of these verbs on the basis of WordNet and FrameNet. For FrameNet, we obtain results up to 67% F-Score.

1 Introduction

A central problem in natural language processing (NLP) is that a considerable portion of the meaning of texts is typically not expressed overtly. It must be provided by the reader through inference, as described, e.g., by Norvig (1987): *An inference is defined to be any assertion which the reader comes to believe to be true as a result of reading the text, but which was not previously believed by the reader, and was not stated explicitly in the text.* The omission of assertions is by no means a sign of erroneous or "sloppy" language use. In fact, pragmatic considerations, notably the maxim of quantity (Grice, 1975) encourage speakers not to express all information.

Omission of material is thus a fundamental fact of language. Its importance for NLP can be appreciated in the framework of textual entailment inference, a semantic meta-task which is defined as the decision, given two short texts T (Text) and H (Hypothesis), whether *readers of T would agree that*

H is (almost) certainly true (Dagan et al., 2009). The reasoning requirements of many NLP tasks can be phrased as textual entailment queries. Examples include answer validation in QA (Peñas et al., 2008); IE (Romano et al., 2006); and multi-document summarization (Harabagiu et al., 2007).

Within textual entailment, methods which do not try to recover missing assertions in T often fail to correctly derive the existence of an entailment relation when H realizes an assertion that T does not. Recent work has verified that this routinely happens for sentences in discourse, as the following example shows (Mirkin et al., 2010):

- (1) **Context:** *China* seeks solutions for its coal mine safety problem. [...]

T: A recent accident has cost more than a dozen miners their lives.

H: An accident *in China* has killed several miners.

The Hypothesis **H** in Example (1) can be inferred almost completely from the Text **T**, with the exception of the prepositional phrase "in China", which is not realized locally in the Text. It can however be inferred from the context sentence.

This paper presents a pilot study in the context of *locational inference*, i.e., the recovery of unexpressed assertions that concerns the spatial configuration of events. Determining location information is an important subpart of inference and plays an important part in Information Extraction (Leidner et al., 2003) as well as Question Answering (Greenwood, 2004) and has been accorded an important role in the analysis of narratives (Herman, 2001; Howald, 2010).

Our ultimate goal is to develop computational methods that can automatically retrieve omitted locations. The present paper takes the first step by defining the task, performing a focused data analysis, and evaluating the ability of the two most widely used resources in computational linguistics, WordNet and FrameNet, for the purpose of identifying motion verbs, for which locational inference is particularly relevant.

Plan of the paper. Section 2 defines the task of locational inference and discusses its challenges. Section 3 presents a corpus annotation study which provides ground truth of motion verbs and location roles as well as an analysis. Section 4 evaluates how well WordNet and FrameNet can be used to tackle the identification of motion verbs.

2 Locational Inference and Motion Verbs

2.1 Locational Inference

Locational inference is the special case of the recovery of unrealized arguments or null instantiations for semantic roles that denote places and directions. Figure 1 shows our concept of a processing pipeline for location inference. The task consists of several subtasks. The first subtask (I.) is the recognition of verbs that actually require locational inference. This subtask again decomposes into two parts: first we verify that the verb in question requires a location (1), and then we check that the current instance of the lemma leaves at least one location unrealized so that it must be inferred (2). The second subtask, recovery (II.), then attempts to identify the missing location from the available locations in context (3., 4.). The modeling part of this paper focuses on the first recognition step, that is, the decision of whether an event requires a location (Step 1).

2.2 Related Work

This task shows similarities to some existing NLP processing paradigms. The most notable is semantic role labeling or SRL (Gildea and Jurafsky, 2002), in the sense that locations can be seen as specific semantic roles. Semantic roles have also been employed as a basis for deciding entailment (Burchardt et al., 2009). The division of locational inference into recognition and recov-

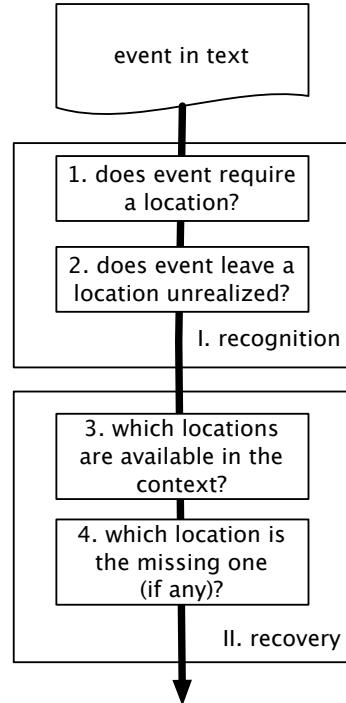


Figure 1: Processing pipeline for locational inference

ery is also reminiscent of the frequent decomposition of SRL into recognition and classification steps. Traditionally, SRL concentrates on locally realized arguments, but awareness is growing that SRL needs to include non-local arguments. Gerber and Chai (2010) found many implicit arguments of nouns to be recoverable from prior context, and a recent SemEval task (Ruppenhofer et al., 2010) explicitly included unrealized arguments.

A second related task is coreference resolution (Soon et al., 2001). The analogy applies in particular to the second part of locational inference (recovery). Similar to coreference resolution, we have to construct a set of contextually available "antecedents" for missing locations and pick the correct one. However, here face "zero anaphora" (Fillmore, 1986), which again makes it difficult to identify features. Silberer and Frank (2012), address the general problem but find it to be very difficult. By focusing on location roles, we hope to simplify the problem through limiting the semantic types of possible antecedents.

2.3 Locational Inference and Motion Verbs

A fundamental question about locational inference is what verbs it applies to. While it can be argued

that (almost) all events are take place somewhere, our intuition was that for many events in a text it is difficult to pin down where exactly they happen. This is true for many states as well as events. Consider this example:

- (2) Although the concept **was** a simple one, Allan **thought** it had potential.

Here, the discourse context does not provide any hint as to the location of the marked verbs. At the same time, the location is not centrally important to the event either; therefore inferring it does not appear crucially important.

The situation is substantially different for some verb classes; notably verbs of motion (including self motion and caused motion), verbs of orientation, and verbs of position. For verbs in these classes, locations play an integral role in their semantics, as can be argued either on the basis of decomposition (Jackendoff, 1990) or of corpus evidence (Baker et al., 1998).

Based on these observations, we decided to focus on these three classes of verbs which require a location in this paper: motion, orientation, and position verbs (henceforth called motion verbs for the sake of simplicity). Thus, *see* and *ask* are not included in the scope of our study, but *arrive* and *sit* are. By taking this stance, we do not mean to imply that locational inference is irrelevant for non-motion verbs. We merely take the decision to focus on these comparatively "clear" cases for the purpose of this pilot study.

2.4 A FrameNet-based Characterization of the Motion Domain

There is no readily available comprehensive definition of motion, orientation, and position verbs in any existing computational resource. Part of the problem are unclear boundaries between motion verbs and change-of-state verbs. For example, in Levin's verb classification (Levin, 1993), class 10 (*Verbs of Removing*) appear to be a good candidate to contain (caused) motion verbs; however, the description of subclass 10.3 (*Clear verbs*) notes that "some of [the verbs'] uses are better characterized as verbs of change of state".

However, in a small pilot experiment we found annotators' intuitions on what constitutes a motion verb to deviate too much for reliable anno-

tation. We decided to ground our notion of motion (plus orientation and position) verbs based on FrameNet. FrameNet (Baker et al., 1998) is a semantic dictionary of English, based on the concept of semantic frames from Fillmore's frame semantics (Fillmore, 1985). Each frame describes a prototypical situation and comes with a set of semantic roles describing the participants and objects of that situation. It also lists a set of lexical units, word senses that can introduce the frame. In this way, frames group verbs into semantic classes. FrameNet includes about 10,000 lexical units, and 800 semantic frames.

More specifically, we started out from FrameNet's MOTION frame which assumes five location roles:

- SOURCE: "If A is the source of a motion event with mover B, then B is at place A at the beginning of the event".
- GOAL: "If A is the goal of a motion event with mover B, then B is at place A at the end of the event".
- PATH: "If A is the path of a motion event with mover B, then B is at place A at sometime during the movement event, but neither at the beginning of the event nor at the end."
- PLACE: "Place identifies the setting in which the Theme's movement takes place without a specified Path. "
- DIRECTION: "If A is the direction of a motion event from source B to goal C, then A is [...] a straight line from B to C."

These five roles appear to be the maximum set of location roles that motion verbs support in English. We then identified all FrameNet frames with at least one of these roles, manually excluding about 10 frames which used the same role names for non-locations (e.g., SOURCE in the case of AUTHORITY as in *militarySRC power*). This step yielded a set of 34 frames, shown in Table 1. A list of the verbs from these frames, together with the lists of rolesets FrameNet provides for the each verb (possibly more than one), formed the basis of our corpus study.

ADORNING	ARRIVING
CAUSE_FLUIDIC_MOTION	CAUSE_MOTION
COTHHEME	DEPARTING
EMANATING	EMITTING
EMPTYING	EVENT
FILLING	FLUIDIC_MOTION
GIVING	IMPORT_EXPORT
LIGHT_MOVEMENT	MASS_MOTION
MOTION	MOTION_DIRECTIONAL
MOTION_NOISE	OPERATE_VEHICLE
PATH_SHAPE	PLACING
POSTURE	PRECIPITATION
QUITTING	RECEIVING
REMOVING	RESIDENCE
RIDE_VEHICLE	SELF_MOTION
SENDING	SMUGGLING
TAKING	TRAVEL

Table 1: List of Motion frames in FrameNet

3 A Corpus Study of Motion Verbs and Location Roles

3.1 Annotation

The goal of the annotation is to produce a ground truth of human’s understanding of location information from text. Annotators were asked to identify all motion verbs in the corpus and annotate their location roles.

For our annotation, we chose the 4000-word short story "The Black Willow" from the "fiction" subset of MASC, the Manually Annotated Subcorpus of the American National Corpus (Ide et al., 2008). We decided on a fiction text since fiction is less stylistically standardized than newswire text and we therefore expect to find more natural narrative structures.

To perform the annotation, we used the MMAX2 Annotation Tool (Müller and Strube, 2006). It is a graphical user interface for creating, browsing, visualizing and querying linguistic annotations on multiple levels. We annotated information on raw text, with no linguistic analysis other than sentence segmentation.

Annotators were provided with the list of motion verbs according to FrameNet (cf. Section 2.4), with the understanding that this list was possibly incomplete. It was complemented by detailed annotation guidelines¹. For verbs, annotators speci-

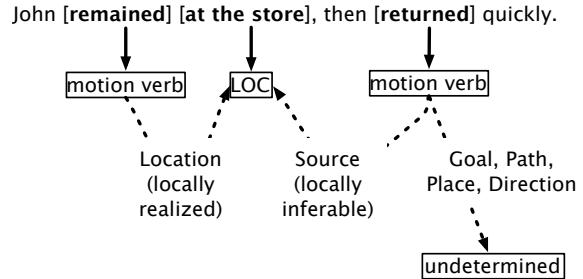


Figure 2: Annotation example

fied whether the verb was in the FrameNet list, and if it was, whether the instance could be covered by any of the role sets offered by FrameNet.

Each location was tagged with one of four possible statuses: *locally realized*, *locally inferable*, *globally inferable*, *world knowledge* and *undetermined*. The first option is for roles realized explicitly in the verb’s predicate-argument structure. The second and third option describe roles which are inferable from context – “locally inferable” is for locations within a three-sentence window (previous, current, and next sentence), and “globally inferable” is for locations that can be found outside this window. “World knowledge” covers cases where an unrealized role can be completed based on world knowledge triggered by the predicate-argument structure (see below for examples). The last option, “undetermined”, indicates that the related location role remains completely undetermined: it is not mentioned explicitly anywhere in the text and cannot be determined based on world knowledge either.

Figure 2 illustrates most of these categories. It shows a sentence with one position verb and one motion verb and one location. The position verb is annotated with just one location role (LOCATION) which is locally realized. The motion verb, in contrast, is tagged with five location roles none of which are locally realized. The SOURCE role is locally inferable from the first verb, and the others are all undetermined, due to the lack of discourse context in the example.

3.2 Annotation Reliability

The annotation was performed by two annotators with very good proficiency in English and graduate-level background in linguistics. The agreement of the two annotators, computed as ex-

¹The guidelines are available from <http://www.nlpado.de/~sebastian/data.shtml>.

Agreement on motion verbs	94%
Agreement on PLACE role	60%
Agreement on SOURCE role	64%
Agreement on GOAL role	73%
Agreement on PATH role	74%
Agreement on DIRECTION role	68%

Table 2: Annotation reliability (exact match)

act match, is shown in Table 2. Agreement on motion verbs, which this paper focuses on, is very good. Agreement on roles is lower. We believe that this is at least in part due to our detailed annotation scheme (cf. the status types described in Section 3.1). This interpretation calls for further investigation in the future. After computing agreement, the two annotators resolved annotation differences to produce a joint gold standard.

3.3 Analysis of the Annotation

Among the 4000 words of "The Black Willow", we found 731 verbs, of which 208 belonged to the motion domain and were thus annotated with information regarding their location(s). Of these 208, 143 (69%) were in the FrameNet-derived list, but 65 were not. Furthermore, There were 9 instances among the 143 where FrameNet did not offer the correct role set (i.e., the correct sense).

The numbers indicate that FrameNet is a solid, but not complete, resource to guide annotation of motion verbs. Annotators cannot be told to rely on the completeness of the lists that they are provided. Examples of missing coverage include cases where the motion verb is not listed in FrameNet, although the correct frame is available (e.g., *lean back*) as well as cases of missing frames. This latter case includes, for example, the frame IMPACT which lists several motion verbs but was not selected because it uses frame-specific roles (*Impacto*, *Impactee*) rather than directly recognizable motion roles.

We also identified a total of 190 realized location phrases. Motion verbs and locations were linked by a total of 268 roles. Thus, each verb is linked to an average of 1.3 realized location roles, either local or inferred, and many of the locations are referred to by more than one event. Given the theoretical maximum of 683 location roles that the 208 verbs could realize according to their FrameNet rolesets, some 40% of the roles

were realized somewhere in the discourse.

The 268 roles decompose into 188 locally realized roles (i.e., within the same predicate-argument structure), 43 locally inferrable ones (in the direct context), and 37 globally inferrable ones (in the complete document context). These numbers indicate that there is indeed considerable need for locational inference: only 188 out of 683 roles are realized locally. Unfortunately, only another 80 can be recovered by finding their "antecedents" in the discourse using the surface-based methods to have in mind (cf. Section 2.1).

Thus, we find that even for motion verbs, a considerable number of location roles remains undetermined even taking the complete document into account. An analysis by location role found PATH to be the role with the highest ratio (74%), and PLACE the one with the lowest ratio (39%). GOAL (56%), SOURCE (54%), and DIRECTION (54%) are located between the extremes. We investigated the background of these fairly high numbers and found the main reason to be that undetermined location roles are often inferable not from the discourse context, but from the meaning of the predicate itself or from other roles. This happens in the example "they kicked up dust [PATH along the roadway]" where the PATH role supports the inference that the PLACE of the kicking event is the roadway, too. In "he left [SOURCE the room]", the predicate triggers the inference that the DESTINATION is simply "some place outside the room". Finally, in "the sun had reached [GOAL the trees]", the sky understood as SOURCE. This inference requires world knowledge about the sun as well as the relative position of the sky and the trees.

4 Automatic Identification of Motion Verb Instances

The first requirement for automating locational inference is an automatic means to identify motion verbs in running text (cf. Step 1 in Figure 1). Given that the definition of the motion domain is essentially a semantic one, the second contribution of this paper is a set of experiments with the goal of recognizing motion verbs in text. We follow a knowledge-based approach, using two standard CL resources, WordNet and FrameNet.

FrameNet was already introduced in Section 2.4. WordNet is an electronic lexical database, in which

English nouns, verbs, adjectives, and adverbs are organized into synonym sets (synsets) at the sense level (Miller et al., 1990). WordNet includes more than 150,000 words grouped into 117,000 synsets, among which there are over 11,000 verbs.

The main challenge in identifying motion verbs for locational inference is word sense disambiguation: Many verbs have motion as well as non-motion senses. For example, *to cross* can be a motion verb (*The ship crossed the ocean*), in which case it would be subject to locational inference, but it can also mean a gesture (*Peter crossed himself*) or a trickery event (*John was crossed by the con man*) – in the last two cases we currently assume that no locational inference takes place. A potential additional complication arises from the fact that the motion domain is a classical source domain for metaphorical mappings (Lakoff and Johnson, 1980). That is, many motion verbs are used to express metaphorical motion, orientation, or position (*Colin moves towards the Labour position / is on the Tory side*). Fortunately, a preliminary corpus analysis indicated that metaphorical usages of motion verbs behave similar to literal usages with regard to locations and locational inference. That is, even though it might be ultimately desirable to distinguish literal and metaphorical motion verbs in order to avoid unwanted inferences, such as *the Labour position* is a physical place, we avoid this distinction in the current study.

Thus, the concrete motion verb prediction task is as follows: Given the information in one of the two resources, we (a) define the set of motion verbs in terms of the resources; (b), we disambiguate the verb instances in the text with the resource; (c) we predict an instance to be a motion verb if it is included in the set from (a).

4.1 Preprocessing

Given that our manual annotation is based on FrameNet, it seems natural to use FrameNet frames for disambiguation. However, the development of standalone FrameNet-based WSD is made difficult by the incomplete coverage of word senses by FrameNet (Erk, 2005) as a consequence of which many instances should be left unassigned. The results of our own annotation confirm this observation (cf. Section 3.3). We therefore performed Word Sense Disambiguation on the basis

of WordNet 3.0 using UKB, a state-of-the-art unsupervised graph-based word sense disambiguation tool (Agirre and Soroa, 2009). Since UKB requires part-of-speech information, so also performed part-of-speech tagging with TreeTagger (Schmid, 1994).

We also evaluated the quality of the preprocessing components. First, we found that TreeTagger only recognized 89% of the gold standard motion verbs as verbs. Most of the missing cases were phrasal verbs which the tagger could not handle correctly; this leads to an upper bound of 89% recall for any model building on the TreeTagger output. Then, we annotated the verbs with WordNet synsets to evaluate UKB. When compared against the gold standard, its exact match accuracy is almost exactly 50%, comparable to the verb results reported by Agirre and Soroa on the Senseval all-words datasets. However, accuracy improves to 73% if we evaluate just the coarse-grained decision motion verbs vs. non-motion verbs.

4.2 Disambiguation Strategies

The numbers reported in the previous subsection leave open the question of whether WSD is actually good enough to serve for the selection of motion verbs in our application.

To explore this question, we will compare three strategies for Word Sense Disambiguation. The first strategy is no disambiguation at all (NoWSD). It classifies a verb instance as a motion verb if any sense of the lemma is in the resource-derived list of motion verbs (cf. above). The second strategy, WSD, classifies an instance as a motion verb if its UKB-assigned synset is in the list of motion verbs. The third strategy, PredomSense, leverages the observation that WSD has a hard time beating the *predominant sense heuristic* (McCarthy et al., 2004) which assigns the predominant, or first, sense to all instances. Here, this strategy means that we treat all verbs as motion verbs whose first sense, according to WordNet, is in the list of motion verbs.

4.3 Motion verbs in FrameNet

Defining motion verbs based on frames. This approach builds directly on the intuition developed in Section 2.4: we characterize the motion domain through a set of motion frames recognizable by the

	Precision	Recall	F ₁ -Score
NoWSD	43	87	58
WSD	68	48	56
PredomSense	73	62	67

Table 3: Motion verb recognition results with FrameNet

use of location roles. The resulting list was shown in Table 1. We created an initial list of motion verbs by listing all verbal lexical units for these frames. However, as discussed in Section 3.3, this list is far from complete. A second problem is that the WSD labels assigned by UKB are WordNet synsets. Both problems can be solved together by mapping map the FrameNet lexical units onto WordNet. We used the mapping constructed by Shi and Mihalcea (2005) to determine the matching synsets for each lexical unit². In order to improve coverage, we added all hyponyms of these synsets. This corresponds to the assumption that any hyponyms of a lexical unit l can evoke the same frame as l . The resulting set comprises 2838 synsets.

Results. Table 3 shows the results for the three strategies defined in Section 4.2. NoWSD (row 1) shows that without disambiguation, 87% of the motion verbs are detected, but the precision is only 43%. The 13% false negatives are either cases of missing frames, or of missing lexical units in FrameNet and WordNet (phrasal verbs). The 57% false positives are due to ambiguous words with non-motion senses. Using WSD (row 2) substantially improves precision, but hurts recall, with a small loss in F-Score. Finally, the predominant sense heuristic outperforms WSD in both precision and recall. It cannot rival NoWSD in recall, but the higher precision yields a net gain in F-Score of 9%, the overall best results.

These numbers show that when the first sense of a verb is a motion sense, then heuristic assumption that instances of this verb belong to the motion domain outperform the UKB-provided disambiguation. (Note however that UKB tries to solve a more difficult task, namely fine-grained sense assignment.) The heuristic is nevertheless far from perfect: Among the 27% false positives, there are

²Newer mappings have been created by, e.g., Tonelli and Pighin (2009).

	Precision	Recall	F ₁ -Score
NoWSD	38	55	45
WSD	83	18	30
PredomSense	87	32	47

Table 4: Motion verb recognition results with WordNet

high-frequency high-ambiguity verbs like *take*, but we also find that many motion verbs specifically have a concrete motion sense but also a more abstract non-motion sense, often in the mental or cognitive domain. Examples are *struggle*, *lean*, *confront*, *follow*.

4.4 Motion Verbs in WordNet

While FrameNet is a comprehensive resource for defining motion verbs, it is available only for a small number of languages and suffers from coverage problems. We therefore also experimented with using only WordNet, which is available for a large number of languages.

Defining motion verbs in the WordNet hierarchy. WordNet uses a troponymy relation in the verbal domain which organizes verb senses into hierarchy structured by specificity. For example, *talk* is a troponym of *communicate*; in turn, *whisper* is a troponym of *talk*. Within this hierarchical organization, we can define the motion domain in WordNet as a (small) set of subtrees by identifying the root of these subtrees. This is similar to how, for example, semantic classes for selectional preferences are often represented in terms of WordNet subtrees (Resnik, 1996). The challenge is to find a set of nodes whose subtrees cover as much as possible of the motion domain while avoiding overgeneration. An inspection of WordNet led us to two nodes that meet these conditions well. The first one is "move, locomote, travel, go" (synset ID 01818343-V), which covers the motion domain. The second one is "to be (occupy a certain position or area; be somewhere)" (synset ID 02629830-V), which covers the orientation and position domains. The WordNet motion list formed by these nodes and their hyponyms comprises a total of 1090 synsets.

Results. The results for the three strategies from Section 4.2 applied to WordNet are shown in Table 4. Generally, we observe that WordNet, works

considerably worse than FrameNet. This is not surprising, given the additional layer of information that FrameNet provides, and the simplistic motion domain model we adopted for WordNet.

WordNot notably suffers from low recall. Even in the NoWSD condition, many verbs that are motion verbs are not included in the WordNet-derived list of motion verbs, the recall being only 55%. The reason appears to be that a number of motion verbs in particular are scattered in WordNet outside our two chosen subtrees. Examples include *crouch*, a hyponym of the synset *to change (to undergo a change)*, and *stand*, a hyponym of the synset *to be (have the quality of being)*. However, these motion verbs are not easy to assign to complete subtrees that can also be designated as motion subtrees.

Not surprisingly, NoWSD has by far the lowest precision of the three conditions, since many instances of verbs that have motion senses but also other senses are mistagged as motion verbs. The precision improves dramatically for the WSD condition, from 38% to 83%. However, the recall takes a further major hit down to 18%, and thus the resulting F-Score is very low. In the PredomSense condition, precision increases even somewhat further, and the decline in recall compared to NoWSD is not quite as pronounced. Consequently, the PredomSense condition shows the overall best F-Score for WordNet-based models. For both FrameNet- and WordNet-based models, therefore, it seems currently preferable to employ a predominant sense heuristic over performing full-fledged word sense disambiguation. The best WordNet-based result, 47% F-Score, is however still 20% below the best FrameNet-based result of 67%. An example for an instance that is wrongly classified as a motion verb even in the high-precision PredomSense condition is *to follow*, whose first WordNet sense concerns motion, but which was used in the corpus for “obeying” (following commands).

5 Discussion and Conclusion

In this paper, we introduced the task of locational inference and sketched a processing strategy. We presented a corpus study which provided evidence for the relevance of locational inference and performed an experiment which compared WordNet and FrameNet for the recognition of the motion

domain (motion, orientation, and position verbs). We found FrameNet to be a useful tool to define these semantic domains. Processing can also proceed when just WordNet is available, but unsurprisingly with lower results. Comparing different word sense disambiguation schemes, the unsupervised WSD system UKB could not beat the simple “predominant sense heuristic”.

Concerning the automatic recognition of motion verbs, one avenue of future research is the refinement of the characterization of motion verbs in FrameNet and WordNet. As we have observed, the location role-based collection of motion frames misses some frames which should be added to the list of frames. As for WordNet, our current definition which limits itself to two subtrees in the WordNet verb hierarchy leads to a very high precision but a low recall. Since WordNet was not designed specifically with the motion domain in mind, many other motion verbs are scattered throughout the verb hierarchy, and their distribution is difficult to describe succinctly. We will experiment with the precision/recall trade-off that arises from adding more synsets to the definition of motion verbs.

In terms of annotation, we also made, but not yet acted upon, the observation that many locations mentioned in a discourse are related hierarchically. For example, a person can be concurrently said to be in a seat, in the carriage and on the road, all of which fill the PLACE role with varying degrees of specificity. These descriptions can be said to be related through bridging.

Finally, we plan to bring these threads together in the realization of a complete pipeline for locational inference (cf. Figure 1). Our current experiments only concern the very first step of this pipeline, and the recovery and assignment of location roles that remain locally unrealized remains the ultimate goal of our research program.

Acknowledgments. This work was partially supported by the EC-funded project EXCITEMENT (FP7 ICT-287923) and the Research Training Group "Coherence in language processing" of Heidelberg University.

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41, Athens, Greece.
- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 86–90, Montreal, QC.
- Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the Impact of Frame Semantics on Textual Entailment. *Journal of Natural Language Engineering*, 15(4):527–550.
- Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.
- Katrin Erk. 2005. Frame Assignment as Word Sense Disambiguation. In *Proceedings of the 6th International Workshop on Computational Semantics*, Tilburg, The Netherlands.
- Charles J Fillmore. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, IV(2):222–254.
- Charles J Fillmore. 1986. Pragmatically controlled zero anaphora. In V Nikiforidou, M VanClay, M Niepokuj, and D Feder, editors, *Proceedings of the Berkeley Linguistics Society*, volume 12, pages 95–107.
- Matthew Gerber and Joyce Y Chai. 2010. Beyond nombank : A study of implicit arguments for nominal predicates. In *Proceedings of ACL*, pages 1583–1592, Uppsala, Sweden.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- Mark A Greenwood. 2004. Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, pages 17–22.
- Paul Grice. 1975. Logic and conversation. In P Cole and J Morgan, editors, *Syntax And Semantics*, pages 1–13. Academic Press, New York.
- Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2007. Satisfying information needs with multi-document summaries. *Information Processing and Management*, 43(6):1619–1642.
- David Herman. 2001. Spatial reference in narrative domains. *Text*, 4:515–541.
- Blake Stephen Howald. 2010. Linguistic spatial classifications of event domains in narratives of crime. *Journal of Spatial Information Science*, 1(1):75–93.
- Nancy Ide, Collin F. Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of LREC 2008*, pages 2455–2461, Marrakech, Morocco.
- Ray S Jackendoff. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.
- George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago, IL.
- Jochen L Leidner, Gail Sinclair, and Bonnie Webber. 2003. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT/NAACL 2003 workshop on the analysis of geographic references*, pages 31–38, Edmonton, Canada.
- Beth Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.
- Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 279–286, Barcelona, Spain.
- George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.
- Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the Role of Discourse References in Entailment Inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala, Sweden.
- Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy New Resources New Tools New Methods*, volume 3 of *English Corpus Linguistics*, pages 197–214. Peter Lang.
- Peter Norvig. 1987. Inference in Text Understanding. In *Proceedings of the National Conference on Artificial Intelligence*, pages 561–565, Seattle, WA.
- Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo. 2008. Testing the Reasoning for Question Answering Validation. *Journal of Logic and Computation*, 18:459–474.
- Philip Resnik. 1996. Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition*, 61:127–159.
- Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a Generic Paraphrase-Based Approach for Relation

- Extraction. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 401–408, Trento, Italy.
- Josef Ruppenhofer, Caroline Sporleder, R. Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.
- Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of the 6th CICLing*, pages 100–111. Springer.
- Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 1–10, Montréal, Canada, 7–8 June. Association for Computational Linguistics.
- Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.
- Sara Tonelli and Daniele Pighin. 2009. New features for FrameNet–WordNet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Boulder, CO, USA.

WordNet-Based Lexical Simplification of a Document

S. Rebecca Thomas

Bard College

Annandale-on-Hudson, NY

USA 12504

thomas@bard.edu

Sven Anderson

Bard College

Annandale-on-Hudson, NY

USA 12504

sanderso@bard.edu

Abstract

We explore algorithms for the automatic generation of a limited-size lexicon from a document, such that the lexicon covers as much as possible of the semantic space of the original document, as specifically as possible. We evaluate six related algorithms that automatically derive limited-size vocabularies from Wikipedia articles, focusing on nouns and verbs. The proposed algorithms combine Personalized Page Rank (Agirre and Soroa, 2009) and principles of information maximization, beginning with a user-supplied document and constructing a customized small vocabulary using WordNet. The best-performing algorithm relies on word-sense disambiguation with sentence-level context information at the earliest stage of analysis, indicating that this computationally costly task is nonetheless valuable.

1 Introduction

This report explores algorithms for the automatic generation of a limited size lexicon from a document, such that the lexicon covers as much as possible of the semantic space of the original document, as specifically as possible. While lexical simplification has typically worked at the sentence level, it is our belief that the document as a whole must be taken into account, and that creating a simpler vocabulary for the document and then applying that vocabulary to the sentences will give more appropriate results. This paper presents some results on the problem of developing a document-level vocabulary.

Tasks for which it would be desirable to generate a smaller, and ideally simpler, vocabulary include full-document text summarization and text paraphrase. In addition, simplification of a vocabulary can be extended to icon sets used by mobile devices or augmentative communication devices. In particular, one type of Augmented and Assistive Communication (AAC) system includes a touchscreen, from which users select icons that may, alone or in combination, represent specific words or phrases (Baker, 1982).¹ An AAC icon set represents a set of concepts customized by a human expert to a particular user. There is tension between ensuring that the collection of icons – the vocabulary – is large enough to be sufficiently expressive and ensuring that it is small enough to allow for efficient navigation and maximal communication speed, a major issue with AAC systems (Trnka et al., 2009). One might design the icon set beginning with a text corpus representing typical utterances of the particular user – perhaps from a log file kept by a communication device – and generate an icon set with more, and more specific, icons for topics of frequent communication. Likewise, for any device with a touchscreen interface, reducing the “vocabulary” of touchable icons without compromising expressivity allows for better usage of limited screen area.

In the next section we outline previous relevant work in simplification. The third section describes six methods for generating a reduced

¹Users with appropriate levels of literacy and motor control may be able to type on an alphabetic keyboard; these are not the users we are primarily concerned with in this work, but see (Wandmacher et al., 2008; Trnka et al., 2009).

lexicon from a document-derived vocabulary and describes two measures for evaluating the quality of the resulting lexicons. The fourth section discusses results obtained with English Wikipedia (EW) articles. The paper concludes with discussion and interpretation of our results.

2 Background

Text simplification is usually implemented in two stages: first syntactic analysis and simplification and then lexical simplification, which reduces the number of difficult words and expressions. The earliest approach to lexical simplification replaces words by simpler synonyms. Synonym difficulty is estimated using the frequency and length of each word: more difficult words usually have lower frequency of occurrence and a greater number of syllables. For example, (Carroll et al., 1998) queries WordNet for synonyms (synsets) and selects the most frequently occurring synonym in a synset as determined by frequency (Kucera and Francis, 1967). The authors wanted to avoid deep semantic analysis; thus, no word-sense disambiguation was performed. Lal (2002) extended this approach by including the syllable count as part of a word’s difficulty metric. More recently, lexical and syntactic simplification have been simultaneously implemented via statistical machine translation (Coster and Kauchak, 2011) and English Wikipedia edit histories (Yatskar et al., 2010). These studies did not directly evaluate the success of lexical simplification.

DeBelder et al. (2010) add a form of word-sense disambiguation to lexical simplification. The authors use the latent variables from a Bayesian network language model to suggest alternate words in a text. An advantage of this approach is that it does not necessarily rely on specialized semantic databases like WordNet, nor does it rely on psychological data to estimate word difficulty. DeBelder and Moens (2010) use this approach to simplify the syntax and lexicon of text for children. They show an improvement of about 12% in accuracy, as assessed by human judges, over the baseline model that simply chooses the most frequent synonym.

Lexical simplification within a sentence context was a SemEval-2012 task (Specia et al., 2012). The best of five submitted systems

employed word frequency as well as context-dependent information. The authors conclude that research is needed that evaluates the role of context in the simplification of all words in a context. The present paper extends preliminary results in (Anderson et al., 2011), describing derivation of a small, simplified vocabulary in which the context is the entire document: lexical choice decisions are made simultaneously and for an entire document, not sentence by sentence.

3 Reducing Vocabulary Size

3.1 Algorithms

We developed and tested six methods for generating the reduced lexicon. Given a starting document in standard English, we tag the text using the Stanford Part of Speech Tagger (Toutanova and Manning, 2000), extract all and only the occurrences of the part of speech we are interested in, and reduce each word occurrence to its base uninflected form using WordNet’s `morphstr` method.² We explored three levels of word-sense disambiguation: none, weak, and strong disambiguation. This step produced a base set from which we generate a reduced lexicon using frequency-based selection, weights generated by disambiguation, or a greedy algorithm described below. We discuss the disambiguation step and then the vocabulary reduction step, concluding with an overview of the resulting six algorithms.

Disambiguation Step

Both strong and weak word-sense disambiguation employed the approach of (Agirre and Soroa, 2009).³ The PageRank algorithm underlying PPR permits synsets to vote for one another depending on WordNet’s graph structure and the weight of each synset. This voting process is iterated until it converges to a stable solution that ranks vertices in a graph based on their relative importance. In weak disambiguation, each word appearing in the starting document is weighted proportional to the count of that word. All other synsets are initialized with weight zero, and PPR is performed once for the entire document. Strong disambiguation processes words one at a time by weighting

²We limit our focus to nouns and verbs, working with each part of speech independently.

³<http://ixa2.si.ehu.es/ukb/>

its neighboring words in the original text. Agirre and Soroa obtained superior disambiguation performance with this method, which they call *w2w*.

Lexicon Reduction Step

The lexical reduction step begins with the set of word senses from the first step and reduces these to a lexicon of any desired size. The two simplest algorithms merely select those N words with the highest frequency or those with the highest PPR weight after convergence.

We contrast simple reduction with a “greedy” information approach. The greedy algorithm constructs a subtree of WordNet containing all the words of a base set constructed by the first step. The WordNet graph for nouns is a tree with the word *entity* as its root. The graph for verbs is not a tree, lacking a common unique ancestor node; there are at least 15 different head nodes, and hundreds of verbs with no unique ancestor (Richens, 2008). We force a tree structure for the verb graph by adding a new node, which is made the parent of all the root nodes in the original verb forest.

We add to each tree node a count: for leaves, this is the number of occurrences of that word sense in the original document; for internal nodes, it is the sum of its own occurrences and its children’s counts.

The objective of lexical reduction is to derive an optimal set of hypernyms, H , for the document. Initially H contains only the root node. The algorithm works its way down the constructed subtree, at each step evaluating every child of every node already in H , and greedily adding the child that maximizes information gain to the growing lexicon; see Figure 1. Note that the sequence of choices of locally optimal children may not lead to a globally optimal solution.

The information gained when a child is added to the hypernym set is computed via

$$-p * \left(\frac{c}{p} \log \left(\frac{c}{p} \right) + \left(\frac{p-c}{p} \right) \log \left(\frac{p-c}{p} \right) \right)$$

where c is the number of word occurrences covered by the child synset and p is the number of word occurrences covered by its parent. At each iteration, this guarantees that the node that maximizes information gain, given prior choices, is added to H ; the number of word occurrences it

covers is subtracted from its parent’s count, as the newly added node is now the covering synset for those occurrences. If the parent synset no longer covers any word occurrences – that is, if its count reaches zero – then it is removed from H . Iterations continue until H reaches its target size.

Algorithms

Combining disambiguation and lexical reduction steps, we implemented and evaluated six of the resulting algorithms to create lexicons of pre-specified size N .

Frequency-based (Baseline) As a baseline, we follow the approach of (Carroll et al., 1998) and employ word frequency as a surrogate measure of simplicity. After disambiguating words using strong disambiguation (*w2w*), we then select as our lexicon those N words which have the greatest combined frequency as measured by the product of document frequency and frequency of occurrence in the Web1T corpus (Brants and Franz, 2006). Only unigram and bigram frequencies were employed. Given any word in the original vocabulary, we can traverse up the tree from it until we reach the first node that is a member of the lexicon; we say this is the lexicon element that *covers* the word. Note that a lexicon H will not necessarily cover every word in the initial document.

No Disambiguation (Greedy) No disambiguation is used. Each word’s number of occurrences is credited to every one of its senses, so the base set consists of all senses of every word (that is the proper part of speech) in the document. Lexical reduction employs the Greedy algorithm.

Personalized PageRank, Top-N (PPR-N) For each word appearing in the starting document, we find the corresponding synset(s) in WordNet and assign them weights proportional to the count of occurrences of that word. All other synsets are initialized with weight zero. After convergence of PPR, the N highest-weight synsets in the result comprise the reduced lexicon H . Again, H may not cover every word in the initial document.

Estimated Proportions of Word Sense Occurrences (PPR-WSD-G) Full PPR-W2W word sense disambiguation is time-consuming, requir-

```

Initialize hypernym set H to contain the root of the tree: {entity}
While H has not reached its final size,
    for each child c of each element p in H,
        compute information gained if c is added to H
    let x be the child that maximizes information gain; insert x into H
    subtract x's count from its parent's count
    if x's parent no longer covers any vocabulary, remove x's parent from H

```

Figure 1: Pseudocode for greedy algorithm

ing hours for longer articles. In the PPR-WSD-G approach, rather than performing full disambiguation of each word occurrence, we created a “context” consisting of the full set of words of the desired part of speech appearing in the document. Weight was assigned to every sense of each word, proportional to the number of occurrences of that word. PPR was run on this context, which we anticipated would concentrate weight in those parts of the tree where multiple weighty synsets reinforced each other, and thus for narrowly focused documents might give a reasonably accurate although approximate estimate of the comparative relevance of different senses of a given word.

For each word in the initial vocabulary, all its WordNet senses were ranked in decreasing order of their PPR-generated weights. The senses whose weights were less than 30%⁴ of the highest weight for this word were eliminated; the count of occurrences of this word was then distributed over the remaining senses, proportional to their PPR weights. The Greedy algorithm was then applied to shrink this set.

Weak Disambiguation (PPR-VOC-G) In this algorithm the entire initial vocabulary list was treated as a context for the PPR algorithm. We chose a base set approximately the same size as the starting list of unique words in the document. The PPR weights were multiplicatively scaled to provide the “counts” for this base set. The Greedy algorithm was then applied.

Strong Disambiguation, Top-N (PPR-W2W-G) The strongest context-based disambiguation (PPR-W2W) creates a vocabulary that is then reduced by application of the Greedy algorithm.

⁴Found via parametric testing.

3.2 Evaluation of Reduced Vocabularies

The ideal reduced lexicon would be capable of representing the same concepts as the original vocabulary, but using far fewer words. However, as lexicon size decreases, the semantic precision of expression tends to decrease as well; consequently, measures of the semantic quality of a lexicon are needed to assess our results.

Our primary evaluation of the algorithms is based on a comparison of lexicons from human-simplified text with those we generate algorithmically from non-simplified text. Specifically, articles and simplified versions of articles on the same topics were obtained from the English Wikipedia (EW) and Simple English Wikipedia (SEW) (Wikipedia, 2010). English Wikipedia articles contain a reasonably large sample of standard English written by diverse authors covering a range of topics. Simplified articles are not usually direct simplifications of original articles from the EW: a simplified article generally covers only a subset of the topics covered by the original article using a simpler grammar and lexicon. Consequently, reducing the lexicon of EW articles to that of SEW articles represents a difficult task for any lexical simplification algorithm.

Text and Simplified Text Corpora

All six algorithms were applied to a set of ten articles found in the SEW. We selected articles from the SEW that were listed as “good” simplified articles by managers of Wikipedia, that had at least three paragraphs of prose, and that had a majority of sections that appeared to correspond to sections in articles in the EW. All articles were manually edited to remove images, tables, and references, and to retain only those topics, usually indicated by sections, found in both the original and simplified article. We applied the Stan-

	Chop	City	Evol	Goth	Hum	Mon	Okla	RRH	Sat	Snake
Noun, EW	326	341	817	634	1141	587	817	399	440	225
Noun, SEW	208	136	511	434	390	338	285	228	224	68
Noun, common	145	52	265	295	266	255	232	119	161	34
Verb, EW	161	99	315	221	311	177	207	164	149	127
Verb, SEW	90	43	172	128	119	102	80	85	87	47
Verb, common	61	19	103	73	77	63	43	52	59	23

Table 1: Unique word counts for EW and SEW articles, and the number of words EW and SEW have in common

ford Part of Speech Tagger (Toutanova and Manning, 2000) to isolate nouns and verbs from all documents; resulting noun lexicons from SEW are 30–68% of the size of the EW’s vocabulary, while for verb lexicons the corresponding figures are 37–58% (see Table 1). The number of unique nouns in the EW articles varied from 225–1141, with the number of unique nouns in SEW articles ranging from 68–511. Similarly, there were 99–315 unique verbs in the EW articles, and 43–172 unique verbs in the SEW articles.

Comparing the noun and verb counts, we note that the number of unique verbs is about half the number of unique nouns. The verb count varies less across the ten articles, suggesting that, as articles grow longer, they incorporate more additional nouns than additional verbs. Table 1 also indicates the degree of overlap among the noun and verb lexicons for each pair of articles.

Evaluation Measures

Let us call the vocabulary (restricted to a single part of speech) of an EW article L , and the vocabulary of the SEW article on the same topic S . Then we apply one of our algorithms, described above, to L in order to produce a reduced lexicon (or hypernym set) H . We use two measures to assess the quality of the hypernym set. Our first measure, affinity, is used to measure the semantic distance between L and the H generated from it, in order to assess whether expressivity has been adequately retained. The second, Symmetric Vocabulary Distance, is used to measure the semantic distance between the automatically-generated H and S , under the assumption that the vocabulary of the human-simplified text (S) is a reasonable proxy for a human-approved reduction of L . We set the desired size of H to be the size of S , so that we can fairly compare the semantic space

covered by H to that covered by S .

Affinity Between Starting Vocabulary and Hypernym Set Intuitively, we aim to generate a precise lexicon, i.e. one in which the semantic distance between the starting vocabulary and the reduced lexicon is minimized; thus the most precise lexicon is the original vocabulary. To operationalize this intuition, we experimented with several distance measures based on path distance in the WordNet tree (Budanitsky and Hirst, 2006) and ultimately adapted a scoring measure proposed in (Widdows, 2003), which finds the distance in the WordNet subtree between a vocabulary word’s sense and its nearest ancestor (hypernym) in the lexicon. Widdows calls the inverse square of this distance the *affinity* score for that word. Suppose there are N vocabulary word senses in the base set L , and $\text{dist}(x)$ is the distance (number of edges plus one) in the tree from a word sense x to its hypernym in the reduced lexicon H , or ∞ if there is no such hypernym. If c is defined to be the weight (number of occurrences in the document) of the current sense, and C is the summed weight of the entire vocabulary, then our distance measure is defined as:

$$\frac{1}{C} * \sum_{i=1}^N \begin{cases} \frac{c}{\text{dist}(i)^2} & \text{if } \text{dist}(i) \neq \infty \\ \frac{-c}{4} & \text{if } \text{dist}(i) = \infty \end{cases} \quad (1)$$

Affinity scores increase as distance between synsets decreases.

Symmetric Vocabulary Distance When comparing two vocabularies, an intuitive measure of difference is the semantic distance between a word in the first vocabulary and the word in the second that is semantically nearest to it. This intuition leads to the following definition of vocab-

ulary distance. Let $d(a, b)$ denote a measure of semantic distance between words a and b , here measured by the count of WordNet edges in the shortest path from a to b . Suppose the vocabulary and reduced lexicon are represented as sets S and H , respectively. We define the distance between a word, $s \in S$ and (the entire) lexicon H to be $d(s, H) = \min_{h \in H} d(s, h)$. Summing over all the elements of S gives a distance measure that is asymmetric, and we therefore define the symmetric distance between S and H by

$$d(H, S) = \frac{\sum_{s \in S} d(s, H) + \sum_{h \in H} d(h, S)}{2}.$$

4 Evaluation Results

Average affinity scores between vocabulary words in L and their nearest (covering) hypernyms in H are shown in Figure 2. Intervals were generated using bootstrap resampling with 95% confidence. Looking at the results for nouns first, confidence intervals indicate a significant difference between the top two algorithms, Greedy and PPR-W2W-G. For purposes of comparison, we note that these “best” scores are lower than those reported by Widdows (2003) for class labels that correctly classify nouns. In that study Widdows found that affinity scores in the range (0.67, 0.91) were indicative of correct class labels for common nouns, but that affinity scores of about 0.57 indicated incorrect labels. Turning to verbs, again the best results are produced by the PPR-W2W-G algorithm, and again this algorithm appears significantly better than the rest of the algorithms, none of which appear significantly better than the others. In all cases, the results for verbs appear better than the results for nouns; this is at least partly explained by the shallowness of the WordNet verb hierarchy, as compared to the noun hierarchy.

The symmetric distances in Table 3 show the distances between automatically reduced lexicons (H) as compared to the vocabulary from human-simplified text (S). Our algorithms construct lexicons that are on average 1-2 edges away from the manually simplified lexicon; these words are much nearer than those in randomly selected lexicons, which are experimentally measured in lexicons of size 100 to 1000 as about 7 to 4 edges distant for nouns and 5 to 3 edges distant for verbs.

Based on the average results, for both nouns and verbs the best performing algorithm is again PPR-W2W-G. The other algorithms have distance scores similar to those of the frequency-based baseline. Word sense disambiguation appears to improve the precision of the reduced lexicon, though the differences are only sometimes clearly significant. As is the case for the affinity measure, the symmetric vocabulary distances for verbs appear better than for nouns, for every algorithm, in some cases significantly so. Here the difference in the percentage of shared unique verbs (60% on average) vs. shared unique nouns (51% on average) between the EW and SEW articles (L and S) may explain some of this apparent superiority of verb results.

Another way to consider these vocabulary distance results is to compare them to the vocabulary distance of the EW and SEW versions of each article. If the algorithms are successful, we expect the average distance between the reduced lexicons of the full article and the original vocabulary of the human-simplified articles (H and S) to be less than the distance between the original and human-simplified (L and S) lexicons themselves. The average distance between the original full and simplified lexicons (H and S) is 1.97 for nouns and 1.45 for verbs, compared with 1.50 and 1.16 respectively between L and S for the reduced lexicons produced by the best performing algorithm (PPR-W2W-G), a 24% and 12% improvement, respectively.

Examples of Vocabulary Substitution

Two examples of vocabulary replacement for verbs are shown below. We use the evolution article from the EW, which has 458 unique verbs (the largest number among our ten articles). The SEW article on the same topic has 272 unique verbs, and so we produce a reduced lexicon of the same size, a 41% reduction, using the best-performing algorithm, PPR-W2W-G.

In the EW article on evolution, we find the sentence

“IF ONE SPECIES CAN OUT-COMPETE ANOTHER, THIS COULD PRODUCE SPECIES SELECTION, WITH THE FITTER SPECIES SURVIVING AND THE OTHER SPECIES BEING DRIVEN TO EXTINCTION.”

Four verbs are marked as such by the W2W word sense disambiguation algorithm; in their

Algorithm	Nouns		Verbs	
	Avg. Distance	95% C.I.	Avg. Distance	95% C.I.
Baseline	0.20	(0.15, 0.26)	0.19	(0.13, 0.26)
Greedy	0.32	(0.23, 0.40)	0.49	(0.46, 0.53)
PPR-N	0.01	(-0.03, 0.06)	0.48	(0.42, 0.54)
PPR-V-G	0.14	(0.11, 0.17)	0.51	(0.46, 0.55)
PPR-WSD-G	0.15	(0.13, 0.17)	0.45	(0.39, 0.52)
PPR-W2W-G	0.65	(0.56, 0.72)	0.76	(0.72, 0.80)

Table 2: Average affinities between vocabulary words (from the EW articles) and their nearest hypernyms; higher is better. The averages and 95% confidence intervals are shown for each algorithm.

Algorithm	Nouns		Verbs	
	Avg. Distance	95% C.I.	Avg. Distance	95% C.I.
Baseline	1.93	(1.68, 2.25)	1.35	(1.20, 1.52)
Greedy	2.04	(1.77, 2.37)	1.31	(1.22, 1.41)
PPR-N	1.85	(1.66, 2.09)	1.58	(1.43, 1.74)
PPR-V-G	1.87	(1.63, 2.12)	1.29	(1.20, 1.40)
PPR-WSD-G	1.90	(1.63, 2.20)	1.65	(1.47, 1.81)
PPR-W2W-G	1.50	(1.33, 1.73)	1.16	(1.04, 1.29)

Table 3: Average distances between reduced size noun and verb lexicons, automatically generated from EW articles, and vocabularies extracted from the SEW human-simplified articles on the same topics; lower is better.

root forms they are *produce, survive, be, drive*. The hypernyms selected for each verb by the PPR-W2W-G algorithm are inserted into the sentence below⁵; we manually modified each verb tense to match:

“IF ONE SPECIES CAN OUT-COMPETE ANOTHER, THIS COULD (PRODUCE) SPECIES SELECTION, WITH THE FITTER SPECIES (LIVING) AND THE OTHER SPECIES (BEING) (MOVED) TO EXTINCTION.”

An example that did not give very satisfying results is:

“FOR EXAMPLE, MORE THAN A MILLION COPIES OF THE ALU SEQUENCE ARE PRESENT IN THE HUMAN GENOME, AND THESE SEQUENCES HAVE NOW BEEN RECRUITED TO PERFORM FUNCTIONS SUCH AS REGULATING GENE EXPRESSION.”

Here, the verbs that are marked as such are *be, recruit, perform, regulate*. The closest hypernym for *recruit, perform*, and *regulate* is the root node we added to the verb graph to make it a tree. The words *recruit* and *regulate* each appears only once in the entire article, and *perform* only twice (both

times with the same sense). Words that appear rarely may not be worth inclusion in the reduced lexicon, especially given a very limited lexicon size. One could attempt to automatically detect such cases, considering infrequency of word use and distance from a word to its nearest hypernym.

Vocabulary Reduction as Simplification

We estimated the difficulty of the various vocabularies derived by the best-performing PPR-W2W-G algorithm by calculating the average Kucera-Francis word frequency over all words in the vocabulary. Multi-word lexical items (e.g., ‘get the better of’) which are not found in the frequency data were excluded. In addition, words with more than one sense are conflated in the frequency counts and are therefore all senses are treated as a single entry for purposes of measuring frequency. In the case of noun vocabularies, the reduced lexicon found by the PPR-W2W-G algorithm had an average frequency of 123, lying nearer the average frequency of nouns in the EW article (109) than that of the simplified SEW article (153). The average verb frequency score of

⁵The hypernyms are synsets rather than words; we simply took the first word from each synset.

the reduced lexicon (243) was nearer the simplified score (290) than the wiki score (163). Thus, in both instances the reduced vocabulary consists of more frequently occurring, and thus arguably simpler, words than the original full vocabulary.

5 Discussion

Our results show that the greedy maximization of information can be combined with word sense disambiguation to yield an algorithm for the automated generation of a reduced lexicon for text documents. Among the six algorithms, we found significant differences between an approach that ignores word sense ambiguity and those that address this ambiguity explicitly. The introduction of multiple senses for every word, most of which are unintended in the original document, introduces sense ambiguity that is not overcome by document word counts, even if those counts are readjusted to weight likely senses. Early word sense disambiguation takes advantage of phrase and sentence context (unavailable at later stages of processing) and results in a smaller tree to be searched.

Comparison of the best algorithm (PPR-W2W-G) with the word-frequency Baseline suggests to the relevance of the semantic hierarchy in addition to word-sense disambiguation. The Baseline algorithm uses full word-sense disambiguation, but not the semantic hierarchy of Wordnet. PPR-W2W-G uses both types of semantic information. The superiority of PPR-W2W-G may also be a by-product of evaluation measures that are based on a Wordnet-based definition of semantic distance. In the future, a better test of the relative importance of the two aspects of semantics may be obtained from other judgements of semantic similarity.

One would intuitively expect that lexical simplification is optimally implemented at the level of the document, not the sentence. Although our algorithm does not explicitly select “simple” words for the lexicon, the combined algorithm yields a lexicon in which most words are only 1-2 edges away from human-simplified counterparts. Since our greedy search of WordNet is top-down, our hypernyms lie above the words drawn from the original document; our intuition that these more general words tend to be simpler

is supported by the higher average frequency of the resulting lexicons. These results do not offer a complete approach to lexical simplification, but suggest a role for contextualized semantics and appropriate knowledge-based techniques.

6 Acknowledgements

We thank Dr. Barry Wagner for bringing the AAC interface problem to our attention and the Bard Summer Research Institute for student support. Students Camden Segal, Yu Wu, and William Wissemann participated in software development and data collection. We also thank the anonymous reviewers for their suggestions.

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 33–41, Morristown, NJ, USA. Association for Computational Linguistics.
- Sven Anderson, S. Rebecca Thomas, Camden Segal, and Yu Wu. 2011. Automatic reduction of a document-derived noun vocabulary. In *Twenty-Fourth International FLAIRS Conference*, pages 216–221.
- Bruce Baker. 1982. Minspeak: A semantic compaction system that makes self-expression easier for communicatively disabled individuals. *Byte*, 7:186–202.
- Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. Linguistic Data Consortium, Philadelphia.
- Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, March.
- John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.
- William Coster and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 665–669.
- Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of the First International Conference on*

- Interdisciplinary Research on Technology, Education and Communication (ITEC 2010)*, pages 43–63. Cambridge Univ Press.
- Henry Kucera and W.Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, Rhode Island.
- Partha Lal and Stefan Rüger. 2002. Extract-based summarization with simplification. In *Proceedings of the Workshop on Text Summarization at the Document Understanding Conference*.
- Tom Richens. 2008. Anomalies in the WordNet verb hierarchy. *Proceedings of the 22nd International Conference on Computational Linguistics - (COLING '08)*, (August):729–736.
- Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 Task 1: English lexical simplification. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 347–355.
- Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70.
- Keith Trnka, John Mccaw, Debra Yarrington, Kathleen F McCoy, and Christopher Pennington. 2009. User Interaction with word prediction: The effects of prediction quality. *Association for Computing Machinery Transactions on Accessible Computing*, 1(3):1–34.
- Tonio Wandmacher, Jean-Yves Antoine, Franck Poirier, and Jean-Paul Dèparte. 2008. SIBYLLE , an assistive communication system adapting to the context and its user. *Association for Computing Machinery Transactions on Accessible Computing*, 1(1):1–30.
- Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, number June, pages 197–204. Association for Computational Linguistics.
- Wikipedia. 2010. Wikipedia, the free encyclopedia. en.wikipedia.org/wiki.
- Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368.

Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs

Ines Rehbein

SFB Information Structure
German Department
Potsdam University

irehbein@uni-potsdam.de

Josef Ruppenhofer

Information Science and
Natural Language Processing
Hildesheim University

ruppenho@uni-hildesheim.de

Caroline Sporleder Manfred Pinkal

Computational Linguistics
Cluster of Excellence MMCI
Saarland University

csporled,pinkal@coli.uni-sb.de

Abstract

This paper presents Release 2.0 of the SALSA corpus, a German resource for lexical semantics. The new corpus release provides new annotations for German nouns, complementing the existing annotations of German verbs in Release 1.0. The corpus now includes around 24,000 sentences with more than 36,000 annotated instances. It was designed with an eye towards NLP applications such as semantic role labeling but will also be a useful resource for linguistic studies in lexical semantics.

1 Introduction

We present SALSA Release 2.0, a lexical-semantic resource for German. SALSA provides annotations of word senses, expressed through the frame-semantic classification of predicates, their semantic roles and syntactic realization patterns. These frame-semantic annotations in the flavor of the Berkeley FrameNet (Baker et al., 1998) are added as a complementary annotation layer to the TIGER treebank (Brants et al., 2002), a syntactically annotated corpus of German newspaper text.

Now that the SALSA project has concluded we do not only want to present the resulting resource but also take the opportunity to revisit some central methodological and analytical issues which came up during the frame development and annotation process. Since SALSA is so centrally related to FrameNet, we will typically cast our discussion as a comparison between SALSA and FrameNet, highlighting key differences. In particular, we discuss the differences in the workflow;

differences in the organization and representation of lexical items; the use of underspecification; and the treatment of metaphor.

The remainder of this paper is structured as follows. In Section 2, we briefly recap the central ideas of Frame Semantics. Section 3 then gives an overview of the size and composition of SALSA. In Section 4, we describe our efforts at quality control in the creation of the resource, focusing on inter-annotator agreement. Section 5 discusses some central methodological and analytical issues that came up in the work of SALSA, situating the discussion against the background of how FrameNet handles the same issues. Finally, we offer conclusions in Section 7.

2 Frame Semantics

SALSA provides annotations in the framework of Frame Semantics (Fillmore, 1982). Frames are representations of prototypical events or states and their participants. In the FrameNet database (Baker et al., 1998), a lexical database of English which implements the ideas of Frame Semantics in the sense of Fillmore (1982), both frames and their participant roles are arranged in various hierarchical relations (most prominently, the is-a relation). FrameNet links these descriptions of frames with the words and multi-words (lexical units, LUs) that evoke these conceptual structures. It also documents all the ways in which the semantic roles (frame elements, FEs) can be realized as syntactic arguments of each frame-evoking word by labeling corpus attestations.

By way of example, consider the *Change position on a scale* frame (Figure 1), evoked in English by lexical units across different word classes such as accelerated.a, advance.v, climb.v, contraction.n, and others. In the German SALSA corpus, the frame is licensed by frame-evoking elements like, e.g., *abstürzen.v*, *erhöhen.v*, *gewinnen.v*, and *klettern.v* (crash, increase, win, climb). The core set of frame-specific roles that apply includes ATTRIBUTE, DIFFERENCE, FINAL_STATE, FINAL_VALUE, INITIAL_STATE, INITIAL_VALUE, ITEM and VALUE_RANGE, out of which only ITEM and ATTRIBUTE are realized in our example.

3 Overview of the SALSA corpus

The SALSA project used the frames of FrameNet Releases 1.2 and 1.3 to perform German annotation on top of the TIGER corpus. Since the English frames were not always available or appropriate, SALSA additionally developed a number of "proto-frames", i.e., predicate-specific frames, to provide coverage for predicate instances that were not covered by FrameNet at the time SALSA analyzed relevant German vocabulary. Figure 1 shows a sentence from the TIGER corpus that is annotated with one original FrameNet frame, *Change position on a scale* with frame elements ITEM and ATTRIBUTE, and with one SALSA proto-frame, *Zahl1-salsa* with its frame element INDIVIDUALS, which is assigned to the same NP as the frame element ITEM of *Change position on a scale*.

SALSA Release 1.0 (Burchardt et al., 2006) was published in October 2007. The total size of the annotations in SALSA Release 1.0 includes

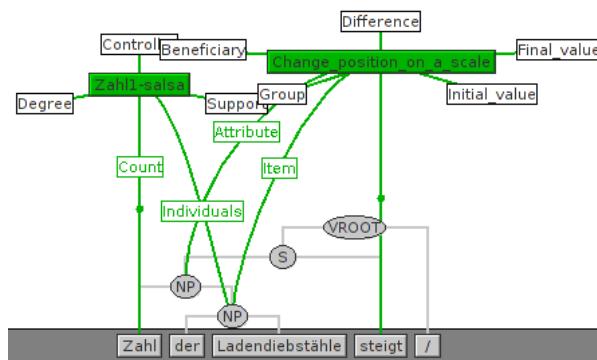


Figure 1: "Number of shoplifting cases increases."

20,380 annotated instances for 493 verb lemmas and 348 annotated instances for 14 noun lemmas (Table 1).

SALSA	Release 1.0		Release 2.0	
	token	type	token	type
verb	20,380	493	20,380	493
noun	348	14	15,871	155
total	20,728	507	36,251	648

Table 1: Annotated instances in SALSA 1.0 and 2.0

SALSA Release 2.0 complements the verb annotations in Release 1.0 with the annotation of more than 15 000 noun instances (Table 1). The selection of nouns that were annotated mostly includes nominalizations (e.g. *Eroberung* (conquest), *Freude* (delight)) and relational nouns like *Bruder* (brother) or *Seite* (side). The annotation scheme follows the one for verbs as closely as possible. There are, however, differences due to

- multi-word expressions, e.g. *guter Hoffnung sein* (expect a baby, be expecting)
- constructions, e.g. *wählen* (choose) – *X n-ter Wahl* (as in e.g. *second quality socks*)
- named entities, e.g. *Der Heilige Krieg* (holy war)
- lack in parallelism of verb and noun senses, e.g. *gut ankommen* ('be well received' – *Arrival)

The goal in selecting the nouns chosen for analysis was to achieve a balanced distribution with lexical units across all frequency bands (Table 2).

An interesting point of comparison is the number of realized frame elements for nouns and verbs in the SALSA corpus. As shown by Table 3, the average number of FEs is higher for verbs at 1.91 than for nouns at 1.45. Note that this is the case despite the fact that during the annotation of

SALSA	Release 1.0	Release 2.0
>500	2	4
301-500	6	17
101-300	41	81
51-100	68	93
31-50	75	89
11-30	99	122
<=10	217	242

Table 2: Frequency distribution of lexical units in SALSA (annotated frame instances)

SALSA	verbs	nouns
0 FE	0	2
1 FE	68	91
2 FEs	376	57
3 FEs	49	4
4 FEs	0	1
avg. # FEs	1.91	1.45

Table 3: Avg. number of realized frame elements (FEs) per lemma for verbs and nouns in SALSA

nouns the annotators were allowed to also assign non-core frame elements when suitable, while the vast number of verbal annotations do not include any non-core elements.¹

In the next section we describe our efforts at quality control during the annotation process.

4 Inter-Annotator Agreement

The annotation process in SALSA involved a thorough quality control, including double annotation by two independent annotators for *all* instances in the corpus. Disagreements in the annotations were resolved by two expert annotators in an adjudication phase. Table 4 shows inter-annotator agreement² for frames and frame elements. Despite showing approximately the same degree of polysemy (Table 4), agreement on noun frames is higher than for verbal targets.

For frame elements, however, percentage agreement for nominal instances is far below the one for verbs. This is mostly due to the annotation of SUPPORT and CONTROLLER, which rank highest among the frame elements on which the annotators disagreed. Percentage agreement for SUPPORT is 43.2%, and the agreement for CON-

¹In the first phase of the project when only verbs were annotated, the policy was to usually forgo annotating non-core FEs in favor of annotating a greater number of targets with core FEs.

²We do not report chance-corrected agreement but percentage agreement for reasons discussed in (Burchardt et al., 2006). In addition, chance-corrected agreement measures like the kappa statistics are often misleading when applied to unbalanced data sets. Consider, e.g., the lemma *Bruder* (brother), where we have 29 annotated instances out of which 28 are assigned the frame *Kinship* by both annotators. However, due to the skewed distribution of word senses the chance-corrected agreement (Fleiss κ) is only 0.491, while the percentage agreement of 96.6% better reflects that the annotators agreed on all but one instance for the lemma *Bruder*.

SALSA	frames	frame elements	# word senses
verbs	84.6	81.0	2.6
nouns	92.9	73.3	2.7
all	87.1	78.2	2.6

Table 4: Percentage agreement for human annotators in SALSA 2.0 on frames and frame elements and avg. number of word senses (frames) per verb/noun lemma

TROL is even lower with 23.6%.³ Another reason for the low coder agreement on frame elements is caused by relaxations in the annotation guidelines which, unlike in Release 1.0, allowed the annotators to assign non-core frame elements for nouns, when appropriate. Apparently, some annotators made more use of this than others. Amongst the 15 highest-ranking frame elements on which annotators disagreed most frequently, we found DESCRIPTOR, DEGREE, PERSISTENT_CHARACTERISTIC, PLACE and TIME, all of which are non-core frame elements. In most cases, these FEs had only been annotated by one annotator, while the other one had assigned the core frame elements only.

Amongst the frames which proved to be most difficult for the annotators are the ones expressing fine-grained distinctions between related meanings, such as the distinction between *Statement* and *Telling*, between *Being_born* and *Birth*, or between *Judgment* and *Judgment.communication*. Other difficult cases involved the annotation of more abstract concepts like *Causation*, which was often mixed up with the *Reason* frame.

5 Comparison of SALSA and FrameNet

Now we want to focus on some central methodological and analytical issues which came to our attention during the annotation process.

5.1 Frame development and workflow

A key difference between SALSA and FrameNet lies in how the vocabulary to be analyzed and annotated is chosen. FrameNet has two modes of working, a lexicographic and a full text one. In full-text analysis mode the goal is to cover running text as densely as possible with frame semantic annotations. When frames are found to be

³Efforts to overcome this difficulty by replacing support and control predicates with fully fleshed out frames and proto-frames did not succeed.

missing, they have to be created “on the fly”. By contrast, in lexicographic work, frames are developed and lemmas identified that can evoke them. In this mode, patterns of polysemy of lemmas typically lead the analysts from the description of the current frame to the description of a related but different one. The database as a whole grows organically.

Correlated with the differences in how the frames are chosen are differences in the annotators’ task. In the lexicographic mode, annotators proceed by focusing sequentially on the different lexical units of the frame. For each LU, they label a set of instances that were extracted from a large corpus based on syntactic pattern or collocates. The annotators deal with only one LU per sentence and they get to select cases where the target lemma at issue clearly evokes the frame of interest.

SALSA’s way of working combined aspects of FrameNet’s lexicographic and full-text modes, as it aimed to achieve full coverage for all the uses of a set of lemmas in the TIGER corpus. The set of lemmas to be analyzed was chosen mainly with reference to the lemmas’ frequency in the corpus. SALSA annotators, like FrameNet full-text annotators, had to classify every instance of a predicate, not being able to only mark up the clearest examples. In terms of the frame inventory, SALSA re-used as many FrameNet frames as possible. In some cases, minor modifications were made to the FrameNet frames, either to accommodate peculiarities of German or to allow for a more consistent annotation when the English FrameNet seemed to make too fine-grained distinctions. Still in other cases, SALSA had to create so-called proto-frames for specific word senses not covered by FrameNet. Since the set of lemmas to be analyzed was mainly frequency-based, the average SALSA lexical unit has fewer known frame-mates than the average FN lexical unit.

SALSA, having 1826 lexical units (1349 verbal ones, 477 nominal ones) and 36,251 annotated frame instances, has defined more than 1,000 different frames. By comparison, FrameNet’s release 1.5, which has more than 10,000 lexical units and includes about 150,000 manually annotated frame instances has only 1019 frames. Table

	Verbs	Nouns	LUs
FN frame	865	163	1,028
modified FN frame	33	35	68
proto-frame	451	279	730
Total (LUs)	1,349	477	1,826

Table 5: Distribution of frames across lexical units

5 shows that many of SALSA’s frames are proto-frames, defined for a specific sense of a specific lemma.

However, the numbers for proto-frames are somewhat misleading. Since SALSA kept the FrameNet frame inventory of Releases 1.2 and 1.3 as its reference point, some of SALSA’s verb-sense specific proto-frames have already been covered by later FrameNet Releases. For instance, the proto-frame *Abnehmer1-salsa* for the lemma *Abnehmer:n* has been superseded by FrameNet’s frame *Commerce_scenario*.

As a measure of the degree to which SALSA could re-use FrameNet’s analyses, consider that of the 1023 frame types used by SALSA, 730 are proto-frames, 256 original FrameNet frames and 37 modified FrameNet frames (these latter recognizable by a frame name ending in “fn salsa”). In other words, about 1 in 8 of the FrameNet frames used was adapted in some way. These adaptations typically concern Frame Elements: they either receive broader definitions, pairs of them are merged, or, more rarely, new ones are introduced into the frame.

Conversely, we can also ask whether the proto-frames that SALSA developed might be of use to FrameNet, too. Since no completely up-to-date record exists for which SALSA proto-frames have been made redundant by new frames created by the FrameNet project, we will consider a 50-item random sample out of the 730 proto-frames. In 15 cases, the proto-frame could be replaced directly by a now existing FrameNet frame, indicating compatible analyses. 15 more proto-frames represent cases where there are very clear English translation equivalents, suggesting that a FrameNet frame would be needed anyway. An example of this is *Attentat* ‘attempt (on sb’s life)’.

For the remaining 20 proto-frames, the question of integration depends on policy decisions and preferences for generality or specificity. As an example, consider the German verb *aufschla-*

gen, which in one of its meanings is typically translatable as *add*. However, it has very narrow selectional and contextual restrictions, being typically used to talk about adding a surcharge to the cost of a good or service. German also has a morphologically related noun *Aufschlag* (and also *Zuschlag*), which translates English *surcharge*. If one decides to have a specific frame in English for this narrow concept, then the German proto-frame may serve as a seed. If English FrameNet decided to only use a more general frame for adding, *aufschlagen* could evoke that, too, but the original definition of the proto-frame would need to be given up. As another example consider the German reflexive verb *sich durchsetzen*, which covers greater semantic ground than its possible translations, among which are *win out*, *get one's way*, and *become accepted*. English FrameNet may not have need for a frame general enough to host *sich durchsetzen*, if it chooses to relate e.g. *get one's way* more specifically to contexts of human competition or argument. On the other hand, the FrameNet hierarchy would probably not be harmed by having a general frame, to which more specific ones can be related. Generally, the challenges in aligning frames and lexical units across the two languages do not seem to be fundamentally different from what is involved in aligning frames and lexical units in one language.

5.2 Organization and representation of lexical units

A basic difference between SALSA and FrameNet is that the coverage of SALSA is limited to the frames (word senses) needed for the TIGER corpus. Senses/frames of a lemma that were not attested are not accounted for. For instance, while the lemma *einräumen* has an attested sense of ‘conceding’ in the TIGER corpus, its sense of ‘filling up, stocking’ is not attested and, consequently, is missing from SALSA. FrameNet, by comparison, does list lexical units in frames anyway even when it cannot provide annotations for lack of attestations in the corpus or lack of resources for performing the annotations.

In SALSA, unlike FrameNet, lemmas have a more prominent role. For one, the annotations are distributed in one file per lemma. More importantly, multi-word expressions are not repre-

sented outright but are treated as senses of one of their component lemmas. For instance, *ins Auge gehen* ‘go wrong, backfire’ is simply a sense of the lemma *Auge*. Note that this treatment of multi-words as part of the treatment of a component lemma is motivated by purely pragmatic considerations. Multiwords in Salsa were included from a decoding perspective, that is, because one of their component lemmas (e.g. *Auge* ‘eye’) was being analyzed and the multi-word uses needed to be covered so as to guarantee the exhaustive treatment of all tokens of the lemma in question. In FrameNet, by contrast, idiomatic multi-word expressions are treated as lexical units and they are included in the resource from an encoding perspective, that is, because they have the right semantics to fit a particular frame. In FrameNet, multi-word lexical units may consist of lemmas that have no other sense in the resource.

Because of the lack of explicit representation, we estimated the percentage of multi-words in SALSA in a 100 item random sample of lexical units: 8 percent of the lexical units are multi-word items.

Another area where SALSA differs subtly from FrameNet is in the handling of support and control predicates, which are recorded in the annotation of frame elements outside of the maximal projection of the frame evoking element (Figures 2 and 3). Table 6 displays the distribution of the special governors that are recognized by SALSA.

	Contr.	Supp. N	Supp. V	Total
instances	692	1644	752	3088
types	451	249	52	663

Table 6: Distribution of Support and Control in SALSA

In SALSA, support predicates can receive two types of treatment. Within frames evoked by nouns, an honorary frame element SUPPORT is used to label support verbs and prepositions (cf. column *Supp. N* in Table 6). An example can be seen in Figure 2. Copular verbs are also treated as instances of SUPPORT predicates, unlike in FrameNet. Unlike Support predicates, Controllers are said to introduce a distinct event from that of the target. They do, however, share at least some frame element with the event of the target. The constituent expressing that shared par-

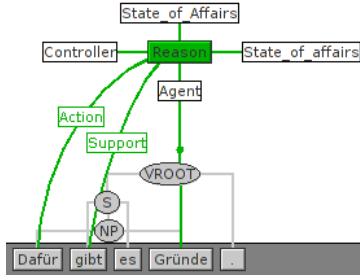


Figure 2: Support verb annotated as honorary FE
“For this, there are reasons.”

ticipant is labeled with a frame element relative to the noun target.

A second treatment that support uses can receive is that they are represented by a support sense of a verb lemma as shown in Figure 4 (cf. column *Supp. V* in Table 6). However, such support-frames have only one frame element, SUPPORTED, for the supported noun. The other arguments of the support verb, let alone those of the noun, are not annotated. Note that this second treatment is motivated by SALSA’s self-imposed goal of analyzing all instances of the lemmas it works with. If support verbs could only be marked from within noun frames, frames for all the nouns that the verb lemmas can support would have to be created, too.

Overall, neither FrameNet nor SALSA meet Heid (1998)’s desiderata of which types of information to collect about noun-verb collocates. Most importantly, although both resources acknowledge the importance of support verbs, neither resource treats support verb-noun combinations as first-class citizens, that is, as separate lexical entries, as noted above. Neither FrameNet nor SALSA give an explicit characterization of the semantics of support verbs, for instance, in terms of (Mel’čuk, 1996) lexical functions. Implicitly, all occurring support verbs appear to be synonyms and pragmatically (including registrally) equivalent. For an attempt to semi-automatically categorise FrameNet support verbs in terms of lexical functions see (Ramos et al., 2008). Morphosyntactic constraints on the support predicate or the supported predicate are not stated and may only be observed from the annotations. For instance, the noun *Opfer* ‘victim’ can occur as part of the structure *zum Opfer fallen*

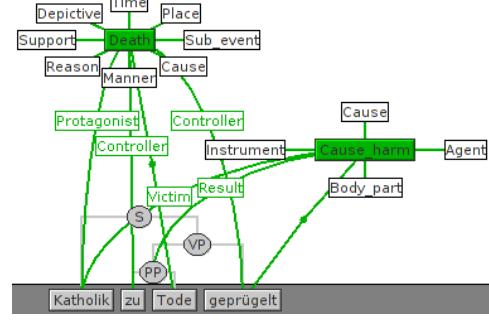


Figure 3: Annotation of Controllers as honorary FEs
“Catholic beaten to death”

‘fall victim’. In this collocation, the contraction *zum* cannot be uncontracted to the combination consisting of the preposition *zu* and the dative determiner form *dem*, which is possible in productively formed combinations.

5.3 Underspecification

One problem for annotating in full-text mode is the fact that all instances of the target lemma have to be resolved. There are many cases where the context does not fully disambiguate the word sense of the lemma of interest. When annotating in the lexicographic mode, it is possible to simply dismiss those cases and focus on the prototypical uses of a particular word meaning. In SALSA, ways had to be found to deal with this issue.

Therefore, underspecification was introduced to increase inter-annotator agreement for cases where the annotators were confronted with a decision where two or more solutions seem equally adequate. Underspecification also accommodates the fact that word meanings are often by no means clear-cut but rather seem to reflect gradience by showing a certain degree of sense overlap (Erk and Padó, 2007).

Underspecification is used in either of two cases. Firstly, two frames (or frame elements) can be underspecified when they both cover part of

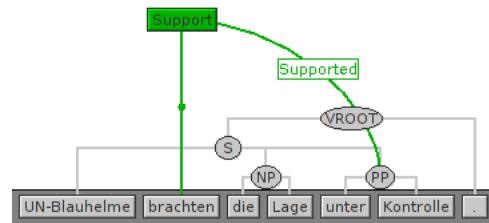


Figure 4: Support verb annotated as verb sense
“UN soldiers brought the situation under control.”

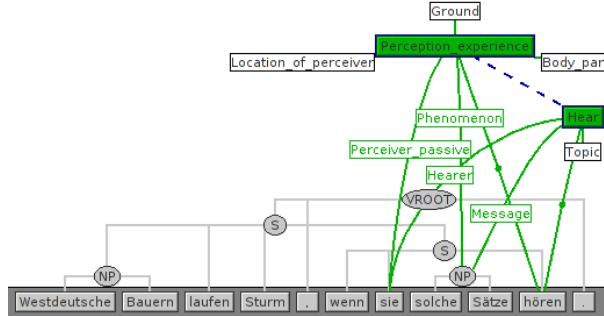


Figure 5: Underspecification in SALSA

the meaning of a predicate (frame element), but fail to represent the whole meaning. The second case applies when it is clear that only one of the two frames can apply, but – due to ambiguity in context – it is not clear which one does represent the intended meaning.

Figure 5 gives an example of frame underspecification where both frames represent a possible meaning of the sentence in (1). It is not clear whether the event of listening was intentional (*Hear*) or whether the farmers are passive listeners who cannot help but undergo the perception of the speech signal (*Perception_experience*).

- (1) Westdeutsche Bauern laufen Sturm , wenn sie solche Sätze hören.
 West-German farmers run storm , when they such sentences hear
 “West-german farmers are up in arms when listening to such sentences.”

Table 7 shows the number of underspecified frames and frame elements for both noun and verb lemmas in SALSA. As expected, the numbers for underspecification are much higher for the verb senses. What comes as a surprise is the large gap between the numbers for verb frames and noun frames (7.34 versus 0.64%). Comparing this with the lower inter-annotator agreement for verbal frames (Table 4), it seems as if the annotators use underspecification as a means to deal with the hard cases, regardless of the fact that it was never intended as such.

SALSA	verbs	nouns
frames	7.34%	0.64%
FEs	1.67%	0.16%

Table 7: Percentage of underspecified frames and frame elements in SALSA 2.0

In the next section we describe how SALSA deals with metaphoric expressions.

5.4 Metaphors

In SALSA, idioms and entrenched metaphorical uses are assigned a frame of their own, as are support uses of many verbal predicates. This is in contrast to WordNet (Miller, 1995) and results in a seemingly higher polysemy when comparing numbers for word senses for both resources (Burchardt et al., 2006).

The entrenchment of metaphors is not always easily ascertained but there are some criteria that are used in combination. One is that entrenched metaphors are not perceived as creative. Another is that entrenched metaphors, especially when they are basic conceptual metaphors, often apply to several frame-mates. For instance, many predicates in the *Change position on a scale* frame such as *rise*, *fall*, *plummet*, *climb* etc. also have uses in the *Motion directional* frame.

Metaphors that do not involve completely entrenched metaphorical meanings are described by a combination of two frames in SALSA: a source frame, expressing the literal meaning of the multi-word, and a target frame describing the understood meaning. This follows the ideas of Lakoff and Johnson (1980) on metaphorical transfer, where a mapping can be defined from the conceptual domain from which the metaphorical expression is drawn (source domain) to the target domain which represents the figurative meaning of the expression. Figure 6 gives an example where the source frame (*Request*) captures the literal meaning of *fordern* (demand, require) and the target frame (*Causation*) expresses the meaning understood by most listeners, namely that there was a CAUSE (the riot) which had an EFFECT (the death of at least four people).⁴

In cases where the target meaning was not easy to capture, only the source frame was annotated. This practice allowed for the annotation to proceed swiftly while, at the same time, assuring that metaphors can be retrieved from the corpus.

⁴Interestingly, this use of German *fordern* is not treated as an established word sense by the Duden online dictionary or by GermaNet (Hamp and Feldweg, 1997), the German counterpart of WordNet, while its usual English translation *claim* has a separate sense in WordNet.

SALSA 2.0	verbs	nouns
	266 (1.3%)	4 (0.02%)

Table 8: Metaphorical expressions annotated in SALSA (numbers in brackets give the percentage of all annotated verbs/nouns annotated as metaphors)

Table 8 gives the number of annotated metaphorical expressions in SALSA. The vast majority of them are evoked by verb lemmas. This, similar to the case of underspecification, might reflect a personal bias of the annotators who created the frames. While the annotator who created the verbal frames tended to flag existing frames as metaphorical, the annotator who created the noun frames had a bias for defining new, more fine-grained frames that captured the metaphorical meaning of the expression. It is not clear to us whether either of the approaches has a significant advantage over the other.

Finally, the above discussion of metaphor does not capture another class of cases. Some frames in SALSA, though inspired by existing FrameNet frames, were created as proto-frames. Here, the differences between SALSA and FrameNet do not concern the Frame elements but typically involve some notion of ‘generalization’. For instance, the verb lemma *ablegen* has a proto-frame *ablegen1-salsa*, which is described as follows:

An Agent causes a Theme to leave a location, the Source. Unlike in the non-generalized version of this frame, neither the location nor the Theme need be a location in the literal sense. The Source is profiled by the words in this frame, just as the Goal is profiled in the Placing frame.

As the definition suggests this proto-frame targets uses of the lemma that talk about things like “getting rid of [lit. doffing, taking off] an image or a habit”. A possible analysis of these cases is that they involve metaphor rather than merely more general meanings. These frames are not explicitly marked but their definition typically refers to the ‘generalization’ of meaning in the frame relative to an existing FrameNet frame or to another proto-frame.

6 Discussion

Having outlined the different approaches of FrameNet and SALSA, a question that naturally comes to mind is which impact these differences have on practical applications. Regarding coverage, SALSA seems to be more domain-specific, as only vocabulary from the newspaper domain has been dealt with, while FrameNet provides more general frames but is to be expected to have coverage gaps on newstext. Underspecification, as applied in SALSA to deal with ambiguous instances, might result in ‘harder’ training and test sets for machine learning applications, while the prototypical instances in FrameNet might be easier to classify. It is still unclear which effect these different training sets will have when used as training data for Semantic Role Labelling systems that are to be applied to new text domains.

7 Conclusions

We presented Release 2.0 of the SALSA corpus, which provides frame-semantic annotations for German nouns and verbs. The corpus now contains more than 36,000 annotated instances from the newspaper domain. In the paper we described the workflow in SALSA, discussed our efforts to ensure annotation quality and reported inter-annotator agreement for frame and frame element annotations. The core of our discussion then focused on methodological choices made in SALSA and compared them to the approach taken by FrameNet. The SALSA corpus is freely available⁵ and can be used as training data for semantics-like NLP applications as well as for linguistic studies in lexical semantics.

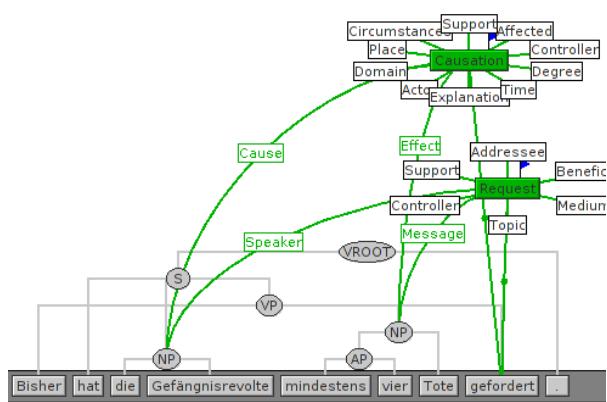


Figure 6: Annotation of metaphors in SALSA
“Up to now, the prison riot has claimed at least 4 lives.”

⁵<http://www.coli.uni-saarland.de/projects/salsa/>

Acknowledgments

This work was funded by the German Research Foundation DFG (grant PI 154/9-3). We gratefully acknowledge the work of our annotators, Lisa Fuchs, Andreas Rebmann, Gerald Schoch and Corinna Schorr. We would also like to thank the anonymous reviewers for helpful comments.

References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. Association for Computational Linguistics.
- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, Sofia, Bulgaria.
- Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.
- Katrin Erk and Sebastian Padó. 2007. Towards a computational model of gradience in word sense. In *Proceedings of IWCS-7*, Tilburg, The Netherlands.
- Charles J. Fillmore, 1982. *Linguistics in the Morning Calm*, chapter Frame Semantics, pages 111–137. Hanshin Publishing, Seoul.
- B. Hamp and H. Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.
- Ulrich Heid. 1998. Towards a corpus-based dictionary of german noun-verb collocations. In *Proceedings of the EURALEX International Congress*, Liège, Belgium.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.
- Igor Mel’čuk. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical functions in lexicography and natural language processing*, volume 31, pages 37–102. John Benjamins, Amsterdam/Philadelphia.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.
- Margarita Alonso Ramos, Owen Rambow, and Leo Wanner. 2008. Using semantically annotated corpora to build collocation resources. In *Proceedings of Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Semantic analysis in word vector spaces with ICA and feature selection

Tiina Lindh-Knuutila and Jaakko J. Väyrynen and Timo Honkela

Aalto University School of Science

Department of Information and Computer Science

PO Box 15400, FI-00076 AALTO, Finland

firstname.lastname@aalto.fi

Abstract

In this article, we test a word vector space model using direct evaluation methods. We show that independent component analysis is able to automatically produce meaningful components that correspond to semantic category labels. We also study the amount of features needed to represent a category using feature selection with syntactic and semantic category test sets.

1 Introduction

The concept of semantic similarity or the broader concept of semantic relatedness is central in many NLP-related applications. The representations that are used range from thesauri to vector space models (Budanitsky and Hirst, 2006). Semantic similarity can be measured as a distance between representations of words; as vector similarity in a vector space model, or as a path length in a structured representation e.g., an ontology. For humans, the notion of semantic similarity often means perceiving two words sharing similar traits: synonyms like *car:automobile*, hypernyms *vehicle:car* or antonyms *short:long* (Turney and Pantel, 2010).

Earlier research has shown that it is possible to learn automatically linguistic representations that reflect syntactic and semantic characteristics of words. In particular, independent component analysis (ICA) can be used to learn sparse representations in which components have a meaningful linguistic interpretation. However, earlier results cannot be considered conclusive especially when semantic relations and semantic similarity

are considered. We present results of a systematic analysis that focuses on semantic similarity using manually built resources as a basis for evaluation of automatically generated vector space model (VSM) representation. We concentrate on word vector spaces based on co-occurrence data. These can be evaluated directly by comparing distances between word pairs or groups of words judged similar by human evaluators. In this article, we describe several test sets used for semantic evaluation of vector space models, and validate our model with them. We then explore how many features are required to distinguish the categories with a feature selection algorithm. Further, we measure how well ICA is able to automatically find components that match semantic categories of words.

2 Methods

2.1 Vector space models

VSMs contain distributional information about words derived from large text corpora. They are based on the idea that words that appear in similar contexts in the text are semantically related, and that relatedness translates into proximity of the vector representations (Schütze, 1993). In information retrieval and many other tasks, topic representations using term-document matrices are often employed. In the same fashion, word vector spaces are built using the more immediate context of a word. Similarity measures for vector spaces are numerous. For VSMs, they have been extensively tested for example by Bullinaria and Levy (2007). In this work, we use the cosine similarity (Landauer and Dumais, 1997), which is

most commonly used (Turney and Pantel, 2010).

The simple way of obtaining a raw word co-occurrence count representation for N target words is to consider C context words that occur inside a window of length l positioned around each occurrence of the target words. An accumulation of the co-occurrences creates a word-occurrence matrix $X_{C \times N}$. Different context sizes yield representations with different information. Sahlgren (2006) notes that small contexts (of a few words around the target word), give rise to paradigmatic relationships between words, whereas longer contexts find words with syntagmatic relationship between them. For a review on the current state of the art for vector space models using word-document, word-context or pair-pattern matrices using singular value decomposition-based approaches in dimensionality reduction, see Turney and Pantel (2010).

2.2 Word spaces with SVD, ICA and SENNA

The standard co-occurrence vectors for words can be very high-dimensional even if the intrinsic dimensionality of word context information is actually low (Karlgren et al., 2008; Kivimäki et al., 2010), which calls for an informed way to reduce the data dimensionality, while retaining enough information. In our experiments, we apply two computational methods, singular value decomposition (SVD) and ICA, to reduce the dimensionality of the data vectors and to restructure the word space.

Both the SVD and ICA methods extract components that are linear mixtures of the original dimensions. SVD is a general dimension reduction method, applied for example in latent semantic analysis (LSA) (Landauer and Dumais, 1997) in the linguistic domain. The LSA method represents word vectors in an orthogonal basis. ICA finds statistically independent components which is a stronger requirement and the emerging features are easier to interpret than the SVD features (Honkela et al., 2010).

Truncated SVD approximates the matrix $X_{C \times N}$ as a product UDV^T in which $D_{d \times d}$ is a diagonal matrix with square roots of the d largest eigenvalues of $X^T X$ (or XX^T), $U_{C \times d}$ has the d corresponding eigenvectors of XX^T , and $V_{N \times d}$ has the d corresponding eigenvectors of $X^T X$.

The rows of $V_{N \times d}$ give a d -dimensional representation for the target words.

ICA (Comon, 1994; Hyvärinen et al., 2001) represents the matrix $X_{C \times N}$ as a product AS , where $A_{C \times d}$ is a mixing matrix, and $S_{d \times N}$ contains the independent components. The columns for the matrix $S_{d \times N}$ give a d -dimensional representation for the target words. The FastICA algorithm for ICA estimates the model in two stages: 1) dimensionality reduction and whitening (decorrelation and variance normalization), and 2) rotation to maximize the statistical independence of the components (Hyvärinen and Oja, 1997). The dimensionality reduction and decorrelation step can be computed, for instance, with principal component analysis or SVD.

We compare the results obtained with dimension reduction to a set of 50 feature vectors from a system called SENNA (Collobert et al., 2011)¹. SENNA is a labeling system suitable for several tagging tasks: part of speech tagging, named entity recognition, chunking and semantic role labeling. The feature vectors for a vocabulary of 130 000 words are obtained by using large amounts of unlabeled data from Wikipedia. In training, unlabeled data is used in a supervised setting. The system is presented a target word in its context of 5+5 (preceding+following) words with a 'correct' class label. An 'incorrect' class sample is constructed by substituting the target word with a random one and keeping the context otherwise intact. The results in the tagging tasks are at the level of the state of the art, which is why we want to compare these representations with the direct evaluation tests.

2.3 Direct evaluation

In addition to indirect evaluation of vector space models in applications, several tests for direct evaluation of word vector spaces have been proposed see e.g., Sahlgren (2006) and Bullinaria and Levy (2007). First we describe the semantic and syntactic category tests. Here, a *category* means a group of words with a given class label. The precision P in the category task is calculated according to (Levy et al., 1998). A centroid for each category is calculated as an arithmetic mean of the word vectors belonging to that category.

¹<http://ronan.collobert.com/senna/>

The distances from each word vector to all category centroids are then calculated, recalculating the category centroid for the query to exclude the query vector. The precision is then the percentage of the words for which the closest centroid matches the category the word is labeled with.

The semantic category test (Semcat) set² is used for example in (Bullinaria and Levy, 2007). This set contains 53 categories with 10 words in each category, based on the 56 categories collected by Battig and Montague (1969). Some word forms appear in more than one category, for example *orange* in FRUIT and in COLOR, and *bicycle* in TOY, and in VEHICLE. We made some slight changes to the test set by changing the British English spelling of a limited number of words back into American English (e.g., *millimetre-millimeter*) to better conform to the English used in the Wikipedia corpus.

We also consider two different syntactic category test alternatives. Bullinaria *et al.* (1997; 2007) use ten narrow syntactic categories, separating noun and verb forms in own categories, whereas Sahlgren (2006) uses eight broad categories. In this article, we compare both of these approaches. As neither of these test sets were publicly available, we constructed our own applying the most common part-of-speech (POS) tags from the Penn Treebank tag set (Marcus *et al.*, 1993) to 3 000 most frequent words in our vocabulary. We call the built test sets Syncat1 and Syncat2. In Syncat1, the 50 most frequent words in 10 narrow POS categories: Singular or mass nouns (NN), plural nouns (NNS), singular proper nouns (NNP), adjectives in base form (JJ), adverbs in base form (RB), verbs in base form (VB), verbs in past participle form (VBN), verbs in ing-form (VBG), cardinal numbers (CD), and prepositions or subordinating conjunctions (IN). In (Levy *et al.*, 1998) the last category contains only prepositions, but the Penn Treebank tagset does not separate between subordinating conjunctions and prepositions. Syncat2 contains 20 most frequent words in seven broader POS categories: nouns, verbs, adjectives, adverbs, prepositions, determiners, and conjunctions. In the open categories; nouns, verbs, adjectives and adverbs, the words can be in any of the tagged forms. The

original experiments (Sahlgren, 2006) contain a category named 'conjunctions', which we created by combining the aforementioned IN-category which contains also prepositions with coordinating conjunctions (CC). The interjection category (UH) from the original work was left out due to the infrequency of such words in the Wikipedia corpus.

In addition to category tests, synonymy can also be directly measured, for example with a multiple choice test, where synonyms should be closer to each other than the alternatives. The most commonly used test for VSMs is the Test of English as a Foreign Language (TOEFL) (Landaier and Dumais, 1997), although other similar tests, such as English as a Second Language (ESL) and SAT college entrance exam (Turney, 2001; Turney, 2005), are also available. In addition, one can easily construct similar multiple-choice tests based on, for example, thesauri and random alternatives. Bullinaria *et al.* (1997; 2007; 2012) use a test which consists of related word pairs, e.g., *thunder-lightning*, *black-white*, *brother-sister*. The distance from a cue word to the related word is then compared to randomly picked words with a similar frequency in the corpus³. In addition to semantic similarity, Deese antonym pairs (Deese, 1954), have been used for VSM evaluation (Grefenstette, 1992). We employ a similar procedure described above for the related words. The precision is calculated by checking how often the cue and the correct answer are closest – with a comparison to eight randomly picked words for each cue word.

2.4 Forward feature selection with entropy

The forward feature selection method is a simple greedy algorithm. At each step, the algorithm selects the single feature that best improves the result measured by an evaluation criteria, without ever removing already selected features. The algorithm stops when all features are selected, or a stopping criterion is triggered. Compared to an exhaustive feature search, which has to evaluate all possible feature combinations, only $d(d+1)/2$ different input sets have to be evaluated, where d is the desired number of features (Sorjamaa, 2010). Reaching the global optimum is not guar-

²<http://www.cs.bham.ac.uk/~jxb/corpus.html>

³<http://www.cs.bham.ac.uk/~jxb/corpus.html>

anteed because the forward feature selection algorithm can get stuck in a local minimum. However, an exhaustive search is not computationally feasible given the large number of features in our case.

In our experiments, Shannon’s entropy, based on category distributions in cluster evaluation, is our selected criterion for feature selection (Celeux and Govaert, 1991). The number of clusters is set equal to the number of categories, with the cluster centroid as the centroid for words in the category. Each word is assigned to the cluster with the closest cluster centroid (Tan et al., 2005, Ch. 8, p. 549). The entropy is highest if each cluster contains only one word from each category, and lowest when each cluster contains only words from one category. In our feature selection task, we measure a single category against all other categories and have only two clusters. We then compare the performance of provided categories (class labels) and groups of words selected randomly.

3 Data and preprocessing

Our data consists of all the documents in the English Wikipedia⁴ that are over 2k in size after removal of the wikimedia tags. In preprocessing, all words are lowercased and punctuation is removed, except for hyphens and apostrophes. The vocabulary consists of 200 000 most frequent types. The context vocabulary used to build the word vectors consists of 5 000 most frequent types. The total number of times the 200 000 types occur together in the corpus is slightly over 326 million, and the least frequent word occurs 26 times in the corpus. The least frequent context word occurs 6 803 times in the corpus.

In previous work, small windows have corresponded the best to paradigmatic relationship between words, which is what most of the direct evaluation tests described above measure. This is why we use a small window of three ($l = 3$) in the experiments, which corresponds to capturing the previous and the following word around a target word. The word frequency in the whole corpus biases the raw co-occurrence frequency counts. Several weighting schemes that dampen the ef-

⁴The October 2008 edition, which is no longer available at the Wikipedia dump download site <http://dumps.wikimedia.org/enwiki/>

fect of the most frequent words have been proposed. We use positive point-wise mutual information (PPMI) weighting (Niwa and Nitta, 1994), which gave best results in the evaluation tests by Bullinaria and Levy (2007).

4 Experiments and results

4.1 Direct VSM evaluation

We first validate our model with the syntactic and semantic tests (Semcat, Syncat1, Syncat2, TOEFL, related word pairs (Distance), and Deese antonyms (Deese)). The results for the semantic and syntactic tests are summarized in Table 1. The results for the our VSM for the semantic categorization, distance, and TOEFL tests are in line with results for the same number of features in (Bullinaria and Levy, 2012). The SENNA results for the syntactic tests beat our simple system and even though the training does not contain semantic information, the results in the semantic test are only slightly worse than our results, which may also be due to the fact that the system has a larger context window. We also reduced the dimensionality of the vector space to 50 with SVD and ICA, and performed the testing. Using only 50 dimensions is not enough to capture all the meaningful information from the 5 000 original dimensions, compared to the 50 dimensions of SENNA, but the SVD and ICA results can be produced in about 10 minutes, whereas SENNA training takes weeks (Collobert et al., 2011). To obtain results equivalent to those with the 5 000 features, we tested a growing number of features: using approximately 500 ICA or SVD components would be enough. We repeated the experiments for the Semcat test and only took those word vectors that represented the 530 words of the test set. In this case, ICA and SVD are able to better represent the dimensions of the interesting subset instead of the whole vocabulary of 200 000 words, which makes the error decrease considerably.

ICA and SVD perform equally well in dimensionality reduction. This is not surprising, as ICA can be thought as (1) whitening and dimensionality reduction followed by (2) an ICA rotation, where (1) is usually computed with PCA. As (2) does not change distances or angles between vectors, any method that utilizes distances or an-

gles between vectors finds very little difference between PCA and ICA. As PCA and SVD are closely related, the same reasoning can be applied there. (Vicente et al., 2007)

4.2 Feature selection

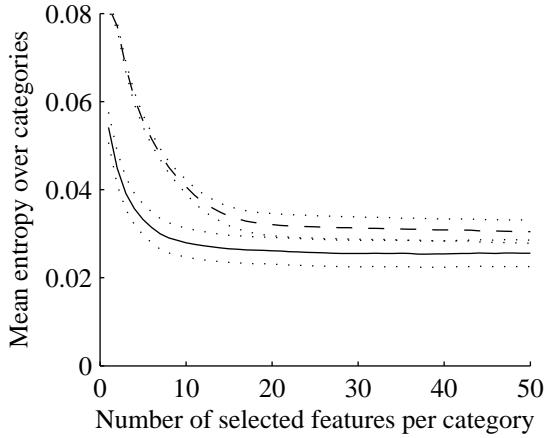


Figure 1: The average means with 95% confidence intervals for entropy for the 53 semantic categories (lower curve) and random categories (upper curve).

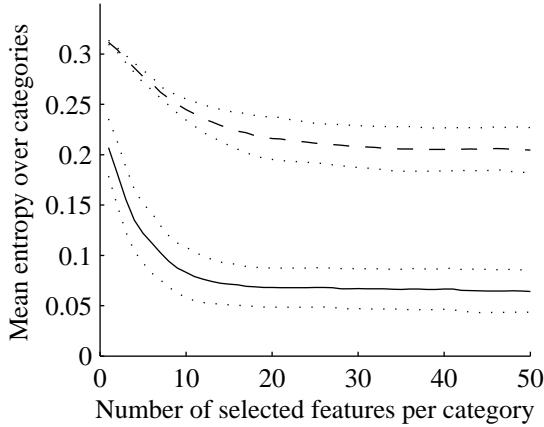


Figure 2: The average means with 95% confidence intervals for entropy for the 10 syntactic categories of Syncat1 (lower curve) and random categories (upper curve).

Figures 1, 2 and 3 show the results for the semantic and syntactic feature selection experiments. To help the visualization, only the confidence intervals around the mean of the semantic/syntactic categories and random categories are shown. The results indicate that each category can be easily separated from the rest by a few features

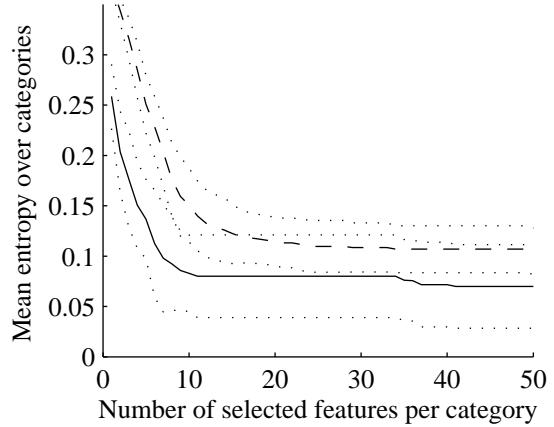


Figure 3: The average means with 95% confidence intervals for entropy for the 7 syntactic categories of Syncat2 (lower curve) and random categories (upper curve).

only. The randomly constructed categories cannot be as easily separated, but the difference to semantic categories diminishes as more features are added.

Syncat1, where each category corresponds to a single part of speech, gives very good separation results, whereas the separation in the case of Syncat2 is less complete. The separation of adjective, adverb and determiner categories cannot be distinguished from the results of random categories. Looking more closely at the most common words labeled as adjectives, we see that most of them can be also used as adverbs, nouns or even verbs, e.g., *first, other, more, such, much, best*, and unambiguous adjectives e.g., *british, large, early* are a minority. Similar reasoning can be applied to the adverbs. Determiners, on the other hand, may exist in so many different kind of contexts, and finding few context words (i.e. features) to describe them might be difficult.

We also looked at the first context word which was selected for each semantic category, and in 40 cases out of 53 the first selected feature was somehow related to the word of the category. We found out that it was either a semantically related noun or verb: A PRECIOUS STONE:*ring*, A RELATIVE:*adopted*; the name or part of a name for the category: A CRIME:*crime*, AN ELECTIVE OFFICE:*elected*; or a word that belongs to that category: A KIND OF CLOTH:*cotton*, AN ARTICLE OF FURNITURE:*table*.

	5 000 feat	ICA50	SVD50	SENNA
Semcat	0.22	0.31/0.19	0.32/0.19	0.25 (R=0.98)
Syncat1	0.17	0.25	0.25	0.10
Syncat2	0.26	0.38	0.37	0.21
TOEFL	0.22 (R=0.95)	0.38 (R=0.95)	0.38 (R=0.95)	0.34 (R=0.91)
Distance	0.11	0.19	0.19	0.11
Deese	0.07	0.12	0.13	0.04

Table 1: The error, $Err = 1 - P$ for different test sets and data sets. Recall is 1 unless otherwise stated. For the Distance test the values reported are mean values over 50 runs. For ICA and SVD, first values for Semcat report the error when the dimensionality was reduced from the full $200\,000 \times 5\,000$ whereas the second values are calculated for the Semcat subset $530 \times 5\,000$ only.

4.3 Feature selection with ICA

As previous results suggest (Honkela et al., 2010), ICA can produce components that are interpretable. An analysis using 50 independent components obtained by ICA was carried out for the vector representations of the 530 words in the Semcat data set, and forward feature selection was then applied for each of the categories. The results are shown in Fig. 4. The semantic categories separate better than the random categories with only a few features in this experiment as well, but as more features are added, the difference decreases. In Fig. 4 we show the reader also some examples of the semantic categories for which the entropy is smallest and largest.

In this experiment, we used 10-fold stratified cross validation, choosing 90 % of the 530 words as the training set and the rest was used as the test set. I.e., when separating one category from all the rest, 9 words were used as the representative of the tested category, 468 words as the representative of the other categories from which we separate, and the remaining 10% as the test data set in the same relation. Reported results are the averaged entropy over the different folds of the cross validation. Between different folds, a few same features were always chosen first, after which there was considerable variation. This seems to indicate that the first 2-3 features selected are the most important when a category is separated from the rest. The results in case of the SENNA data and SVD, left out due to space constraints are similar.

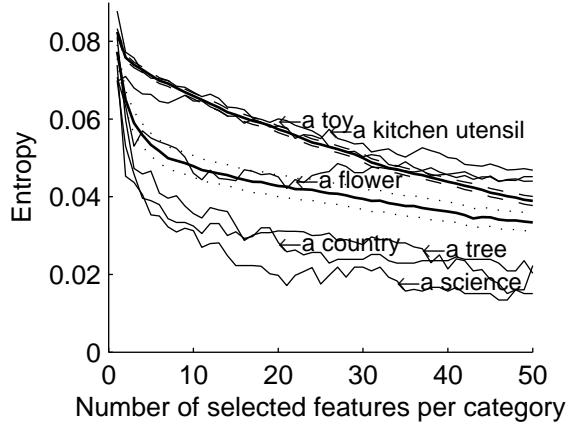


Figure 4: Entropy for sample semantic categories using ICA components in feature selection, mean entropy for random categories with 95% confidence interval shown on dashed lines and mean entropy over semantic categories with 95% confidence interval shown with dotted lines over random categories.

4.4 Semantic information from independent components

Earlier research has shown that ICA is able to produce cognitively meaningful components (Hansen et al., 2005) that correspond to, for instance, noun concepts (Chagnaa et al., 2007), phonological categories (Calderone, 2009), personal traits (Chung and Pennebaker, 2008) and syntactic categories (Honkela et al., 2010). In this work, we studied the interpretability of the ICA components in a semantic task, using the Semcat set, and compare the results to those obtained by SVD and SENNA. For this, we study the activations of the components in the S -matrix. For each component, we take a group of ten words, for which the component has the highest activation

and compare that group to the given categories.

The ICA components are usually skewed in one direction. To see which words have most activation, it is often enough to check the skewness of the distribution. We then selected the 10 words for which the component activation was largest in that direction. We applied two criteria, *strict* and *lax* (similar to Sahlgren (2006)) to see how well the components corresponded to the semcat category labels. To fill the *strict* criterion a minimum of 9/10 words should belong to the same category, and the *lax* criterion is defined as majority, i.e., a minimum of 6/10 words must belong to the same category.

The selection of the subset of word vectors can be done either before or after dimensionality reduction. It is obvious that if ICA or SVD is applied after the subset selection (subset+ICA), the components are a better representation of only those words, than if the subset selection is carried out after a dimension reduction for the complete matrix of 200 000 word vectors (ICA+subset). As FastICA produces results which may vary from a run to another due to random initialization, we verified that the results stayed consistent on subsequent runs.

The results are shown in Table 2. In subset+dimensionality reduction case, ICA is able to find 17 categories out of 53 with the strict criterion, and 37 categories with the lax criterion. Those categories that filled the strict criterion had also the smallest entropy in the feature selection described earlier. We repeated the above-described analysis evaluating separately the 10 smallest and 10 largest values of each SVD component. SVD found found only two categories which passed the strict test. For the relaxed condition, 19 categories passed. In the dimensionality reduction+subset case, ICA is slightly better with the lax criterion, whereas SVD finds three categories with the strict criterion. These results can also be compared to SENNA results, in which no categories passed the strict test, and 4 categories passed the lax test.

As a separate experiment for subset + dimension reduction, we checked whether the features that best represented these categories were also those that were first selected by the feature selection algorithm. We found out that for 15 cate-

gories, the most prominent feature was also selected first and for the two remaining categories, the feature was selected second. For the relaxed condition, for the 37 categories, the best feature was selected first in 27 cases and second in 6 categories, and third in 2 categories.

	Strict	Lax
subset+ICA	17/53	37/53
ICA+subset	1/53	12/53
subset+SVD	2/53	19/53
SVD+subset	3/53	8/53
SENNNA	0/52	4/52

Table 2: Fraction of categories which filled the strict and lax condition for ICA, SVD and SENNA

A further analysis of the categories that failed the relaxed test suggests several reasons for this. A closer analysis shows that words from certain categories tend to occur together, and these categories contain a common superordinate category: For example NONALCOHOLIC and ALCOHOLIC BEVERAGES are all beverages. Among the top 10 activations of a component, there are five words from each of these categories, and among 20 highest activations for this component, 18 of them come from one of these two categories. Similarly, a component shows high activations for words that form the female half of the RELATIVE: *aunt*, *sister*, *mother*, together with *daisy*, *tulip*, *rose*, *lily* from FLOWER, which are also used as female names. There are more overlapping categories in which words may also belong to another category than for which they are assigned, for example TYPE OF DANCE and TYPE OF MUSIC; SUBSTANCE FOR FLAVORING FOOD and VEGETABLE; the TYPE OF SHIP and VEHICLE; and FOOTGEAR and ARTICLE OF CLOTHING and the TYPE OF CLOTH. These results are in line with the earlier result with the feature selection, where unambiguous categories separated better than the more ambiguous ones. There are four categories which cannot be described by a single component in any way: KITCHEN UTENSIL, ARTICLE OF FURNITURE, CARPENTER'S TOOL and TOY. The words in these categories have activations in different ICA components, which suggests that the most common usage is not the one invoked by the given category. For example, in TOY category,

words *bicycle* and *tricycle* go with the words from VEHICLE and *block* has an active feature which also describes PARTS OF A BUILDING or FURNITURE.

5 Discussion

The semantic category test set is based on studies of human similarity judgment, which for even with a large group of responses, is quite subjective. The ICA analysis shows that meaning of the surface forms for some words based on the corpus data are different than the one it is labeled with. For example, *bass* from the FISH category had a strong activation for a feature which represented MUSICAL INSTRUMENT. This is obviously a downside of our method which relies on the bag-of-words representation without taking into account the sense of the word.

We saw that the dimension reduction applied as drastically as we did worsens the evaluation results considerably. The 50-dimensional feature vectors of SENNA produce better results in many of the tasks excluding the TOEFL and the semantic categorization, but definite conclusions on the performance cannot be made, as the SENNA is not trained with the exactly same data. Another downside of SENNA is the very long training. In this paper, we opted to have the simplest word space without taking into account word senses, elaborate windowing schemes or such. The current paper does not address the feature selection as a means for reducing the dimensionality as such, but it is an interesting direction for future work. Karlsgren et al. (2008) suggest studying the local dimensionality around each word, as most vectors in a high-dimensional vector space are in an orthogonal angle to each other. We found out that the first features are most important in representing a semantic category of 10 words, and an unreported experiment with 300 ICA features showed that the features included last had a negative impact to the separation of the categories. Cross-validation results showed that the selected ICA features were also useful with a held-out set.

6 Conclusions

This paper describes direct evaluation tests for word vector space models. In these tests, ICA

and SVD perform equally well as dimensionality reduction methods. Further, the work shows that only a small number of features was needed to distinguish a group of words forming a semantic category from others. Our experiments with the random categories show that there is a clear difference between the separability between most of the semantic categories and the random categories. We found the gap surprisingly large.

Some of the semantic categories separated very badly, which were analyzed to stem from differences in frequency for the different senses of the word collocations. Our premise is that a good latent word space should be able to separate different cognitively constructed categories with only a few active components, which is related to the sparse coding generated by ICA. Further, we have shown that we could find interpretable components that matched semantic categories reasonably well using independent component analysis. Compared to SVD, ICA finds a fixed rotation where the components are also maximally independent, and not only uncorrelated. This facilitates the analysis of the found structure explicitly, without relying on implicit evaluation methods. The interpretability of the ICA components is an advantage over SVD, demonstrated by the quantitative *strict/lax* evaluation.

The main motivation of this work is to support the development towards automatic processes for generating linguistic resources. In this paper, we focus on independent component analysis to generate the sparse linguistic representations, but similar conclusions can be made with closely related methods, such as non-negative matrix factorization (NMF).

Acknowledgments

We gratefully acknowledge the support from the Department of Information and Computer Science at Aalto University School of Science. In addition, Tiina Lindh-Knuutila was supported by the Finnish Cultural Foundation and Jaakko J. Väyrynen and Timo Honkela were supported by the META-NET Network of Excellence.

References

- W.F. Battig and W.E. Montague. 1969. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80(3, part 2.):1–45.
- A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.
- J.A. Bullinaria and J.P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.
- J.A. Bullinaria and J.P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44.
- B. Calderone. 2009. Learning phonological categories by independent component analysis. *Journal of Quantitative Linguistics*, 17(2):132–156.
- G. Celeux and G. Govaert. 1991. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176.
- A. Chagnaa, C.-Y. Ock, C.-B. Lee, and P. Jaimai. 2007. Feature extraction of concepts by independent component analysis. *International Journal of Information Processing Systems*, 3(1):33–37.
- C. K. Chung and J. W. Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Research in Personality*, 42(1):96–132.
- R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12:2493–2537.
- P. Comon. 1994. Independent component analysis—a new concept? *Signal Processing*, 36:287–314.
- J.E. Deese. 1954. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.
- G. Grefenstette. 1992. Finding the semantic similarity in raw text: The Deese antonyms. Technical Report FS-92-04, AAAI.
- Lars Kai Hansen, Peter Ahrendt, and Jan Larsen. 2005. Towards cognitive component analysis. In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 148–153, Espoo, Finland, June.
- T. Honkela, A. Hyvärinen, and J.J. Väyrynen. 2010. WordICA — emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16:277–308.
- A. Hyvärinen and E. Oja. 1997. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492.
- A. Hyvärinen, J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. John Wiley & Sons.
- J. Karlsgren, A. Holst, and M. Sahlgren. 2008. Filaments of meaning in word space. In *Proceedings of the European Conference on Information Retrieval*, pages 531–538.
- I. Kivimäki, K. Lagus, I.T. Nieminen, J.J. Väyrynen, and T. Honkela. 2010. Using correlation dimension for analysing text data. In *Proceedings of ICANN 2010*, pages 368–373. Springer.
- T.K. Landauer and S.T. Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.
- J.P. Levy, J.A. Bullinaria, and M. Patel. 1998. Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology*, 10:99–111.
- M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Y. Niwa and Y. Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 304–309.
- M. Patel, J.A. Bullinaria, and J.P. Levy. 1997. Extracting semantic representations from large text corpora. In *Proceedings of Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 199–212. Springer.
- M. Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University, Department of Linguistics.
- H. Schütze. 1993. Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.
- A. Sorjamaa. 2010. *Methodologies for Time Series Prediction and Missing Value Imputation*. Ph.D. thesis, Aalto University School of Science.
- P.-N. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
- P.D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings*

- of the Twelfth European Conference on Machine Learning (EMCL2001)*, pages 491–502, Freiburg, Germany. Springer-Verlag.
- P.D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint conference on Artificial intelligence (IJCAI-05)*, pages 1136–1141.
- M. A. Vicente, P. O. Hoyer, and A. Hyvärinen. 2007. Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(5):896–900.

Aggregating Skip Bigrams into Key Phrase-based Vector Space Model for Web Person Disambiguation

Jian Xu, Qin Lu, Zhengzhong Liu

Department of Computing

The Hong Kong Polytechnic University, Hong Kong

{csjxu, csluqin, hector.liu}@comp.polyu.edu.hk

Abstract

Web Person Disambiguation (WPD) is often done through clustering of web documents to identify the different namesakes for a given name. This paper presents a clustering algorithm using key phrases as the basic feature. However, key phrases are used in two different forms to represent the document as well context information surround the name mentions in a document. In using the vector space model, key phrases extracted from the documents are used as document representation. Context information of name mentions is represented by skip bigrams of the key phrase sequences surrounding the name mentions. The two components are then aggregated into the vector space model for clustering. Experiments on the *WePS2* datasets show that the proposed approach achieved comparable results with the top 1 system. It indicates that key phrases can be a very effective feature for WPD both at the document level and at the sentential level near the name mentions.

1 Introduction

Most of current search engines are not suited for web persons disambiguation because only pages related to the most popular persons will be easily identified. Web Persons disambiguation (WPD) targets at identifying the different namesakes for a given name (Artiles et al., 2010). Normally WPD involves two steps. The first step uses clustering methods to cluster different namesakes and the second step works on each cluster to extract the descriptive attributes of each namesake to form their profiles. This paper focuses on the clustering algorithms in WPD.

Most of the previous researches attempted to use a combination of different features such as, tokens, named entities, URL or title tokens, n-gram features, snippets and other features (Chen et al., 2009; Chong et al., 2010). Traditionally, document clustering based on a single representation space using the vector space model (VSM) is often the choice (Salton and McGill, 1983). However, how to find a good balance between the selection of a rich set of features and degradation performance due to more noise introduced is an important issue in VSM.

This paper presents a clustering algorithm based on using key phrases only. The use of key phrases is based on the hypothesis that key phrases, or sometimes referred to as topic words (Steyvers and Griffiths, 2007), are better semantic representations of documents (Anette Hulth, 2003). We also argue that the key phrases surrounding the name mentions can represent the context of the name mentions and thus should be considered as another feature. This is an important distinction on clustering for WPD compared to the purpose of other document clustering algorithms. In this paper, key phrases are thus used in two parts. In the first part, key phrases are used as the single feature to be represented by the VSM for clustering. In the second part, the key phrases in a sequential representation surrounding a name mention are identified using skip bigrams. Finally, the skip-bigrams are concatenated to the bag of key phrase model to serve as the aggregated key phrase-based clustering (AKPC) algorithm.

For key phrase extraction, a supervised learning algorithm is used and trained through the English Wikipedia personal article pages so as to avoid laborious manual annotation. To

incorporate the context information at sentential level into WPD, the name mentions in the document are first located and then the key phrases that surround the name mentions are extracted. These key phrases are arranged into sequences from which the skip bigrams are then extracted.

Different from the previous skip bigram statistics which considers pairs of words in a sentence order with arbitrary gaps (Lin and Och, 2004a) and compares sentence similarities through the overlapping skip bigrams, the skip bigrams in this paper are weighted by an exponentially decay factor of their full length in the sequence, hence emphasizing those occurrences of skip bigrams that has shorter skips (Xu et al., 2012). It is reasonable to assume that if two sentences are similar, they should have many overlapping skip bigrams, and the gaps in their shared skip bigrams should be similar as well. Besides, a different weighting scheme for skip bigrams in this paper is used. It combines the penalizing factor with the length of gaps, named skip distance (SD). The longer the skip distance is, the more discount will be given to the skip bigrams.

The rest of this paper is organized as follows. Section 2 describes the related works of web person disambiguation. Section 3 presents key phrase extraction algorithm. Section 4 describes the skip bigrams. Section 5 gives the performance evaluation of the aggregated key phrase-based clustering (AKPC) algorithm. Section 6 is the conclusion.

2 Related Work

Web Person Disambiguation, as a task was defined and contested in the WePS workshops 2007, 2009, 2010 (Artiles et al., 2007, 2009, 2010). In WePS workshops, both development data (including training data and golden answer) and testing data are provided. The searched results include snippets, ranking, document titles, their original URLs and HTML pages (Artiles et al., 2009).

Some harvested the tokens from the web pages external to the WePS development data (Chen et al., 2009; Han et al., 2009), and others used named entities (Popescu et al., 2007). Some

algorithms used external resources such as Google 1T corpus and Wikipedia to tune the weighting metrics. For example, Chen et al. (2009) used the Google 1T 5-gram data to learn the bigram frequencies. Chong et al. (2010) used Wikipedia to find phrases in documents.

Key phrases give a semantic summarization of documents and are used in text clustering (Hammouda et al., 2005), text categorization (Hulth and Megyesi, 2006) and summarization (Litvak and Last, 2008). For key phrase extraction, supervised and unsupervised approaches are both commonly used. Wan and Xiao (2008) proposed the CollabRank approach which first clustered documents and then used the graph-based ranking algorithm for single document key phrase extraction. Zha (2002) applied the mutual reinforcement principle to extract key phrases from a sentence based on the HITS algorithm. Similarly, Liu et al. (2010) considered the word importance related to different topics when ranking key phrases. Li et al. (2010) proposed a semi-supervised approach by considering the phrase importance in the semantic network. Frank et al. (1999) and Witten et al. (2000) used the Naive Bayes approach to extract key phrases with known key phrases. Similarly, Xu et al. (2012) proposed to use the anchor texts in Wikipedia personal articles for key phrase extraction using the Naive Bayes approach.

Skip bigram statistics are initially used to evaluate machine translation. It measures the overlap between skip bigrams between a candidate translation and a set of reference translations (Lin and Och, 2004a). The skip bigram statistics uses the ordered subsequence of words as features for sentence representation in machine translation evaluation. It counts the matches between the candidate translation and a set of reference translations. However, there is no attempt to use key phrases to create skip bigrams for WPD.

3 Key Phrase Extraction

In VSM based clustering, different algorithms use different set of features to represent a document such as tokens, name entities (Popescu et al., 2007; Chen et al., 2009; Han et al., 2009).

The choice of features directly affects both the performance and the efficiency of their algorithms. Simple features can be more efficient, but may suffer from data sparseness issues. However, more features may also introduce more noise and degrade the performance and efficiency of the algorithm. This paper investigates the use of key phrase as the single feature for WPD. Key phrases are similar to topic words used in other applications as semantic representations (Steyvers and Griffiths, 2007). The difference is that key phrases are bigger in granularity, which can help reduce the dimensionality of the data representation. More importantly, key phrases are better representation of semantic units in documents. For example, the key phrase *World Cup* denotes the international football competition, if it is split into *World* and *Cup*, its semantical meaning would be altered.

Extraction of key phrases can take different approaches. If training data is available, some learning algorithms can be developed. However, annotation is needed to prepare for the training data which can be very time consuming. To avoid using manual annotation, we resort to using anchor text in Wikipedia as training data for key phrase extraction. Anchor texts in Wikipedia are manually labeled by crowds of contributors, thus are meaningful and reliable. Figure 1 is an excerpt of the Wikipedia personal name article for the American president *Abraham Lincoln*:

Abraham Lincoln  (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. He successfully led his country through its greatest constitutional, military and moral crisis – the American Civil War – preserving the Union while ending slavery, and promoting economic and financial modernization. Reared in a poor family on the western frontier, Lincoln was mostly self-educated. He became a country lawyer, a Whig Party, Illinois state legislator in the 1830s, a one-term member of the United States House of Representatives in the 1840s, but he failed in two attempts to be elected to the United States Senate in the 1850s. After opposing the expansion of slavery in the United States in his campaign debates and speeches,^[1] Lincoln secured the Republican Party nomination and was elected president in 1860.

Figure 1: Excerpt of a Wikipedia article

In this excerpt, *American Civil War*, *Whig Party*, *United States Senate*, *Illinois state legislator*, *Republican Party* and other anchor texts can be used as key phrases. Using the Wikipedias personal names articles, key phrase extraction algorithm can then be employed to train the prediction model. In

this work, the extraction algorithm uses the Naive Bayes (NB) learning strategy for training through the use of anchor texts in Wikipedia personal names articles to extract key phrases. The list of personal names in Wikipedia is first obtained from DBpedia¹ which are used to obtain the relevant Wikipedia personal articles. The NB algorithm creates the key phrase prediction model using the extracted key phrases during the training process. Similar to the supervised key phrase extraction approaches (Witten et al., 2000; Xu et al., 2012), our key phrase extraction is summarized as follows.

- Preprocessing: Clean the Wikipedia articles including html tags removal, text tokenization, lemmatization and case-folding;
- Anchor text extraction: Extract the anchor texts based on the embedded hyperlinks;
- Candidate phrase generation: Use ngram-based method to generate candidate phrases which can contain up to 3 words as a phrase. They cannot start and end with stop words;
- Annotation: Label the candidate phrases with anchor text as positive instances and others as negative instances;
- Feature value generation and discretization: Compute (1) candidate phrases TF*IDF values, and (2) the distance values by the number of words preceding the candidate phrases divided by the document length in words. If there are multiple candidate phrases in the same document, the value of its first appearance will be used;
- Classification: Use the Naive Bayes learning algorithm to produce the key phrase prediction model.

The NB classification for positive prediction is formally defined as:

$$P(yes|k) = \frac{Y}{Y+N} \times P_{tf*idf}(t|yes) \times P_{dist}(d|yes)$$

where k is a phrase, Y and N denote positive and negative instances. Positive instances are

¹<http://wiki.dbpedia.org>

those candidate phrases that are anchors in Wikipedia and negative ones are those candidate phrases which are not anchors. t is the discretized TF*IDF value and d refers to the discretized distance value.

4 Key Phrase-based Skip Bigrams

Skip-bigrams are pairs of key phrases in a sentence order with arbitrary gaps. They contain the sequential and order-sensitive information between two key phrases. Xu et al. (2012) extracted skip bigrams based on the words to measure sentence similarities. In this paper, we used the sequences of key phrases surrounding a name mention. To use the skip bigrams, the key phrase sequences are first extracted within a context window of the name mentions. Figure 3 shows the key phrases surrounding the mention of *Amanda Lentz*.

Tumbling World Cup, 28 Amanda Lentz won the 5th Tumbling World Cup Final, defeating current World Champion Elena Blouina from Russia by one tenth of a point. Amanda, who is reigning U.S. Tumbling Champion, talks with USA Gymnastics and shares here secrets to success.

Figure 2: Key Phrases for person *Amanda Lentz*

In this short text, the key phrases in the red circles (their extraction will be described in Section 3). To find the skip bigrams, we first pinpoint the person name *Amanda Lentz*, find the key phrases surrounding this name mention by specifying the window size, and then create a key phrase sequence as follows:

tumbling world_cup amanda_lentz tumbling world_cup world_champion russia tumbling champion usa_gymnastics

From the above key phrase sequences, the skip bigrams are extracted. Without loss of generality, let us consider the following examples of key phrase sequences S_1 and S_2 around a name mention:

$$S_1 = k_1 k_2 k_1 k_3 k_4 \text{ and } S_2 = k_2 k_1 k_4 k_5 k_4$$

where k_i denotes a key phrase. It can be used more than once in a key phrase sequence. Hence, S_1 has the following skip bigrams:

$$(k_1 k_2, k_1 k_1, k_1 k_3, k_1 k_4, k_2 k_1, k_2 k_3, k_2 k_4, k_1 k_3, k_1 k_4, k_3 k_4)$$

S_2 has the following skip bigrams:

$$(k_2 k_1, k_2 k_4, k_2 k_5, k_2 k_4, k_1 k_4, k_1 k_5, k_1 k_4, k_4 k_5, k_4 k_4, k_5 k_4)$$

In the key phrase sequence S_1 , we have two repeated skip bigrams $k_1 k_4$ and $k_1 k_3$. In the sequence S_2 , we have $k_2 k_4$ and $k_1 k_4$ repeated twice. In this case, the weight of the recurring skip bigrams will be increased. Now, the question remains of how to weigh the skip bigrams.

Given Ω as a finite key phrase set, let $S = k_1 k_2 \cdots k_{|S|}$ be a sequence of key phrases for a name mention, $k_i \in \Omega$ and $1 \leq i \leq |S|$. A skip bigram of S , denoted by u , is defined by an index set $I = (i_1, i_2)$ of S such that $1 \leq i_1 < i_2 \leq |S|$ and $u = S[I]$. The skip distance of $S[I]$, denoted by $l_u(I)$, is the skip distance of the first key phrase and the second key phrase of u in S , calculated by $i_2 - i_1 + 1$. For example, if S is the key phrase sequence of $k_1 k_2 k_1 k_3 k_4$ and $u = k_1 k_4$, then there are two index sets, $I_1 = [3, 5]$ and $I_2 = [1, 5]$ such that $u = S[3, 5]$ and $u = S[1, 5]$, and the skip distances of $S[3, 5]$ and $S[1, 5]$ are 3 and 5, respectively. In case a name mention occurs multiple times in a document, the key phrase sequences for the name mentions are concatenated in their occurrence order to form one compound sequence. In the following discussions, S refers to the compound key phrase sequence if there are multiple name mentions.

The weight of a skip bigram u for a given S with all its possible occurrences, denoted by $\phi_u(S)$, is defined as:

$$\phi_u(S) = \sum_{I: u=S[I]} \lambda^{l_u(I)}$$

where λ is the decay factor, in the range of $[0, 1]$, that penalizes the longer skip distance $l_u(I)$ of skip bigrams. That is to say, the longer the skip distance is, more discount will be given to the skip bigrams.

By doing so, for the key phrase sequence S_1 , the complete key phrase set is $\Omega = \{k_1, k_2, k_3, k_4\}$. The weights for the skip bigrams are listed in Table 1:

These extracted skip bigrams with their corresponding weights will be concatenated into the key phrase-based vector space model. Suppose two documents are represented by the

u	$\phi_u(S_1)$	u	$\phi_u(S_1)$
k_1k_2	λ^2	k_2k_1	λ^2
k_1k_1	λ^3	k_2k_3	λ^3
k_1k_3	$\lambda^4 + \lambda^2$	k_2k_4	λ^4
k_1k_4	$\lambda^5 + \lambda^3$	k_3k_4	λ^2

Table 1: Skip Bigrams and their Weights in S_1

key phrase vectors V_{S_1} and V_{S_2} ,

$$V_{S_1} = (k_1, k_2, k_3, k_4)'$$

$$V_{S_2} = (k_1, k_2, k_4, k_5)'$$

The symbol prime denotes the transpose of the row vectors. Once the skip bigrams are extracted, they are concatenated into their vector spaces and thus the V_{S_1} and V_{S_2} are expanded into

$$V_{S_1} = (k_1, k_2, k_3, k_4, k_1k_2, k_1k_1, \\ k_1k_3, k_1k_4, k_2k_1, k_2k_3, k_2k_4, k_3k_4)'$$

$$V_{S_2} = (k_1, k_2, k_4, k_5, k_2k_1, k_2k_4, \\ k_2k_5, k_1k_4, k_1k_5, k_4k_5, k_4k_4, k_5k_4)'$$

The V_{S_1} and V_{S_2} vectors are enriched after concatenation and if they share more overlapping skip bigrams with similar skip distances, the similarity between V_{S_1} and V_{S_2} will be increased.

5 Experiments

The evaluation of the algorithm is conducted using the test data of *WePS2* workshop 2009² which has 30 ambiguous names. Each ambiguous name has 150 search results from the various domains including US census, Programme Committee members for the annual meeting of ACL, and so on (Artiles et al., 2009).

Because the number of clusters is not known beforehand, the parameter configuration for clustering is of great importance for clustering web persons. In this paper, the *WePS1* development data³ is used to select the optimal threshold. This development data contains 47 ambiguous names. The number of clusters per name has a large variability from 1 to 91 different people sharing the name (Artiles et al., 2009). In the preprocessing step, the software Beautiful Soup⁴ is used to clean the html texts and the OpenNLP tool⁵ to tokenize cleaned texts.

²<http://nlp.uned.es/weps/weps-2>

³<http://nlp.uned.es/weps/weps-1>

⁴<http://www.crummy.com/software/BeautifulSoup/>

⁵<http://incubator.apache.org/opennlp/>

5.1 Key Phrase Extraction

For key phrase extraction, no training data from WePS is used. Instead, the training data are from the Wikipedia personal names articles. The personal names in Wikipedia are available from the DBpedia. 245,638 personal names are used in this paper with their corresponding Wikipedia articles. These persons come from different walks of life, thus providing a wide coverage of terms across different domains. Through the Wikipedia personal name article titles, the Wikipedia Miner tool⁶ is used to obtain the anchor text within the article page. With the article pages as documents and the related key phrases (anchor texts), the key phrase prediction model is trained first. Then the key phrases in the WePS testing data are extracted. In case of overlapping key phrases, longer key phrases will be used. For example, *president of united states* and *united states* are both key phrases. But, when *president of united states* appears in the context, it will be used even though both *present of united states* and *united states* are extracted simultaneously.

The key phrases extracted for the persons *AMANDA_LENTZ* and *BENJAMIN_SNYDER* are listed here as an example:

AMANDA_LENTZ: *IMDb, North Carolina, Literary Agents, published writers, High School, Family History, CCT Faculty, Campus Calendar, Women Soccer, World Cup, Trampoline, ...*

BENJAMIN_SNYDER: *Biography Summary, Artist, National Gallery of Canada, Fine Arts Museum, history of paintings, modern art work, University of Manitoba, Special Collections, portfolio gallery, ...*

It is quite obvious that above extracted key phrases are informative and useful for the WPD task. Compared to using topic words, the use of key phrases reduces the document dimension significantly, thus reducing runtime cost. When dealing with internet documents which can be in very large quantity, reduction of runtime cost can make the algorithms more practical.

⁶<http://wikipedia-miner.cms.waikato.ac.nz/>

5.2 Evaluation Metrics for WPD

The algorithm is evaluated by the purity, inverse purity scores, and B-Cubed precision and recall (Artiles et al., 2007, 2009). The purity measure is defined as

$$Purity = \sum_i \frac{C_i}{n} maxPre(C_i, L_j)$$

$$Pre(C_i, L_j) = \frac{C_i \cap L_j}{C_i}$$

where C_i denotes the i^{th} cluster produced by the system, L_j denotes the j^{th} manually annotated category and n the number of clustered documents. $Pre(C_i, L_j)$ refers to precision of a C_i for the category L_j . Inverse purity focuses on the cluster with the maximum recall for each category, defined by,

$$Inv.Purity = \sum_i \frac{L_i}{n} maxPre(L_i, C_j)$$

To take into consideration of both precision and recall in evaluating clustering performance, the harmonic mean of both purity and inverse purity is defined as follows:

$$F = \frac{1}{\alpha \frac{1}{Purity} + (1 - \alpha) \frac{1}{Inv.Purity}}$$

where $\alpha = \{0.2, 0.5\}$ used in the WePS workshops (Artiles et al., 2009, 2010). If smaller α gives more importance to inverse purity, indicating a higher weight to recall. In the case of $\alpha = 0.5$, equal weighting is given to precision and recall.

B-Cubed metrics calculate the precision and recall related to each item in the clustering result. The B-Cubed precision (BEP) of one item represents the amount of items in the same cluster that belong to its category, whereas the B-Cubed recall (BER) represents how many items from its category belong to its cluster. They are,

$$BEP = Avg_e[Avg_{e' | C(e) \cap C(e') \neq 0}[Mult.Pre(e, e')]]$$

$$BER = Avg_e[Avg_{e' | L(e) \cap L(e') \neq 0}[Mult.Recall(e, e')]]$$

e and e' are two documents, $C(e)$ and $L(e)$ denote the clusters and categories related to e . The multiplicity precision $Mult.Pre(e, e')$ is 1 when e and e' in the same cluster share the same category. Therefore, the B-Cubed precision of one item is its averaged multiplicity precision with the other items in the same categories. The multiplicity recall $Mult.Recall(e, e')$ is 1 when

e and e' in the same category share the same cluster. Similarly, the harmonic mean of B-Cubed precision and recall is defined by,

$$F = \frac{1}{\alpha \frac{1}{BEP} + (1 - \alpha) \frac{1}{BER}} \quad \alpha = \{0.2, 0.5\}$$

5.3 Document Clustering for WPD

The clustering algorithm used in this work is the hierarchical agglomerative clustering algorithm in single linkage (Manning et al., 2008). Documents are represented by key phrase vectors and their similarities are computed using the cosine metric. The weight for a key phrase is calculated with the consideration of both TF and ITF as well as the link probability as defined before (similar to that used in (Xu et al., 2012)).

$$W_k = \log(TF(k) + 1) * (\log IDF(k) + Pr_{link}(k))$$

where $TF(k)$ denotes the term frequency of k , $IDF(k)$ is the inverse document frequency of k and $Pr_{link}(k)$ is the link probability of k . $Pr_{link}(k)$ is defined as $Pr_{link}(k) = \frac{C_{link}(k)}{C_{occur}(k)}$. $C_{link}(k)$ is the number of hyperlinks anchored to k in Wikipedia, and $C_{occur}(k)$ is the number of occurrences of k in the Wikipedia articles. That means some extracted key phrases appear in the Wikipedia articles, but are not linked to, thus their importance decreases.

As the number of clusters cannot be predetermined, we use the *WePS1* development data to select optimal parameters which give the best B-Cubed and purity F-measures. The parameter configurations are listed in Table 2.

SD	DF(λ)	WS	CP
3	0.5	20	0.182

Table 2: Parameter Configurations

SD denotes the skip distance which is used to specify how many gaps can be allowed in a skip bigram; DF refers to the decay factor λ which is used to penalize the non-continuous skip bigrams. WS is the window size to specify the maximum number of key phrases surround a name mention, and CP denotes the cut-off point for the number of clusters in the hierarchical dendrogram.

In the following experiments, $APKPB$ refers to the clustering algorithm purely using key phrases ($PKPB$ stands for pure key phrase

based approach) and A_{SKIP} denote the clustering algorithm using skip bigrams. The aggregated algorithm is denoted by A_{AKPC} . Table 3 and Table 4 show the comparison of A_{AKPC} the algorithm with the top-3 systems in *WePS* 2009 in terms of purity measure and B-Cubed measure, respectively.

Runs	F-measures			
	$\alpha=0.5$	$\alpha=0.2$	Pur.	Inv_Pur.
T1: PolyUHK	0.88	0.87	0.91	0.86
T2: UVA_1	0.87	0.87	0.89	0.87
T3: ITC_UT_1	0.87	0.83	0.95	0.81
A_{AKPC}	0.88	0.87	0.86	0.87

Table 3: Performance Comparison of A_{AKPC} using Purity scores

Runs	F-measures			
	$\alpha=0.5$	$\alpha=0.2$	BEP	BER
T1:PolyUHK	0.82	0.80	0.87	0.79
T2:UVA_1	0.81	0.80	0.85	0.80
T3:ITC_UT_1	0.81	0.76	0.93	0.73
A_{AKPC}	0.82	0.80	0.85	0.80

Table 4: Performance Comparison of A_{AKPC} using B-Cubed scores

Table 3 and Table 4 show that in comparison to the top 1 system, the proposed A_{AKPC} has the same performance in terms of F-measure for both purity score and B-cubed score. In terms of B-Cubed recall, A_{AKPC} achieves the highest score, implying that the number of categories has been well guaranteed by our clustering solutions. Admittedly, our system loses 2 percent in terms of B-Cubed precision. However, when comparing to the features used in the top 3 systems, the top 1 system by *PolyUHK* (Chen et al., 2009) incorporates tokens, title tokens, n-gram and snippet features into its system using VSM. The *PolyUHK* system has to tune the unigram and bigram weights through the Goodgle 1T corpus which is external to the *WePS* training data. The second best *UVA_1* system (Balog et al., 2009) employs all tokens of in the training document only documents, and the third best *ITC_UT_1* system (Ikeda et al., 2009) uses named entities, compound nouns and URL links features. The A_{AKPC} algorithm in this paper simply uses key phrase and limited amount

of skip bigrams around the name mentions. The Key phrase extraction algorithm are trained by Wikipedia article. Even though this takes additional computation power, it can be done once only. In the testing phase, extraction of key phrases is much faster than the other systems and the dimension of the key phrases in the VSM is also much smaller than the other systems.

To measure the effectiveness of the two sub-algorithms A_{PKPB} and A_{SKIP} , performance of the two algorithms are also evaluated separately as independent clustering algorithms shown in Table 5 and Table 6 for B-cubed measures and purity measures, respectively. Note that when evaluating the two algorithms, the cut-off points need to be readjusted from the *WEPS1* development data. The cut-off point for A_{PKPB} remains unchanged as 0.182 and the A_{SKIP} cut-off point is set to 0.055.

From Table 5 and Table 6, it can be seen that for both A_{PKPB} and A_{SKIP} , if used separately, do not perform as well as A_{AKPC} . However, A_{PKPB} has a better performance than A_{SKIP} when used alone. This implies that key phrases, as a single feature in clustering algorithm, are better features than using skip bigrams of key phrases surrounding the mentions. This is easy to understand as the context windows for the name mentions used in skip bigram model do not have as large a coverage of key phrases as that in the whole documents. However, A_{SKIP} gives the second best performance in B-Cubed precision and purity. This is why the overall B-Cubed precision and purity are improved after its aggregation.

In terms of purity in Table 5, A_{PKPB} has the same performance as the top 1 and top 2 systems in term F0.2 and 1 percent better when compared to the top 2 system in terms of inverse purity. Our system, however, loses 1 percent in F0.5 score and 4 percent in purity score when compared to the top 1 system and A_{SKIP} achieves a second best purity score among the top 3 systems.

It is most important to point out that the A_{PKPB} algorithm in Table 6, however simple, has a competitive performance in comparison to the top 1 system. A_{PKPB} has the best results in terms of F0.2 for B-cubed score, implying that

Runs	F-measures			
	$\alpha=0.5$	$\alpha=0.2$	Pur.	Inv_Pur.
T1: PolyUHK	0.88	0.87	0.91	0.86
T2: UVA_1	0.87	0.87	0.89	0.87
T3: ITC_UT_1	0.87	0.83	0.95	0.81
A_{PKPB}	0.87	0.87	0.87	0.88
A_{SKIP}	0.79	0.74	0.93	0.71

Table 5: Performance Comparison of A_{PKPB} and A_{SKIP} using Purity scores

two documents in the same manually annotated categories share the same cluster produced by our system. In terms of B-Cubed score, even though A_{PKPB} loses one percent in F0.5, the performance gain is three percent in B-Cubed recall when compared to the *PolyUHK* system. In terms of B-Cubed precision, our system is not as good as the top three systems. However, our system strikes a better balance between B-Cubed precision and purity score, which means that our system's clustering solutions are consistent with manually annotated categories.

Runs	F-measures			
	$\alpha=0.5$	$\alpha=0.2$	BEP	BER
T1:PolyUHK	0.82	0.80	0.87	0.79
T2:UVA_1	0.81	0.80	0.85	0.80
T3:ITC_UT_1	0.81	0.76	0.93	0.73
A_{PKPB}	0.81	0.81	0.82	0.82
A_{SKIP}	0.69	0.63	0.91	0.60

Table 6: Performance Comparison of A_{PKPB} and A_{SKIP} using B-Cubed scores

In order to demonstrate the performance improvement by aggregating the skip bigrams into the vector space model, we looked at our designs with and without aggregating skip bigrams. Table 7 shows the evaluation results.

Runs	F-measures		Purity		B-Cubed	
	B-Cubed	Purity	P	IP	BEP	BER
A_{PKPB}	0.81	0.87	0.87	0.88	0.82	0.82
A_{AKPC}	0.82	0.88	0.89	0.87	0.85	0.80

Table 7: Performance Comparison of A_{PKPB} and A_{AKPC} using both Purity and B-Cubed scores

In Table 7, both B-Cubed and purity F0.5 scores have been increased by 1 percent. The B-Cubed precision is improved by 3% and purity

is increased by 2%, which means that the A_{AKPC} gives a much more reliable clustering solution. It is common in most information retrieval cases that algorithms with high precision will have a compromise on their recall performance. In this paper, we have gained 3% and 2% improvement in B-Cubed precision and Purity, but lost 2% and 1% in B-Cubed recall and inverse purity, respectively.

6 Conclusions and Future Work

This paper proposed the *AKPC* algorithm to use key phrases as document representations and skip-bigram of key phrases as contextual information in Web person disambiguation. Results show that the proposed *AKPC* algorithm gives a competitive performance when compared to the top three systems in *WePS* 2009.

Further investigation also shows that clustering based on key phrases as single features is very effective. It employs a supervised approach to extract meaningful key phrases for person names. The extraction of key phrases in the training phase is fully automatic and no manual annotation is needed as the training data is from Wikipedias anchor text. The weighting scheme takes into consideration of both the traditional TF*IDF and the Wikipedia link probability. Experiments show that the proposed key phrase based clustering algorithm using VSM is both effective and efficient. Unlike the tokens used by most of previous researches, key phrases are more meaningful and are more capable of separating people of the same namesake.

Further extension of this work includes aggregating order-sensitive skip bigrams into key phrase-based vector space model to enrich context information in the inclusion of web persons disambiguation. Experiments show that the precision of clustering solutions is increased. We combined the decay factor with the skip distance to assign a reasonable weight for skip bigrams and studied the effectiveness of varying skip distance and decaying factor. In future work, we will explore skip ngrams for a larger n. Moreover, we will explore the use of efficient combination of key phrases with skip ngrams.

References

- A. Hulth. 2003. Improved Automatic Keyword Extraction Given more Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216-223.
- A. Hulth and B. Megyesi. 2006. A Study on Automatically Extracted Keywords in Text Categorization. In *CoLing/ACL 2006, Sydney*.
- Y. Chen, Sophia Yat Mei Lee and Chu-Ren Huang. 2009. PolyUHK: A Robust Information Extraction System for Web Personal Names. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.
- Chong Long and Lei Shi. 2010. Web Person Name Disambiguation by Relevance Weighting of Extended Feature Sets. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.
- Decong Li, Sujian Li, Wenjie Li, Wei Wang, and Weiguang Qu. 2010. A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 296-300.
- D.C. Manning, P. Raghavan, and H. Schutze. 2008. *Hierarchical Clustering. Introduction to Information Retrieval*. Cambridge University Press, New York, 2008, 377-401.
- E. Frank, G. W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific Keyphrase Extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99)*, pages 668-673.
- G. Salton and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- Hongyuan Zha. 2002. Generic Summarization and Key Phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Tampere*, pages 113-120.
- I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin and C.G. Nevill-Manning. 2000. KEA: Practical Automatic Keyphrase Extraction. In *Working Paper 00/5, Department of Computer Science, The University of Waikato*.
- J. Artiles, J. Gonzalo and S. Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64-69.
- J. Artiles, J. Gonzalo and S. Sekine. 2009. WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.
- J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine and E. Amigo. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.
- Jian Xu, Qin Lu, and Zhengzhong Liu. 2012. PolyUCOMP: Combining Semantic Vectors with Skip bigrams for Semantic Textual Similarity. *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*.
- Jian Xu, Qin Lu and Zhengzhong Liu. 2012. Combining Classification with Clustering for Web Person Disambiguation. *WWW 2012 Companion, April 16-20, 2012, Lyon, France*.
- K. M. Hammouda, D.N. Matute and M.S. Kamel. 2005. CorePhrase: Keyphrase Extraction for Document Clustering. In *Proceedings of MLDM. 2005*.
- K. Balog, J. He, K. Hofmann, et al. 2009. The University of Amsterdam at WePS2. *2nd Web People Search Evaluation Workshop. 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.
- Lin Chin-Yew and Franz Josef Och. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.
- M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. 2009. Person Name Disambiguation on the Web by Two-Stage Clustering. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.
- Marina Litvak and Mark Last. 2008. Graph-based Keyword Extraction for Single-document Summarization. In *Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, pages 17-24.
- M. Steyvers and T. Griffiths. 2007. Probabilistic Topic Models. *T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum*.
- O. Popescu and B. Magnini. 2007. IRST-BP: Web People Search using Name Entities. *Proceedings of the Fourth International Workshop on Semantic*

- Evaluations (SemEval-2007), June (2007)*, pages 195-198.
- Xiaojun Wan and Jianguo Xiao. 2008. CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969-976.
- Xianpei Han and Jun Zhao. 2009. ASIANED: Web Personal Name Disambiguation Based on Professional Categorization. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, pages 2-5.
- Zhiyuan Liu, Wenyi Huang, Yabin Zheng and Maosong Sun. 2010. Automatic Keyphrase Extraction via Topic Decomposition. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 366-376.

Named Entity Recognition: Exploring Features

Maksim Tkachenko

St Petersburg State University
St Petersburg, Russia

maksim.tkachenko@math.spbu.ru andrey.simanovsky@hp.com

Andrey Simanovsky

HP Labs Russia
St Petersburg

Abstract

We study a comprehensive set of features used in supervised named entity recognition. We explore various combinations of features and compare their impact on recognition performance. We build a conditional random field based system that achieves 91.02% F_1 -measure on the CoNLL 2003 (Sang and Meulder, 2003) dataset and 81.4% F_1 -measure on the OntoNotes version 4 (Hovy et al., 2006) CNN dataset, which, to our knowledge, displays the best results in the state of the art for those benchmarks respectively. We demonstrate statistical significance of the boost of performance over the previous top performing system. We also obtained 74.27% F_1 -measure on NLPBA 2004 (Kim et al., 2004) dataset.

1 Introduction

Recognition of named entities (e.g. people, organizations, locations, etc.) is an essential task in many natural language processing applications nowadays. Named entity recognition (NER) is given much attention in the research community and considerable progress has been achieved in many domains, such as newswire (Ratinov and Roth, 2009) or biomedical (Kim et al., 2004) NER. Supervised NER that uses machine learning algorithms such as conditional random fields (CRF) (McCallum and Li, 2003) is especially effective in terms of quality of recognition.

Supervised NER is extremely sensitive to selection of an appropriate feature set. While many features were proposed for use in supervised NER

systems (Krishnan and Manning, 2006; Finkel and Manning, 2009; Lin and Wu, 2009; Turian et al., 2010), only limited studies of the impact of those features and their combinations on the effectiveness of NER were performed. In this paper we provide such a study.

Our contributions are the following:

- analysis of the impact of various features taken from a comprehensive set on the effectiveness of a supervised NER system;
- construction of a CRF-based supervised NER system that achieves 91.02% F_1 -measure on the CoNLL 2003 (Sang and Meulder, 2003) dataset and 81.4% F_1 -measure on the OntoNotes version 4 (Hovy et al., 2006) CNN dataset;
- demonstration of statistical significance of the obtained boost in NER performance on the benchmarks;
- application to NER of a DBpedia (Mendes et al., 2011) markup feature and a phrasal clustering (Lin et al., 2010) feature which have not been considered for NER in previous works.

The remainder of the paper is structured in the following way. In Section 2 we describe related work on feature analysis. In Section 3 we give a brief introduction to the benchmarks that we use. In Section 4 we discuss various features and their impact. Section 5 describes the final proposed system. Section 6 contains a summary of the performed work and future plans.

2 Related Work

The majority of papers on NER describe a particular method or feature evaluation and do not make a systematic comparison of combinations of features. In this paper those works are mentioned later when we discuss a particular feature or a group of features. In this section we present several works that deal with multiple features and thus are close to our study.

Design questions of NER systems were considered by (Ratinov and Roth, 2009). They used a perceptron-based recognizer with greedy inference and evaluated two groups of features: non-local dependencies (e.g. context aggregation) and external information (e.g. gazetteers mined from Wikipedia). Their recognizer was tested on the CoNLL 2003 dataset, a newswire dataset ($F_1 = 90.80\%$), the MUC7 dataset, and their own web pages dataset.

The authors of (Turian et al., 2010) systematically compared word representations in NER (Brown clusters, Collobert and Weston embeddings, HLBL embeddings). They ignored other types of features.

(Saha et al., 2009) presented a comparative study of different features in biomedical NER. They used a dimensionality reduction approach to select the most informative words and suffixes and they used clustering to compensate for the lost information. The MaxEnt tagger developed by them obtained $F_1 = 67.4\%$ on NLPBA 2004 data.

3 Benchmarks

In this paper we present the results obtained on three benchmarks: CoNLL 2003, OntoNotes version 4, and NLPBA 2004 dataset.

CoNLL 2003 is an English language dataset for NER. The data comprises Reuters newswire articles annotated with four entity types: person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). The data is split into a training set, a development set (testa), and a test set (testb). Performance on this task is evaluated by measuring precision and recall of annotated entities combined into F_1 -measure. We used BILOU (begin, inside, last, outside, unit) annotation scheme to encode named entities. Previ-

ous top performing systems also followed that scheme. We study feature behavior on this benchmark; our system is tuned on the test and development sets of it.

OntoNotes version 4 is an English language dataset designed for various natural language processing tasks including NER. The dataset consists of several sub-datasets taken from different sources including Wall Street Journal, CNN news, machine-translated Chinese magazines, Web blogs, etc. We provide the results obtained by our final system on OntoNotes subsets in order to compare them with earlier works. It has its own set of named entity classes but it has a mapping of those to CoNLL classes. We use the latter for systems comparison. We used the same test/training split as in (Finkel and Manning, 2009).

NLPBA 2004 dataset (Kim et al., 2004) is an English language dataset for bio-medical NER. It consists of a set of PubMed abstracts nad has a correspoding set of named entites (protein, DNA, RNA, cell line, cell type).

4 Feature Set

We performed feature comparison using our system which is a CRF with Viterbi inference. We have also tested greedy inference and have found out that the system performs worse and its results are lower than those of a perceptron with greedy inference that we modeled after (Ratinov and Roth, 2009).

In each of the following subsections we consider a particular type of features. In Subsection 4.1 we deal with local knowledge features which can be extracted from a token (word) being labeled and its surrounding context. Subsection 4.2 describes evaluation of external knowledge features (part-of-speech tags, gazetteers, etc.). Discussion of non-local dependencies of named entities is included in Subsection 4.3. Sub-section 4.4 contains further improvements of performance and specific features that do not fall into previous categories; they help to overcome common errors on the CoNLL 2003 dataset.

4.1 Local Knowledge

Our baseline recognizer uses only local information about a current token. It is not surprising that a token-based CRF performs poorly, espe-

Features	Dev	Test
CoNLL-2003		
w_0	25.24%	22.04%
w_{-1}, w_0, w_1	83.41%	74.82%
$w_{-1}, w_0, w_1,$ $w_{-1} \& w_0, w_0 \& w_1$	81.20%	72.26%
$w_{-2}, w_{-1}, w_0, w_1, w_2$	82.31%	73.73%
NLPBA 2004		
w_0	-	61.67%
w_{-1}, w_0, w_1	-	65.51%
$w_{-1}, w_0, w_1,$ $w_{-1} \& w_0, w_0 \& w_1$	-	66.01%
$w_{-2}, w_{-1}, w_0, w_1, w_2$	-	65.45%

Table 1: Evaluation of context in NER; w — token, $a \& b$ — conjunction of features a and b .

cially when we try to model non-unit named-entity chunks¹. A system which only selects complete unambiguous named entities that appear in training data works better (Tjong Kim Sang and De Meulder, 2003). Table 1 contains performance results of context features. w_0 is a current token, w_1 is a following token and w_{-1} is a preceding one. The larger context we consider the worse F_1 -measure we get. Such behavior indicates that token/class dependency statistics in the training corpus is not enough to deduce which context is important. The quality maximum is observed when we use a sliding window of three tokens. The context can be smoothed by semi-supervised learning algorithms (Ando and Zhang, 2005) in order to compensate for the lack of statistics.

Better results are obtained if we ignore surrounding tokens but use more features based only on the current token. We used suffixes, prefixes² and orthographic features (shape³) of the current token (see Table 2). Different word-based features give us better evidence of a particular word being a part of a named entity (the gain

¹BILOU scheme is not appropriate for one-token features; an adequate result should be around $F_1 = 52\%$ as in (Klein et al., 2003).

²We used prefixes and suffixes of the length up to 6 to reduce the number of features. For example, suffixes like *-burg*, *-land* are highly correlated with location entities (Maastricht, Nederland)

³Informally, the shape feature is a result of mappings like Bill → Xxx, Moscow-based → Xxx-xx, etc.

Features	Dev	Test
CoNLL-2003		
w_0	25.24%	22.04%
$w_0 + \text{suffixes and prefixes}$	87.41%	78.59%
$w_0 + s_0$	86.70%	79.16%
$w_0 + s_{-1}, s_0, s_1,$ $s_{-1} \& s_0, s_0 \& s_1, s_{-1} \& s_0 \& s_1$	87.67%	81.37%
All Local Features	88.91%	82.89%
NLPBA 2004		
w_0	-	61.67%
$w_0 + \text{suffixes and prefixes}$	-	66.22%
$w_0 + s_0$	-	62.01%
$w_0 + s_{-1}, s_0, s_1,$ $s_{-1} \& s_0, s_0 \& s_1, s_{-1} \& s_0 \& s_1$	-	65.85%
All Local Features	-	66.83%

Table 2: Evaluation of local features in NER; w — token, s — shape, $a \& b$ — conjunction of features a and b .

in F_1 is about 4%) than the context does. It is also useful to extend the shape feature onto surrounding words. The token-based features do not outperform the context features in the biomedical domain but still provide useful information. Biomedical entities are different from newswire entities in terms of shape features; for instance, lower-cased entities (e.g. *persistently infected cells*) are common in the former domain. Domain-specific modifications are required for the shape function (e.g., the regular shapes of the proteins *CD4* and *CD28* are not the same).

4.2 External Knowledge

Most NER systems use additional features like part-of-speech (POS) tags, shallow parsing, gazetteers, etc. Such kind of information requires external knowledge: unlabeled texts, trained taggers, etc. We consider POS tags (Section 4.2.1), words clustering (Section 4.2.2), phrasal clustering (Section 4.2.3), and encyclopedic knowledge (Section 4.2.4). F_1 measures obtained in the experiments covered in this section are shown in Tables 3 and 4; the discussion follows below.

4.2.1 Part-of-Speech Tagging

POS tags are widely used in NER but recently proposed systems omit this information (Ratinov

and Roth, 2009; Lin and Wu, 2009). POS tagging is itself a challenge and this preprocessing task can take a lot of time. We find that the impact of these features depends on a POS tagger. We replace the original POS tag annotation with the annotation produced by OpenNLP tagger⁴, however, even high-quality POS tags lead to a decrease of F_1 -measure.

4.2.2 Words Clustering

The authors of (Ando and Zhang, 2005; Suzuki and Isozaki, 2008; Turian et al., 2010) showed that utilization of unlabeled data can improve the quality of NER. We divide the recognizers that use unlabeled text into two groups. The first group consists of semi-supervised systems which directly use labeled and unlabeled data in their training process (Ando and Zhang, 2005; Suzuki and Isozaki, 2008). The second group includes systems that use features derived from unlabeled data (Ratinov and Roth, 2009; Lin and Wu, 2009). (Turian et al., 2010) have shown that recognizers of the first group tend to perform better presumably because they have task-specific information during the training process. However, a simpler way to improve NER quality is to include word representations as features into learning algorithms.

Brown clusters were prepared by the authors of (Turian et al., 2010) by clustering the RCV1 corpus which is a superset of the CoNLL 2003 dataset⁵. Clark clusters were induced by us with the original Clark’s code⁶ on the same RCV1 corpus but without preprocessing step used in (Turian et al., 2010). Brown clusters were successfully applied in NER (Miller et al., 2004; Ratinov and Roth, 2009). We consider Clark’s algorithm since it shows competitive results in unsupervised NLP (Christodoulopoulos et al., 2010; Spitskovsky et al., 2011) and it is also successfully used in NER (Finkel and Manning, 2009). A combination of different word representations (Turian et al., 2010) gives better results. We also applied latent Dirichlet allocation (LDA) to create probabilistic word clustering in the same way as

Features	Dev	Test
CoNLL-2003		
p_0	45.63%	43.98%
$w_0 + p_0$	83.07%	73.42%
b_0	80.98%	75.51%
$w_0 + b_0$	89.35%	82.17%
c_0	67.47%	64.06%
$w_0 + c_0$	86.47%	79.29%
l_0	45.20%	44.24%
$w_0 + l_0$	82.28%	72.63%
g_0	79.90%	76.72%
$w_0 + g_0$	88.36%	81.98%
$b_0 + c_0 + l_0$	86.40%	80.76%
$b_0 + c_0 + l_0 + g_0 + p_0$	89.26%	84.66%
$w_0 + b_0 + c_0 + l_0 + g_0 + p_0$	90.87%	87.00%
NLPBA 2004		
p_0	-	18.29%
$w_0 + p_0$	-	62.81%
b_0	-	30.65%
$w_0 + b_0$	-	63.70%
c_0	-	15.53%
$w_0 + c_0$	-	63.41%
l_0	-	12.81%
$w_0 + l_0$	-	63.42%
g_0	-	43.30%
$w_0 + g_0$	-	63.41%
$b_0 + c_0 + l_0$	-	40.65%
$b_0 + c_0 + l_0 + g_0 + p_0$	-	58.47%
$w_0 + b_0 + c_0 + l_0 + g_0 + p_0$	-	63.52%

Table 3: Evaluation of phrasal and word clusterings in NER; w — token, p — POS tag, c — Clark clusters, b — Brown clusters, l — LDA clusters, g — phrasal clusters. Subscript index stands for the token which clustering label is used. -1 stands for the previous token; $+1$ stands for the next token; 0 stands for the current token.

⁴<http://incubator.apache.org/opennlp/>

⁵The resource is available at <http://metaoptimize.com/projects/wordreprs/>

⁶The code is available at <http://www.cs.rhul.ac.uk/home/alexc/>

was done in (Chrupala, 2011) and used the most probable cluster label of a word as a feature.

Brown’s algorithm is a hierarchical clustering algorithm which clusters words that have a higher mutual information of bigrams (Brown et al., 1992). The output of the algorithm is a dendrogram. A path from the root of the dendrogram represents a word and can be encoded with a bit sequence. We have used prefixes of the length of 7, 11, 13 of such encodings as features (these numbers were selected on the CoNLL development set with a recognizer that used *token + Brown* feature), which gave us 10368 clusters.

Clark’s algorithm groups words that have similar context distribution and morphological clues starting with most frequent words (Clark, 2003). We induced 200 clusters according to (Finkel and Manning, 2009) and used them as features.

We used LDA with 100 clusters.

We experimented with a combination of the current token feature and its cluster representation as well as with the representation alone. One interesting observation is that the small space of features (without any tokens) gives results comparable to those of the system which uses tokens with context (on CoNLL 2003 dataset). This trend continues with phrasal clusters (see Table 3).

4.2.3 Phrasal Clustering

Word clustering can be extended onto phrases. Presumably, phrases are far less ambiguous than single words. That consideration was applied to NER by (Lin and Wu, 2009) who presented a scalable k-means clustering algorithm based on Map-Reduce. It is not possible to reproduce their result exactly because they employed private data. In our experiments we used a 1000 soft clusters derived from 10 million phrases from a large web n-gram corpus by a similar k-means algorithm (Lin et al., 2010). N-grams that have high entropy of context are considered phrases. The resource⁷ contains phrases and their cluster memberships (up to twenty clusters) along with the similarity to each cluster centroid. We omit similarity information and treat cluster id’s as features

⁷<http://webdocs.cs.ualberta.ca/~bergsma/PhrasalClusters/>

(with the corresponding prefixes of the BILOU-scheme) for each word in a phrase.

The combination of phrasal clusters and Brown clusters performs well and is only slightly worse than their combination with tokens. Thus, context-based clustering with enough statistical information is able to detect named entity mentions. The NLPBA 2004 dataset is from another domain and the above-mentioned effect is not fully preserved but the clustering still improves performance.

4.2.4 Encyclopedic Knowledge

A simple way to guess whether a particular phrase is a named entity or not is to look it up in a gazetteer. Look-up systems with large entity lists work pretty well if entities are not ambiguous. In that case the approach is competitive against machine learning algorithms (Nadeau, 2007). Gazetteer features are common in machine-learning approaches too and can improve performance of recognition systems (Nadeau and Sekine, 2007; Kazama and Torisawa, 2007). Nowadays there are a lot of web resources which are easily adaptable to NER such as Wikipedia⁸, DBpedia⁹, and YAGO¹⁰. We employed Wikipedia and DBpedia.

Wikipedia was successfully used for NER before in (Kazama and Torisawa, 2007; Joel Nothman, 2008; Ratinov and Roth, 2009). Wikipedia contains redirection pages which take the reader directly to a different page. They are often created for synonymous lexical representations of objects or they denote variations in spelling, e.g., the page entitled *International Business Machines* is a redirection page for *IBM*. Disambiguation pages is another kind of Wikipedia pages. Disambiguation pages contain links to other pages which titles are homonyms. For instance, the page *Apple (disambiguation)* contains links to *Apple Inc.* and *Apple (band)*.

We used the titles of Wikipedia pages with their redirects as elements of gazetteers. To get class labels for each Wikipedia page, we employed a classifier proposed by (Tkatchenko et

⁸<http://www.wikipedia.org/>

⁹<http://dbpedia.org/>

¹⁰<http://www.mpi-inf.mpg.de/yago-naga/yago/>

Features	Dev	Test
CoNLL-2003		
w_0 + Wikipedia gaz.	56.35%	53.98%
w_0 + Wikipedia gaz. + disambig.	84.73%	77.72%
w_0 + DBpedia gaz.	84.06%	75.40%
w_0 + DBpedia gaz. + disambig.	83.62%	75.14%
w_0 + Wikipedia & DBpedia gaz.	85.21%	78.16%
NLPBA 2004		
w_0 + DBpedia gaz.	-	61.66%

Table 4: Evaluation of encyclopedic resources in NER; w — token.

al., 2011). We combined the semi-automatically derived training set with manually annotated data which enabled us to improve classification results (in order to omit redundant and noisy features we selected classes that correspond to named entity classes in the dataset: PERSON, GPE, ORGANIZATION, FACILITY, GEOLOGICAL REGION, EVENT). Since this markup does not contain NLPBA classes we used Wikipedia only with CoNLL classes.

DBpedia is a structured knowledge base extracted from Wikipedia. The DBpedia ontology consists of 170 classes that form a subsumption hierarchy. The ontology contains about 1.83 million of classified entities (Bizer et al., 2009). We used this hierarchy to obtain high quality gazetteers.

We developed a simple disambiguation heuristic to utilize resources provided by Wikipedia and DBpedia. A disambiguation page defines a set of probable meanings of a particular term. If an ambiguous term co-occurs with an unambiguous term from the same meaning set, the former will be resolved to the meaning of the latter and labeled as an element of the corresponding gazetteer. The results of application of gazetteers is shown in Table 4. Because ambiguity resolving (or wikification) is out of the scope of this paper we do not use it in further experiments including the final system.

We also applied *additional gazetteers* taken from Illinois University parser (Ratinov and Roth,

2009).

4.3 Non-local Dependencies

The same tokens often have the same labels. However, sometimes they may have different labels, for example, *HP* and *HP LaserJet 4200* are not entities of the same type (it is likely that we annotate them as COMPANY and PRODUCT respectively). The latter case is often governed by non-local information. Techniques like two-stage prediction (Krishnan and Manning, 2006), skip-chain CRF (Sutton and McCallum, 2006), and a recognizer which uses Gibbs sampling and penalizes entities with different labels (Finkel et al., 2005) were proposed to account for non-local dependencies in NER. Non-local information is also partially propagated by phrasal and word clusterings. We implemented two approaches which take into account non-local dependencies: two-stage prediction (Krishnan and Manning, 2006) and context aggregation (Ratinov and Roth, 2009).

Two-stage prediction is an approach in which we use the output of a first recognizer to train a second one. For instance, document-based and corpus-based statistic of given token labels is used to re-assign a label to a token (Krishnan and Manning, 2006).

The idea of context aggregation (Ratinov and Roth, 2009) is that if a current token occurs more than once within a window of 200 tokens, we add features to the current token. The features are previous, next, and current tokens of all those extra occurrences. We also performed aggregation of cluster labels for all word and phrasal clusterings that we considered.

We have not performed a separate evaluation of non-local dependencies and tested them only in the final system.

4.4 Minor Features

If we combine the features discussed above (except the non-local dependencies) we get a drastic performance improvement (see Table 5). However, we developed features which correct common errors found on the development and training sets of our benchmarks. Those features were (1) hyphen feature that indicates if a particular token contains a hyphen; (2) sub-tokens feature

Features	Dev	Test
CoNLL-2003		
All features	93.78%	91.02%
NLPBA 2004		
All features	-	74.27%

Table 5: Evaluation of feature combinations.

that adds all sub-tokens of a current token which is hyphened, e.g. *Moscow-based* has sub-tokens *Moscow* and *based*; (3) text-break (expected and unexpected line breaks) feature capturing splits in text; (4) numbers generalization feature, we considered masks of numbers instead of specific numbers according to (Ratinov and Roth, 2009), e.g. 10-12-1996 → *DD*-*DD*-*DDDD*, 1999 → *DDDD*; (5) a conjunction of the Brown cluster of a current token with the preceding token; (6) capitalized context, which captures additional context of capitalized words, namely, we add a feature that encodes two previous tokens. We use an umbrella term *minor features* to describe this error-fixing list of features.

We also added a two stage prediction in which the first CRF (tuned for a specified task), the boundary recognizer, tries to detect entity boundaries and the second CRF utilizes the output of the first CRF and considers all tokens inside an entity. For example, if we have a phrase ... *blah-blah the Test and County Board Cricket blah-blah-blah* ... and the boundary recognizer has detected that *Test and County Board Cricket* is a potential entity, the second CRF adds all tokens of the potential entity as features when it classifies each token within the entity.

The combination of all proposed features is shown in Table 5. We tested the two-stage prediction approach on this configuration but have not found improvements.

5 Final System

Our final system utilizes all above mentioned features except the two-stage prediction. Each feature set improves performance of the recognizer. We tried to perform optimization by deleting features one by one in order to get the best performing configuration with a smaller set of features. We find that the sequence of deletion steps de-

pends on the initial search space (e.g. if we start optimization procedure without Clark clusters, it will delete the text-break feature; otherwise, it will delete hyphen and sub-tokens features). Table 6 shows the quality of the system with particular features omitted. You can see that the performance of the recognizer is not dramatically reduced in most cases. We believe that it is possible to come up with a smaller feature space or to do feature reweighing (Jiang and Zhai, 2006) in order to improve NER quality and processing speed.

Most of considered features are local and are extracted from a token or its local context. First of all, the behavior of context tokens as features is preserved for both datasets. A small sliding window of three tokens is good enough. Second, the word-based features behavior is not persistent and depends on the specificity of entities. Nevertheless, names contain morphological clues that distinguish them from common words. Comparing token-based with word-based features you might see that token-derived information gives a gain of at least four points of F_1 -measure for newswire corpus and can be on the same level for biomedical domain. Third, clustering could be considered as feature reduction process (Saha et al., 2009); it helps to overcome the lack of statistics. Using only clustering representations hypothesis on the reduced space of features can be useful in recognition and works even better than token-based features. Last but not least, gazetteers are still useful for NER, especially when we have such freely available resources as Wikipedia and DB-Pedia. Disambiguation approaches in gazetteer matching could bring radical improvements.

Two tables compare our results with the best reported systems on the CoNLL 2003 (Table 7), OntoNotes version 4 (Table 8), and NLPBA (Table9) datasets.

We used approximate randomization test (Yeh, 2000) with 100000 iterations to compare our system to (Ratinov and Roth, 2009). The test checks if a randomly sampled mixture of the outputs of the baseline algorithm and the one being tested performs better than the baseline algorithm. Our improvement over the top performing competitor is statistically significant with p-value 0.0001. Unfortunately, we could not compare with (Lin and Wu, 2009) because their system uses propri-

Feature	Dev	Test	Test+
—	93.78	91.02	74.27
Capitalized context*	93.82	90.66	74.24
Clark aggregation	93.80	90.66	74.01
Hyphen*	93.78	90.84	74.11
Brown + token*	93.66	90.79	74.22
Sub-tokens*	93.66	90.74	74.20
POS tag	93.65	90.82	74.07
Numbers gen.*	93.65	90.65	74.25
Text break features*	93.60	90.68	74.06
Additional gazetteers	93.56	90.17	74.24
Context aggregation	93.55	90.42	74.10
Brown aggregation	93.52	90.29	74.27
Current token	93.51	90.74	74.21
Clark cluster	93.49	90.35	74.23
Affixes	93.47	90.63	74.08
DBpedia gazetteers	93.46	90.53	74.27
Tokens in window	93.35	90.37	74.30
LDA cluster	93.34	90.46	74.18
Brown cluster	93.26	90.25	74.20
Phrasal cluster	93.12	90.43	74.75
Tokens in entity	93.12	90.43	74.11
Wikipedia gazetteers	93.12	90.43	n/a
Shape	92.83	89.75	74.02

Table 6: Evaluation of omitting of features on the CoNLL 2003 development (Dev) and test (Test) sets and on NLPBA test set (Test+). All F_1 values are in %. “*” indicates minor feature

	Finkel	Ratinov	Our system
ABC	74.91%	72.84%	76.75%
CNN	78.70%	79.27%	81.40%
MNB	66.49%	73.10%	71.52%
NBC	67.96%	65.78%	67.41%
PRI	86.34%	79.63%	83.72%
VOA	88.18%	84.93%	87.12%

Table 8: Comparison of F_1 -measures of recognizers on the OntoNotes version 4 benchmark. Finkel — (Finkel and Manning, 2009); Ratinov — (Ratinov and Roth, 2009)

System	F_1 -measure
(Wang et al., 2008)	77.6%
Our system	74.27%
(Zhou and Su, 2004)	72.6%
(Finkel et al., 2004)	70.1%
(Settles, 2004)	69.8%
(Saha et al., 2009)	67.4%

Table 9: Comparison of recognizers on NLPBA 2004 benchmark.

etary data and its output is also unavailable.

6 Conclusions

In this paper we have analyzed a comprehensive set of features used in supervised NER. We have considered the impact of various individual features and their combinations on the effectiveness of NER. We have also built a CRF-based supervised NER system that achieves 91.02% F_1 -measure on the CoNLL 2003 dataset and 81.4% F_1 -measure on the OntoNotes version 4 CNN dataset and demonstrated that the performance boost over the earlier top performing system is statistically significant on the benchmarks. We have also considered novel features for NER, namely a DBpedia markup and a phrasal clustering from Google n-grams corpus.

We plan to extend the work on clustering features which we find very promising for NER. We have obtained a large proprietary newswire corpus from a media corporation and plan to utilize it in our further experiments on enhancing NER in the newswire domain. We also consider exploring features useful for specific entity classes.

System	F_1 -measure
Our system	91.02%
(Lin and Wu, 2009)	90.90%
(Ratinov and Roth, 2009)	90.80%
(Ciaramita and Altun, 2005)	90.80%
Tjong Kim Sang 2003	90.30%
(Suzuki and Isozaki, 2008)	89.92%
(Ando and Zhang, 2005)	89.31%
(Florian et al., 2003)	88.76%

Table 7: Comparison of recognizers on the CoNLL 2003 benchmark. Tjong Kim Sang stands for (Tjong Kim Sang and De Meulder, 2003).

References

- Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL*. The Association for Computer Linguistics.
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165.
- Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.
- Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: how far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.
- M. Ciaramita and Y. Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*.
- Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL ’03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jenny R. Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL ’09, pages 326–334, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. 2004. Exploiting context for biomedical entity recognition: From syntax to the web. In *In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications (JNLPBA-2004)*.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL ’05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada.
- Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90 In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.
- Jing Jiang and Chengxiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *In Human Language Technology Conference*, pages 74–81.
- James R. Joel Nothman. 2008. Transforming Wikipedia into Named Entity Training Data. pages 124–132.
- Jun’ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.
- Jin D. Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA ’04, pages 70–75, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 180–183. Edmonton, Canada.
- Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. In *Proceedings of*

- the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1030–1038, Suntec, Singapore, August. Association for Computational Linguistics.
- Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New Tools for Web-Scale N-grams.
- Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.
- Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
- David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification.
- David Nadeau. 2007. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision*. Ph.D. thesis, Ottawa, Ont., Canada, Canada. AAINR49385.
- L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6.
- Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2009. Feature selection techniques for maximum entropy based biomedical named entity recognition. *J. of Biomedical Informatics*, 42:905–911, October.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA. Association for Computational Linguistics.
- Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, JNLPBA '04*, pages 104–107, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Valentin I. Spitkovsky, Hiyan Alshawi, Angel X. Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*.
- Charles Sutton and Andrew McCallum. 2006. An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.
- Jun Suzuki and Hideki Isozaki. 2008. Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data. In *Proceedings of ACL-08: HLT*, pages 665–673, Columbus, Ohio, June. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.
- Maksim Tkatchenko, Alexander Ulanov, and Andrey Simanovsky. 2011. Classifying wikipedia entities into fine-grained classes. In *ICDE Workshops*, pages 212–217.
- Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Haochang Wang, Tiejun Zhao, Hongye Tan, and Shu Zhang. 2008. Biomedical named entity recognition based on classifiers ensemble. *IJCSA*, 5(2):1–11.
- Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences.
- GuoDong Zhou and Jian Su. 2004. Exploring deep knowledge resources in biomedical name recognition. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 99–102, Geneva, Switzerland, August 28th and 29th. COLING.

Mention Detection: First Steps in the Development of a Basque Coreference Resolution System

Ander Soraluze, Olatz Arregi, Xabier Arregi, Klara Ceberio and Arantza Díaz de Ilarrazá
IXA Group. University of the Basque Country
ander.soraluze@ehu.es

Abstract

This paper presents the first steps in the development of a Basque coreference resolution system. We propose a mention detector system based on a linguistic study of the nature of mentions. The system identifies mentions that are potential candidates to be part of coreference chains in Basque written texts. The mention detector is rule-based and has been implemented using finite state technology. It achieves a F-measure of 77.58% under the Exact Matching protocol and of 82.81% under Lenient Matching.

1 Introduction

Coreference resolution is the key to understanding texts, and is therefore crucial in advanced NLP applications where a higher level of comprehension of the discourse leads to better performance, such as in Information Extraction (McCarthy and Lehnert, 1995), Text Summarisation (Steinberger et al., 2007), Question Answering (Vicedo and Ferrández, 2006), Machine Translation (Peral et al., 1999), Sentiment Analysis (Nicolov et al., 2008) and Machine Reading (Poon et al., 2010).

In coreference resolution, an entity is an object or set of objects in the world, while a mention is the textual reference to an entity (Doddington et al., 2004).

It is very common to divide the coreference resolution task into two main subtasks: mention detection and resolution of references (Pradhan et al., 2011). Mention detection is concerned

with identifying potential mentions of entities in the text and resolution of references involves determining which mentions refer to the same entity. Although Mention Detection has close ties to named-entity recognition (NER henceforth), it is more general and complex task than NER because besides named mentions, nominal and pronominal textual references also have to be identified (Florian et al., 2010).

Our goal is to create an accurate mention detector as the basis for developing an end-to-end coreference resolution system for Basque. We especially emphasise a linguistically sophisticated understanding of the concept of mention, which we then encode by means of regular expressions using finite state technology.

As Basque is a less resourced language, there is a considerable lack of linguistic resources, which makes it particularly challenging to develop highly accurate tools for tasks like mention detection.

This paper is structured as follows. After reviewing related work in section 2, we describe in section 3 the linguistic features related to mentions in Basque texts. Section 4 presents an overview of the system, describing the resources used in preprocessing and the rules defined for mention detection. The main experimental results are outlined in section 5 and analysed in section 6. Finally, we review the main conclusions and preview future work.

2 Related Work

Coreference resolution has received some attention in the context of overall information extrac-

tion. It was first included as a specific task in the programmes of the Sixth and Seventh Message Understanding Conferences (MUC-6, 1995; MUC-7, 1998). In addition, during the period 2000–2001, the Automatic Content Extraction (ACE) program effort was devoted to Entity Detection and Tracking (EDT), which consists of finding and collecting all mentions of an entity into equivalence classes that represent references to the same entity.

However, MUC and ACE were specifically designed for shared tasks on information extraction, and thus annotation decisions regarding coreferent elements were designed to task needs at the cost of linguistic accuracy (van Deemter and Kibble, 1995; Recasens, 2010).

Recently, conferences exclusively devoted to coreference resolution have been organised. SemEval-2010 Task 1 was dedicated to coreference resolution in multiple languages (Recasens et al., 2010). One year later, in the CoNLL-2011 shared task (Pradhan et al., 2011), participants had to model unrestricted coreference in OntoNotes corpora (Pradhan et al., 2007).

To resolve coreferences, one must first detect the mentions that are going to be linked in coreference chains. As several researchers point out (Stoyanov et al., 2009; Hacioglu et al., 2005; Zheкова and Kübler, 2010), the mention detection step is crucial for the accuracy of end-to-end coreference resolution systems. Errors in mention detection propagate and reduce the level of accuracy in the performance of subsequent steps. Therefore, improving the ability of coreference resolvers to identify mentions would likely improve the state-of-the-art, as indicated by studies where the impact of mention detection for the performance of a coreference resolution system has been quantified. For example, Uryupina (2008) reports that 35% of recall errors in their coreference resolution system are caused by missing mentions and in Uryupina (2010) adds that 20% of precision errors are due to inaccurate mention detection. Similarly, in (Chang et al., 2011), a system that uses gold mentions outperforms one using predicted or system mentions by a large margin, from 15% to 18% in F1 score, while Kim et al. (2011) point out that mention detection is also essential in specialised domains like

biomedicine. They further observe that using gold mentions versus system mentions can change the coreference resolution performance substantially in terms of the MUC score, from 87.32% to 49.69%.

Mention detection itself is a very challenging task since expressions can have complex syntactic and semantic structures. Considering the influence that mention detection has in coreference resolution, it is clear that this task deserves more attention.

Concerning the technology used by mention detectors, two main types can be distinguished: rule-based approaches and machine learning models.

In SemEval-2010 the majority of systems used rule-based approaches (Zheкова and Kübler, 2010; Uryupina, 2010; Attardi et al., 2010; Broscheit et al., 2010); the same was true in CoNLL-2011, where only four systems used trained models. While trained models seem to be able to balance precision and recall, and to obtain a higher F-score on the mention detection task, their recall tends to be substantially lower than that achievable by rule-based systems. The low recall has negative repercussions for the whole coreference resolution system because the system has no way to recover missed mentions. Indeed, the best-performing system in CoNLL-2011, (Lee et al., 2011), was completely rule-based.

3 Linguistic Analysis of Mentions

Although coreference is a pragmatic linguistic phenomenon highly dependent on the situational context, it shows some language-specific patterns that vary according to the features of each language. With reference to the subtask of mention detection, in this section we establish what mentions we regard as potential ones to be included in a coreference chain.

In general, we take into account noun phrases (NP), focusing on the largest span of the NP. In the case of nouns complemented by subordinate clauses and coordination, we also extract the embedded NPs of larger NPs as possible candidates for a coreference chain. We propose the following mention classification:

1. **Pronouns:** In Basque, no separate forms ex-

ist for third person pronouns versus demonstrative determiners; demonstrative determiners are used as third person pronouns (Laka, 1996). Therefore, we mark the mentions formed by demonstratives used as pronouns.

- (a) *LDPko buruek Mori hautatu zuten apirilean Keizo Obuchi orduko lehen ministroa ordezkatzen, [hark] tronbosia izan ostean.*

“The heads of LDP chose Mori in April, to replace the Prime Minister Keizo Obuchi, [who (he)] suffered a thrombosis.”

2. **Possessives:** We consider two types of possessives: NPs containing a possessive determiner, even if it is not the head of that NP as in (b), and possessive pronouns as in (c).

- (b) *Epitieren kasuan [[bere] helburua lortu dezakela dirudi eta baliteke denboraldia Lehen Mailan hastea.*

“In the case of Epitie, it seems that he could achieve [[his] aim] and possibly start the football season in the Premier League.”

- (c) *Escuderok euskal musika tradizionala eraberritu eta indartu zuen. [Harenak] dira, esate baterako, Illeta, Pinceladas Vascas eta Eusko Salmoa obrak.*

“Escudero renewed and gave prominence to traditional Basque music. The works Illeta, Pinceladas Vascas and Eusko Salmoa, for example, are [his]”

3. **Verbal nouns:** Verbs that have been nominalised and function as the head of the mention, with the corresponding case marking suffix. The whole clause governed by the verbal noun has to be annotated.

- (d) *[Instalazio militarrak ixtea] eskatuko dute.*

“They will ask for [closing the military installations].”

4. **NPs as part of complex postpositions:** Basque has a postpositional system, and therefore we mark the independent NP that

goes before the complex postpositions. In (e) the postposition is *aurka* (“against”), and we annotate the noun (in that case, a proper noun) that precedes it.

- (e) *Joan den astean [Moriren] aurka aurkeztutako zentsura moziok piztu zuen LDPko krisia.*

“Last week the vote of no confidence against [Mori] caused the crisis in the LDP.”

5. **NPs containing subordinate clauses:** The head of these mentions is always a noun complemented by a subordinate clause. In (f) the head noun is complemented by a subordinate clause of the type that is, for Basque, called a *complementary clause*. We take the whole stretch of the NP (both the subordinate clause and the head noun) as a mention.

In addition, relative clauses can add information to nouns as in (g). In that case the boundaries of the mention are set from the beginning of the relative clause to the end of the NP.

- (f) *[DINAk Argentinan egindako krimenak ikertzeo baimena] eman du Txileko Gorte Gorenak.*

“The Supreme Court of Chile has given [permission to investigate the crimes DINA committed in Argentina].”

- (g) *[Igandeko partiduak duen garrantzia] dela eta, lasai egotea beharrezkoa dutela esan zuen Lotinak.*

“Lotina said that it is necessary to stay calm because of [the importance that Sunday’s match has].”

6. **Ellipsis:** In Basque the ellipsis is a broad phenomenon.

At a morphosyntactical level, a noun-ellipsis occurs when the suffixes attached to the word correspond to a noun, although the noun is not explicit in the word. We consider this type of ellipsis in the case of verbs that take suffixes indicating noun-ellipsis, as in example (i). The POS given by the analyser

indicates the presence of the ellipsis phenomenon, which is implied by the presence of both the verb (*sailkatu zen-* “finished”) and the ellipsis (-Ø-ak ‘who...’). All the information corresponding to both units is stored and treated as a noun.

- (i) *[Bigarren sailkatu zenak] segundo bakarra kendu zion.*

“[Ø¹ who finished in second place] only had a second’s advantage.”

At sentence level, the subject, object or indirect element of the sentence can be elided. The morphological information about these elements (number, person...) is given by the verb. We do not mark these elliptical pronouns as mentions (j).

- (j) Ø *Ez zuen podiumean izateko itxaropen handirik.*

“[He] did not have much hope of being on the podium.”

7. **Coordination:** In the case of coordination, nominal groups of a conjoined NP are extracted. We also regard as mentions the nested NPs (*siesta*, “a nap” and *atsedena* “a rest”) and the whole coordinated structure (*siesta eta atsedena* “a nap and rest”).

- (h) *Bazkal ondoren [[siesta] eta [atsedena]] besterik ez zuten egin.*
“After lunch they did nothing but have a [[nap] and [rest]].”

Finally, we want to remark that phrases composed solely of an adjective or an adverb are not included in this annotation, because the majority of coreference relations occurs between NPs.

4 System Overview

In this section, we first introduce the preprocessing tools that our mention detector uses and then explain the rules implemented to identify the mention types described in section 3.

4.1 Preprocessing

The mention detector receives as input the results of two analysers: IXAti (Aduriz and Díaz de

¹In this case Ø refers to someone.

Ilarraza, 2003), which identifies chunks based on rule-based grammars, and ML-IXAti, which identifies clauses by combining rule-based grammars and machine learning techniques (Arrieta, 2010).

Both IXAti and ML-IXAti make use of the output produced by several tools implemented in our research group ²:

- **Morphosyntactic analyser:** *Morpheus* (Alegria et al., 1996) performs word segmentation and PoS tagging. The module that identifies syntactic functions is implemented in the *Constraint Grammar* formalism (Karlsson et al., 1995).
- **Lemmatisation and syntactic function identifier:** *Eustagger* (Alegria et al., 2002) resolves the ambiguity caused at the previous phase.
- **Multi-words items identifier:** The aim is to determine which of two or more words are to be considered multi-word expressions (Alegria et al., 2004).
- **Named entity recogniser:** *Eihera* (Alegria et al., 2003) identifies and classifies named entities (person, organisation, location) in the text.

4.2 Defined Rules

Preprocessing by generic NLP tools, while helpful, did not by itself succeed in correctly setting the boundaries of potential mentions. To address this deficiency, we developed a mention detector consisting of a set of hand-crafted rules which have been compiled into Finite State Transducers (FST).

The use of Finite State Technology enables the processing of large datasets at a high processing speed and with low memory usage. Using *foma*³ (Hulden, 2009), an open source platform for finite-state automata and transducers, we defined 7 FSTs, composed of 24 hand-crafted rules.

Our FSTs match NPs and clauses provided by the preprocessing tools and identify the mentions and their boundaries. They are organised into

²<http://ixa.si.ehu.es/Ixa>

³The regular expression syntax used in *foma* can be consulted at <http://code.google.com/p/foma/>

seven categories according to the linguistic analysis described in section 3.

1. **Pronouns:** Although the IXAti tool returns most pronouns, it does miss a few. This first transducer, which is the simplest one, identifies the pronouns missed in the preprocessing step.
2. **Possessives:** This FST identifies possessive pronouns and possessive determiners that are nested in another NP. In example (b) in section 3, the NP *bere helburua* “his aim” is obtained in the preprocessing step. The FST extracts from this NP a new mention, *bere* “his”.
3. **Verbal nouns:** The FST identifies verbal nouns and sets the boundaries of mentions so as to group them with the linguistic elements related to them. The transducer first identifies a verbal noun in a sentence (*ixtea* “closing” in (d) in section 3) and sets the right side boundary after it. Then, the left side boundary is established at the closest clause tag proposed by the ML-IXAti tagger. It is worth mentioning that simple NPs that are nested in the verbal noun mention are marked by the syntactic analyser.
4. **Postpositional phrases:** In the case of complex postpositions a FST has been defined to modify the postpositional structure in order to obtain as a mention only the NP that is part of that structure. See (e) in section 3.
5. **NPs containing subordinate clauses:** When the FST finds a head noun of a NP complemented by a subordinate clause, it sets the right side boundary after the head (*baimena* “permission” in (f) in section 3). To set the left side boundary the transducer uses the closest boundary tag proposed by the clause tagger (*DINAk* “DINA”).
Relative clauses are treated in a manner similar to complementary ones: The mention boundaries are set from the beginning of the relative clause to the end of the NP.
6. **Ellipsis:** The FST uses the syntactic analysis to identify verbs with an elided noun. The

right side boundary is established after the verb and the left side boundary is obtained using the closest clause tag.

7. **Coordination:** The identification of mentions that are components of coordination structures has been the most difficult task because it is not evident in which cases additional mentions should be obtained. This FST extracts mentions surrounding coordinating conjunctions (*eta, edo...* “and, or,” etc.). It is applied only to mentions that other FSTs have identified previously and that contain a coordinating conjunction.

To better illustrate the behaviour of the mention detector we use one example to analyse the entire process of the module. Figure 1 shows the input that the system receives for the sentence *Armada britainiarak Ipar Irlandan dituen bi kuartel eta beste bi begiratoki eraitsi dituzte*. “Two military barracks and two other viewing points that the British army has in Northern Ireland have been demolished”. The figure shows the phrase tags (BNP, ENP, BVP and EVP) and clause boundaries (CB) provided by the two tools, IXAti and ML-IXAti. In addition to the NPs provided –which are themselves counted as mentions– we want to obtain from this sentence the new mention *Armada britainiarak Ipar Irlandan dituen bi kuartel eta beste bi begiratoki*.

Based on the input received, we define the rule in Figure 2 to perform relative clause detection. First, a relative verb (RV) is defined as the composition of a verb with a relative suffix. RV is used in the definition of the relative clause mention (RM).

The rule identifies the verb containing a relative suffix (*dituen* “that has”) which is tagged with VREL. Next, the right side boundary is established in the NP or coordinated NPs that follow the relative verb (*bi kuartel eta beste bi begiratoki* “Two military barracks and two other viewing points”). Finally, the left side boundary of the mention is established at the closest clause boundary ({CB}) to the left (*Armada* “army”). Following these steps, the system obtains a new correct mention and tags the whole structure to show where the mention begins (<MENTION>) and ends (</MENTION>).

IXAti	BNP	ENP	BNP	ENP	REL	BNP	ENP	PJ	BNP	ENP	BVP	EVP
ML-IXAti	{CB{CB					CB}						CB}

Figure 1: The input received by the mention detector. BNP = Begin-NP, ENP = End-NP, BVP = Begin-VP, EVP = End-VP, VREL = Relative Verb, PJ = Conjunction, CB = Clause Boundary

```
define RV Verb & $[ ``VREL'' ];
define RM [CB W+ RV NP [ [and|or] NP]*] @-> "<MENTION>" . . . "</MENTION>";
```

Figure 2: A simplified rule to recognise relative clauses. NP = Noun Phrase, CB = Clause Boundary, W=Word, RV = Relative Verb, RM = Relative Mention

5 Experimental setup

Many authors propose that mention identification and coreference resolution task evaluations should be carried out separately. Recasens (2010) notes that the task of mention identification is clearly distinct from coreference resolution, and that therefore, one single metric giving the overall result is less informative than evaluating each task individually. Separate evaluations allow for more detailed observation of the shortcomings in each task and thus can help determine whether a system performs well at identifying coreference links but poorly at detecting mention boundaries, or *vice versa*. In this vein, (Popescu-Belis et al., 2004) propose to consider mention identification its own task and to separate its evaluation from the evaluation of coreference resolution.

The evaluation of mention identification reveals how efficiently the system detects the mentions that are to be resolved in the coreference resolution step. Commonly used measures are precision, recall and F-measure. To calculate these, the set of manually annotated mentions (GOLD) and those extracted by the mention detector (SYS) are compared.

Typically mentions are considered correct if their span is within the span of the gold mention and contains the head word (Kummerfeld et al., 2011). This matching type is known as *Lenient Matching* or *Partial Matching*. However, stricter variations of evaluation metrics have also been applied. CoNLL-2011 Shared Task (Pradhan et al., 2011), for example, used the *Strict Matching* method, which considers only exact matches to be correct. We evaluated our mention detector using both *Lenient Matching* and *Strict*

Matching.

The corpus used to develop the system is part of EPEC (the Reference Corpus for the Processing of Basque) (Aduriz et al., 2006). EPEC is composed of articles published in 2000 in *Euskaldunon Egunkaria*, a Basque language newspaper. We divided the dataset used into two main parts: one part to develop the system, with 278 mentions and the other to test it, with 384 mentions. These two parts have been manually tagged by an expert linguist.

To set the baseline, we counted as mentions the chunks (except verbal ones) proposed by the chunk identifier and compared them with gold mentions. Table 1 shows the scores obtained by our mention detector under *Exact Matching* and *Lenient Matching*, respectively, in comparison with the baseline.

	Baseline			Mention Detector		
	P	R	F ₁	P	R	F ₁
EM	63.37	70.33	66.65	76.85	78.59	77.58
LM	72.01	79.75	75.65	81.96	83.97	82.81

Table 1: The baseline and system scores. EM=Exact Matching, LM=Lenient Matching

6 Discussion

Using *Exact Matching* the system obtains a F-measure of 77.58%; using *Lenient Matching* the score is considerably better, 82.81%. The difference between the scoring protocols is due to *Lenient Matching* being less strict than *Exact Matching*.

We can affirm that the improvement obtained by our mention detector is significant. The sys-

tem outperforms the baseline by 11 points when using *Exact Matching*, and by 7 points when the evaluation is carried out using *Lenient Matching*. We are positive that the improvement achieved in mention detection will benefit the whole process.

We carried out a qualitative evaluation to clarify why the precision and recall scores are so similar, and found that in the majority of cases the proposed mentions are exactly the same as the gold mentions. Therefore, we can argue that our system obtains high-quality mentions.

The qualitative evaluation also afforded us an opportunity to better understand the cause of errors committed by the mention detector. We observed that most of the errors are provoked by the automatic preprocessing tools. The main cause of these errors is that the span of the NPs returned by these tools exceed the gold mention's boundaries. It is obvious that these mentions will be penalised using the two scoring protocols.

Our system uses automatically processed resources; neither gold syntactic analysis nor gold NP boundaries are provided. The use of gold resources would lead to near-perfect performance in the mention detection task. Nevertheless, our main goal is to create an end-to-end coreference resolution system able to perform the entire process even when no gold resources are provided.

7 Conclusions and Future Work

We present a mention detector implemented by means of finite state transducers. The mention detector is based in a deep linguistic analysis of mentions in Basque. As many authors have affirmed, the mention detection task is crucial to the performance of a coreference resolution system. Yet the question of defining a mention is usually treated only superficially. Our hypothesis was that carrying out a detailed linguistic study of mentions and properly defining the linguistic features of mentions would produce improved results.

The scores our system achieves are very promising, especially considering that they represent the first steps in Basque coreference resolution. The scores are 77.58% under the *Exact Matching* scoring protocol and 82.81% under the *Lenient Matching* protocol.

In the future we aim to enhance the performance of the mention detector by defining a num-

ber of specific rules able to obtain mention boundaries more accurately. In addition, once IXAti and ML-IXAti are improved, our mention detector's performance will also improve substantially because the majority of errors in our system are caused by the input provided by these two tools.

We also intend to use the mention detector to automatically tag the entire EPEC corpus. The automatic tagging process will be post-edited by expert linguists and the mentions corrected, thus resulting in a gold standard corpus. We expect this corpus to be a valuable resource for taking the next steps in coreference resolution.

Acknowledgments

This work has been supported by Ander Soraluze's PhD grant from Euskara Errektoreordetza, the University of the Basque Country (UPV/EHU).

References

- Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., Diaz-De-Illarraz, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. (2006). Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. pages 1–15. Rodopi. Book series: Language and Computers.
- Aduriz, I. and Díaz de Ilarraz, A. (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. *Inquiries into the lexicon-syntax relations in Basque*, pages 1–21. University of the Basque Country.
- Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004). Representation and Treatment of Multiword Expressions in Basque. In *ACL workshop on Multiword Expressions*, pages 48–55.
- Alegria, I., Artola, X., Sarasola, K., and Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203.
- Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de*

- Información*, (JOTRI 2003), pages 198–203, Madrid, Spain.
- Alegria, I. n., Aranzabe, M., Ezeiza, N., Ezeiza, A., and Urizar, R. (2002). Using finite state technology in natural language processing of basque. In *Implementation and Application of Automata*, volume 2494 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin / Heidelberg.
- Arrieta, B. (2010). *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean*. PhD thesis, University of the Basque Country.
- Attardi, G., Rossi, S. D., and Simi, M. (2010). TANL-1: Coreference resolution by parse analysis and similarity clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 108–111, Stroudsburg, PA, USA.
- Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Verstey, Y., and Zanoli, R. (2010). BART: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 104–107, Stroudsburg, PA, USA.
- Chang, K.-W., Samdani, R., Rozovskaya, A., Rizzolo, N., Sammons, M., and Roth, D. (2011). Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 40–44, Stroudsburg, PA, USA.
- Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation. In *Proceedings of Language Resources and Evaluation Conference*, (LREC 2004), pages 837–840.
- Florian, R., Pitrelli, J. F., Roukos, S., and Zitouni, I. (2010). Improving mention detection robustness to noisy input. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (EMNLP 2010), pages 335–345, Stroudsburg, PA, USA.
- Hacioglu, K., Douglas, B., and Chen, Y. (2005). Detection of entity mentions occurring in english and chinese text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (HLT 2005), pages 379–386, Stroudsburg, PA, USA.
- Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (EACL 2009), pages 29–32, Stroudsburg, PA, USA.
- Karlsson, F., Voutilainen, A., Heikkila, J., and Antila, A. (1995). *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.
- Kim, Y., Riloff, E., and Gilbert, N. (2011). The Taming of Reconcile as a Biomedical Coreference Resolver. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 89–93, Portland, Oregon, USA.
- Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. (2011). Mention Detection: Heuristics for the OntoNotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 102–106, Stroudsburg, PA, USA.
- Laka, I. (1996). A Brief Grammar of Euskara, the Basque Language. <http://www.ehu.es/grammar>. University of the Basque Country.
- Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford’s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 28–34, Stroudsburg, PA, USA.
- McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for conference resolution. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, (IJCAI 1995), pages 1050–1055, San Francisco, CA, USA.
- MUC-6 (1995). Coreference Task Definition (v2.3, 8 Sep 95). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 335–344, Columbia, Maryland, USA.
- MUC-7 (1998). Coreference Task Definition

- (v3.0, 13 Jul 97). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.
- Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment Analysis: Does Coreference Matter? In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, pages 37–40.
- Peral, J., Palomar, M., and Ferrández, A. (1999). Coreference-oriented interlingual slot structure & machine translation. In *Proceedings of the Workshop on Coreference and its Applications*, (CorefApp 1999), pages 69–76, Stroudsburg, PA, USA.
- Poon, H., Christensen, J., Domingos, P., Etzioni, O., Hoffmann, R., Kiddon, C., Lin, T., Ling, X., Mausam, Ritter, A., Schoenmakers, S., Soderland, S., Weld, D., Wu, F., and Zhang, C. (2010). Machine reading at the University of Washington. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, (FAM-LbR 2010), pages 87–95, Stroudsburg, PA, USA.
- Popescu-Belis, A., Rigouste, L., Salmon-Alt, S., and Romary, L. (2004). Online Evaluation of Coreference Resolution. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1507–1510, Lisbon, Portugal.
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 1–27, Stroudsburg, PA, USA.
- Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, (ICSC 2007), pages 517–526, Washington, DC, USA. IEEE Computer Society.
- Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation*. PhD thesis, University of Barcelona.
- Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 1–8, Stroudsburg, PA, USA.
- Steinberger, J., Poesio, M., Kabadjov, M. A., and Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680.
- Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore.
- Uryupina, O. (2008). Error Analysis for Learning-based Coreference Resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Uryupina, O. (2010). Corry: A system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 100–103, Uppsala, Sweden. Association for Computational Linguistics.
- van Deemter, K. and Kibble, R. (1995). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26:629–637.
- Vicedo, J. and Ferrández, A. (2006). Coreference In Q&A. In *Advances in Open Domain Question Answering*, volume 32 of *Text, Speech and Language Technology*, pages 71–96. Springer.
- Zhekova, D. and Kübler, S. (2010). UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 96–99, Stroudsburg, PA, USA.

Phonetically aided Syntactic Parsing of Spoken Language

Zeehan Ahmed, Peter Cahill and Julie Carson-Berndsen
Centre for Next Generation Localisation (CNGL)
School of Computer Science and Informatics
University College Dublin, Ireland
zeehan.ahmed@ucdconnect.ie,
{peter.cahill,julie.berndsen}@ucd.ie

Abstract

The paper presents a technique for parsing a speech utterance from its phonetic representation. The technique is different from a conventional spoken language parsing techniques where a speech utterance is first transcribed at word-level and a syntactic structure is produced from the transcribed words. In a word-level parsing approach, an error caused by a speech recognizer propagates through the parser into the resultant syntactic structure. Furthermore, sometimes transcribed speech utterances are not parseable even though lattices or confusion networks are used. These problems are addressed by the proposed phonetically aided parser. In the phonetically aided parsing approach, the parsing is performed from a phonetic representation (phone sequence) of the recognized utterance using a joint modeling of probabilistic context free grammars and a n-gram language model. The technique results in better parsing accuracy than word-level parsing when evaluated on spoken dialog parsing task in this paper.

1 Introduction

Syntactic parsing is an important step towards an effective understanding of a language. It is widely used in major natural language processing tasks such as machine translation, information extraction & retrieval, language understanding etc. Parsing has also been tried as a language model in speech recognition because of the fact that long distance syntactic constraints produced as a result of parsing are stronger than n-gram

language model constraints. As a result, parsing can improve speech recognition results. However, parsing natural language requires sophisticated resources like phrase-structure grammar, dependency grammar, link grammar, categorical grammar etc. and a parser for these grammars.

In terms of complexity (time, space, and ambiguity), text is much easier to parse than speech. Apart from time and space complexity, ambiguity (confusion) in the speech signal greatly affects the processing and the results. Such confusions in speech signals are mostly handled with phonetic models and language models. The n-gram language model has been widely used for this purpose in large vocabulary speech recognition and translation systems. However, failure to capture long distance relationships and the problem associated with sparse data are the major drawbacks for the n-gram model.

This paper presents a technique for spoken language parsing from a phone sequence as opposed to conventional approach of parsing from a word sequence. The ultimate objective of this work is the integration of syntax into phonetic representation-based speech translation (Jiang et al., 2011). In this work, the role of syntax is analyzed with respect to the improvement in source-side language recognition and computational requirements. In our hypothesis, parsing may act as a language model constraint over phonetic space that can result in improvement in recognition errors as well as syntactic accuracy which is directly related to translation quality.

In this paper, the phonetic knowledge is used as an aid to word-based parsing technique to recover

from recognition error caused by a speech recognizer. To parse a speech utterance from a phone sequence, the probabilistic context free grammars (PCFG) are extended with phonetic knowledge called a probabilistic phonetic grammar (PPG). The PPG alone is not good at parsing a 1-best phone sequence generated by a phone recognizer because of the inherent errors made by the recognizer. A phone confusion network (PCN) could be a better choice for this model. However, the PCN makes the PPG more ambiguous and leads to decrease in the performance of the system. Therefore, the parser for PPG is augmented with n-gram language model to gain better accuracy. The PPG + N-gram model is then referred to as the joint parsing model in this paper. The proposed parsing approach has multiple advantages.

- it allows the syntactic model to be applied at phonetic-level which improves the recognition rate.
- it facilitates the parsing of utterances when it is not possible to parse using a word-based parser due to syntactic errors in recognition.
- N-gram and syntactic language models can be simultaneously applied on the phone sequence for better recognition and syntactic accuracy.

The proposed parsing model takes speech as input in the form of a phone confusion network (PCN) and produces a word sequence, a syntactic parse structure or both corresponding to the speech utterance. The joint parsing model is applied on syntactic parsing of spoken dialogs for evaluation. The model allows the effective parsing of highly ambiguous confusion networks which results in better performance than the state-of-the-art. The conceptual architecture of the systems is shown in figure 1.

The rest of the paper is organized as follows. The next section presents syntactic parsing as a language model in speech recognition and describes different techniques for syntactic parsing of speech. Section 3 describes the phonetic grammars. Section 4 presents the proposed joint parsing model. Section 5 presents the comprehensive evaluation and description of the spoken dialogs

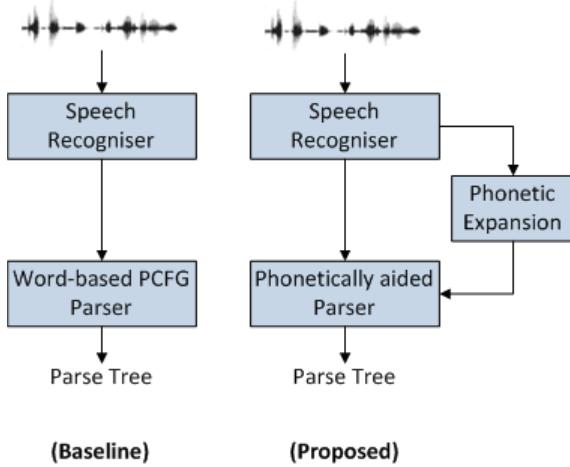


Figure 1: Conceptual Architecture of Baseline and Proposed Systems.

parsing task. Conclusions are finally drawn in section 6.

2 Parsing as a Language Model

Syntactic parsers are typically designed for the purpose of text processing, where the input is deterministic (i.e. at any point during parsing, the following word is known with certainty). Parsing spoken language is more complex than text parsing because of the fact that input is not deterministic. For spoken language parsing, the parser is employed not only to produce the syntactic structure but also to generate the correct word sequence.

Parsing of spoken language for further language processing is performed in two steps (Hall and Johnson, 2004); first a speech recognizer is used to transcribe speech into a word sequence and then a text-based parser is used to produce a syntactic structure. This method, however, fails to perform when there are even a few errors in the recognition output. The alternative approach is to generate a N-best list, a speech lattice or a confusion network and let the parser decide the correct word sequence according to the parsing constraints. This way of using parsing is referred to as syntactic language modeling in speech recognition.

Syntactic language modeling has been widely used for speech recognition. The syntactic language model can be implemented using gram-

mar network or PCFG. Grammar network is the simplest approach and can be applicable only for small vocabulary command and control tasks. PCFG on the other-hand can be used for large vocabulary recognition tasks. The work of (Chelba and Jelinek, 1998) is very prominent for syntactic language modeling in speech recognition. In (Chelba and Jelinek, 1998), the syntactic information has been applied for capturing the long distance relationship between the words. The lexical items are identified in the left context which are then modeled as a n-gram process. Reduction in WER is reported in (Chelba, 2000) by re-scoring word-lattices using scores of a structured language model. Collins (Collins et al., 2004) model of head-driven parsing also reports that the head-driven model is competitive with the standard n-gram language model. Lexicalized probabilistic top-down parsing has been tried as a language model for re-scoring word-lattices in (Roark, 2001). Different variants of bottom up chart parsing has also been applied for re-scoring word-lattices for speech recognition (Hall and Johnson, 2003; Chappelier and Rajman, 1998). (Hockey and Rayner, 2005) reports that even for non-sparse data, PCFGs can perform better than n-gram language models for both speech recognition and understanding tasks.

An example for parsing lattices and confusion networks (CN) has been provided in (Chappelier et al., 1999). Unfortunately, lattice parsing is more difficult than CN parsing in terms of both time and space complexity. According to (Chappelier et al., 1999), the simplest approach to parse a lattice is to find the topological ordering of the nodes and allocate the chart space for CYK parsing (Tomita, 1985) for the number of nodes in the lattice. However, this approach seems impractical even for an average size lattice with a couple of thousand nodes. In another recently proposed approach (Köprü and Yazıcı, 2009), the lattice is treated as a finite state network which is first determinized and minimized and then, with better chart initialization, a much larger lattice can be parsed. However, the approach still fails in some cases as described in the paper. Another reasonable approach for lattice parsing is presented in (Goldberg and Elhadad, 2011) where nodes are indexed from their start position and

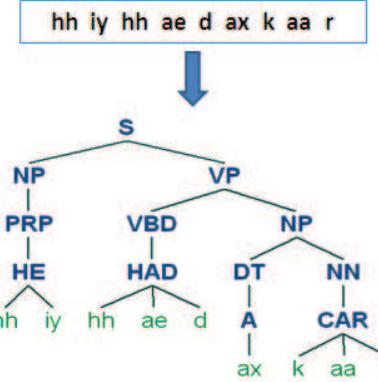


Figure 2: Parsing phone sequence with PPG

then the chart is initialized according to node position. Lexicalized grammar (Collins et al., 2004; Klein and Manning, 2003; Charniak, 2001) is also a more sophisticated approach for parsing lattices.

CN parsing, on the other hand, is simpler than lattice parsing. However, compared to a lattice, it adds additional hypotheses to the search space which are not presented in the original lattice. This might lead to incorrect results during parsing. However, the proposed joint parsing model places tight constraints over CN parsing and results in better hypothesis search than simple confusion network parsing.

3 Phonetic Grammars

The probabilistic phonetic grammar (PPG) is the same as a PCFG except that it is extended with phonetic knowledge as shown in figure 2. The PPG has a form similar to a PCFG i.e. $\langle p, A \rightarrow \alpha \rangle$ where p is the probability of the production and the probability of all the productions having non-terminal A on the left hand side sum to 1.

In a PCFG, the probability of a parse tree is calculated as the product of individual probabilities of productions. Suppose that

$$S \xrightarrow{r_1} \alpha_1 \xrightarrow{r_2} \alpha_2 \dots \xrightarrow{r_n} \alpha_n = w$$

is a derivation of a sentence w from the start symbol S , then the probability of this derivation D is given by

$$P(D) = \prod_{i=1}^n P(r_i) \quad (1)$$

G			
(1.0)	S	\rightarrow	NP VP
(0.7)	NP	\rightarrow	PRP
(0.3)	NP	\rightarrow	DT NN
(0.5)	PRP	\rightarrow	he
(0.5)	NN	\rightarrow	car
(1.0)	DT	\rightarrow	a
(1.0)	VP	\rightarrow	VBD NP
(1.0)	VBD	\rightarrow	had

G_p			
(1.0)	S	\rightarrow	NP VP
(0.7)	NP	\rightarrow	PRP
(0.3)	NP	\rightarrow	DT NN
(0.5)	PRP	\rightarrow	HE
(1.0)	HE	\rightarrow	hh iy
(0.5)	NN	\rightarrow	CAR
(1.0)	CAR	\rightarrow	k aa r
(1.0)	DT	\rightarrow	A
(1.0)	A	\rightarrow	ax
(1.0)	VP	\rightarrow	VBD NP
(1.0)	VBD	\rightarrow	HAD
(1.0)	HAD	\rightarrow	hh ae d

Table 1: PCFG G and augmented PPG G_p from G .

The probability of w is then the sum of the probabilities of all derivations of w i.e.

$$P(w) = \sum_D P(D) \quad (2)$$

The probability $P(r_i)$ can be estimated using either a maximum likelihood method (MLE) in a supervised manner or using an unsupervised method like the inside-outside algorithm (Baker, 1979). The MLE technique is used in this paper for PPG estimation. Given a corpus annotated on the syntactic level, the probability of each production can be estimated from the corpus as follows

$$P(A \rightarrow \alpha_i) = \frac{C(A \rightarrow \alpha_i)}{\sum_i C(A \rightarrow \alpha_i)} \quad (3)$$

3.1 Parsing With PPG

Parsing a phone sequence using PPG is similar to parsing any sentence using PCFG. For example, consider the grammar G in table 1. For phone sequence parsing, the pronunciation model needs to be incorporated into the grammar G . The grammar G is augmented into grammar G_p ¹ using the pronunciation dictionary as shown in table 1. The grammar G_p can then be used for phone sequence parsing as shown in figure 2.

¹The symbols represented in lowercase are the phonetic representation of sounds called phones.

3.2 Obtaining Phonetic Knowledge

The phone sequence of speech can be derived from either using general purpose phone recognizer or converting the word recognition output into phones. The difference here is the language model applied in the recognition process which actually affects the phone recognition rate for the subsequent spoken language processing. For the first approach, a higher order phone language model is preferred for a better phone recognition rate (Bertoldi et al., 2008). This approach can be beneficial for the languages which do not have diversity in pronunciation system and do not differ considerably in orthographic and pronunciation systems. While, in the second approach, the phonetic knowledge can be used as an aid to word-based parsing. The languages like English, which have vast diversity in pronunciation systems can benefit from this approach. As this work uses phonetic knowledge as an aid to word-based parsing, the second approach is followed for obtaining phonetic knowledge.

3.3 Dealing with Phonetic Confusion

The biggest problem with parsing using only the dictionary pronunciation in G_p is that 1-best recognized outputs are not 100% correct which makes the grammar G_p impractical for phone sequence parsing. A little pronunciation variation/error in the example input sentence will cause the sentence to be rejected by G_p . For this purpose, the phone confusion matrix (PCM) approach presented in (Bertoldi et al., 2008; Jiang et al., 2011) is used to transform the 1-best phone sequence into phone confusion network (PCN).

The PCM is extracted by aligning the recognition outputs of the development set with transcriptions and then calculating the confusion score (insertion/deletion/substitution) for each phone as given in equation 4. In this paper, only insertions and substitutions are taken into account for PCN generation.

$$Conf(i, j) = \frac{M_{ij}}{\sum_i M_{ij}} \quad (4)$$

where $\sum_i M_{ij}$ is the total number of time j appears in the transcriptions, and M_{ij} is the times that phone i is aligned with phone j . Pruning is performed during PCN generation. Any

phonetic confusion that has the confusion score $Conf(i, j)$ less than particular confusion threshold (CT) limit, is not included in the final PCN. Experiments are performed for different CT values in this paper.

4 Joint Parsing Model

The PPG alone is not good enough to handle confusion information present in PCN. The PPG parsing is further extended with an n-gram statistical model because of its ability to handle lexical relationships between words effectively. The proposed model is referred to as the joint parsing model. In the proposed joint parsing model, the probability of a sentence is calculated using a joint probability of PPG and n-gram probability distributions i.e. for a sentence S the joint probability $P(\hat{S})$ is given by

$$\begin{aligned} P(\hat{S}) &\approx \operatorname{argmax}_S P(S, D_S) \\ &\approx \operatorname{argmax}_S P(S) * P(D_S) \end{aligned} \quad (5)$$

where, $P(S)$ is the probability provided by n-gram language model i.e. if w_1, w_2, \dots, w_m represents the sequence of words in sentence S , then

$$P(S) = \prod_{i=1}^m P(w_i | w_{i-(n-1)}, \dots, w_{i-1}) \quad (6)$$

and, $P(D_S)$ is the probability of derivation of the sentence S provided by the PPG model as described by equation 1.

4.1 Parsing Algorithm

Parsing a phone sequence with the joint parsing model is similar to parsing with a PPG with additional computation of n-gram probability at each sub-tree formation. This type of parsing is normally used in syntax-based machine translation systems (Weese et al., 2011; Dyer et al., 2010) where tree is built for source language, target language or both and n-gram is applied on target side. With the additional computation of n-gram probability at each sub-tree formation, the algorithm needs to keep the left and right context (n-1 lexical item) of the sub-tree. The run time complexity of the joint parsing model then increases

by a factor of $O(n - 1)$ for CYK parsing algorithm where n is the n-gram size. It is because at every sub-tree formation, only $(n - 1)$ operations are needed to compute the local sub-tree n-gram probability.

Therefore, in its simplest form, the PCFG parsing with CYK algorithm has the run-time complexity of $O(m^3 * |G|)$ where m is the length of sentence and $|G|$ is the size of grammar. In this algorithm, it is assumed that each cell of the CYK grid contains only top scoring unique non-terminals. It is because PCFGs are normally ambiguous and can have multiple possibilities for driving same left-hand side non-terminal. If multiple possibilities are considered for every unique non-terminal in each cell then the accuracy of system may improve but the complexity of the algorithm becomes exponential. Such case is avoided here to keep the algorithm within computational limits. Finally, the run time complexity of the joint parsing model employed here becomes

$$O((n - 1) * m^3 * |G|)$$

5 Evaluation

The technique is evaluated on the IWSLT 2010 corpus². The corpus contains spoken dialogs related to the travel domain. The corpus is composed of three datasets; training, development and test sets. The selected training set contains 19,972 sentences and development set contains 749 sentences which is used for calculating PCM. While, the test set contains 453 sentences and it comes from 1-best ASR output having word error rate (WER) of 18.3%.

IWSLT 2010 is a bilingual corpus (English-Chinese) for a speech translation task. The focus of the work is on the English side of the corpus. In this experiment, the objective is to use the phonetically aided parsing technique for parsing IWSLT 2010 spoken dialogs and compare the performance of the technique with state-of-the-art word-based parsing systems. For comparison, the systems are evaluated based on WER (for measuring the quality of recognized sentence) and F-measure (for measuring the quality of syntactic structure). Three systems are developed in this re-

²<http://iwslt2010.fbk.eu/>

1-best Word Parsing			
W.E.R	Precision	Recall	F-measure
23.6	41.33	41.48	41.40

N-best Word Parsing			
W.E.R	Precision	Recall	F-measure
23.2	39.14	40.38	39.75

(a) 1-best and N-best parsing results.

CT	PPG Parsing				Joint PPG and N-gram Parsing			
	W.E.R	Precision	Recall	F-measure	W.E.R	Precision	Recall	F-measure
1.000	21.6	42.15	42.23	42.19	21.0	42.65	42.81	42.73
0.100	21.2	42.50	42.42	42.46	20.5	43.08	43.10	43.09
0.050	22.2	42.80	42.02	42.41	20.5	43.06	43.08	43.07
0.010	23.6	43.68	38.83	41.11	17.9	43.65	43.66	43.65
0.005	23.7	43.57	38.57	40.92	17.8	43.76	43.74	43.75
0.001	23.7	43.45	38.50	40.83	17.8	43.56	43.54	43.55

(b) PCN Parsing Results.

Table 2: Parsing Results on IWSLT 2010 Spoken Dialogs.

gard; a word-based PCFG parsing system, a PPG parsing system and a joint parsing model system.

5.0.1 Word-based PCFG Parsing System

The word-based PCFG is extracted from the training set. Since the syntactic annotations are required for extracting PCFG, the syntactic annotations for training set are derived using the Stanford Parser (Klein and Manning, 2003). The grammar is extracted using a MLE technique as defined by equation 3. This system is used for parsing 1-best and N-best word output from an automatic speech recognizer where $N = 20$.

5.0.2 PPG Parsing System

The PPG is then created from word-based PCFG using the CMU³ pronunciation dictionary. Each production in PCFG that contains a word, a number of productions are added into PCFG corresponding to the number of pronunciations in the dictionary as shown in table 1. The PPG parsing system operates on the PCN without n-gram language modeling.

5.0.3 Joint Parsing Model System

The joint parsing model system uses the joint modeling of PPG and a 3-gram language model. The 3-gram language model is estimated from training set using (Stolcke, 2002) toolkit. This

system takes the PCN as input and produces the syntactic structure as an output.

5.0.4 Discussion

Table 2 shows the results for the experiment. The WER represents the measure of the error in recognition of word sequences while Precision, Recall and F-measure highlight the accuracy of syntactic structure produced by the parser.

It should be noted that the simple word-based parsing with PCFG degrades the WER as well as F-measure. This is due to the fact that some of the recognized sentences are not syntactically correct and therefore, cannot be parsed by the PCFG. It should also be noted that although parsing N-best list results in little bit improvement in WER, it degrades performance of syntactic parser as highlighted by the F-measure. The main reason for this is that parser chooses wrong sentence for parsing from N-best list when the best sentence is not possible to parse because of few recognition errors. On the other-hand, the PPG experiment with higher confusion threshold values also deviates from original WER (18.3%). As a result, it also degrades parsing performance. This shows that the PPG system alone is not able to handle extra confusion information effectively when the grammar is ambiguous. While, the performance of the system using joint probabilistic modeling of N-gram and PPG improves (both in terms of WER and Precision & Recall) with the increase

³<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

in confusion in the network. This is because both PPG and n-gram work together to effectively search the best hypothesis in the confusion network. Currently, the proposed system gives 2.3% absolute improvement in F-measure and 0.5% improvement in WER. Further improvements are expected if broader phonetic space is considered instead of using the PCM approach. Therefore, it would be desirable to apply the joint parsing model on actual output from phone recognizer or converting a word-lattice into PCN which is going to be the future direction of the work. Furthermore, the F-measure for the experiment is not as good as expected. This is because the Sparseval (Roark et al., 2006) is very strict on deletion and different ways of segmentation of speech utterance as shown in figure 3. In future, the plan is to investigate the tree-based alignment techniques (Demaine et al., 2009) for the evaluation of the syntactic accuracy (precision, recall and F-measure) of spoken language parsing.

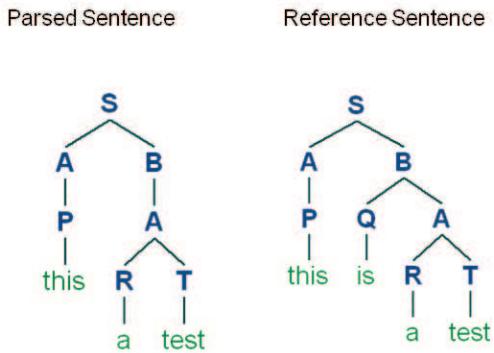


Figure 3: Using Sparseval, Recall = 25% and Precision = 25% even though only one word is missing in recognized sentence.

6 Conclusions

The paper presented a technique of parsing spoken language from phone sequences. Previous approaches to parsing spoken language work on word-level output from a recognizer and use 1-best outputs, N-best lists, lattices or confusion networks for parsing. The presented technique for spoken language parsing takes phone sequence as input and is based on joint probabilistic modeling of an n-gram statistical model and a PCFG model.

Since the 1-best phone output is not accurate, the speech is presented in the form of a phone confusion network (PCN). It has been shown that parsing PCN with PCFG and n-gram model results in better parsing than using simple word-based PCFG parsing approach. The strength of the joint parsing model lies in its ability to better recover from recognition error during parsing which is highlighted by the improvement in word error rate (WER) of recognized output.

Parsing is an important step in natural language processing. Various speech understanding and translation systems rely heavily on parsing accuracy. The proposed technique can be effectively used for these applications as it produces better accuracy than conventional spoken language parsing systems. The accuracy of the system can be further improved, if verified syntactic annotations are used for training data as opposed to using a text based parser to generate annotations. Considering broader phonetic search space than simply using phone confusion matrix approach, may also result in improvement in system accuracy. The future plan is to use the log-linear model and integrate a minimum error rate training approach to adjust the role of acoustic, n-gram, syntactic models on the target dataset.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

References

- James K. Baker. 1979. Trainable grammars for speech recognition. In *Proceedings of the Spring Conference of the Acoustical Society of America, Boston, MA*, pages 547–550.
- Nicola Bertoldi, Marcello Federico, Giuseppe Falavigna, and Matteo Gerosa. 2008. Fast speech decoding through phone confusion networks. In *INTERSPEECH*, pages 2094–2097, Brisban, Australia.
- J.-C. Chappelier and M. Rajman. 1998. A practical bottom-up algorithm for on-line parsing with

- stochastic context-free grammars. Technical report, Swiss Federal Institute of Technology.
- J.-C. Chappelier, M. Rajman, R. Arages, and A. Rozenknop. 1999. Lattice parsing for speech recognition. In *Proceedings of 6th Conference on Traitement Automata du Langage Naturel TALN*, pages 95–104.
- Eugene Charniak. 2001. Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL ’01, pages 124–131, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ciprian Chelba and Frederick Jelinek. 1998. Exploiting syntactic structure for language modeling. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL ’98, pages 225–231, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ciprian Chelba. 2000. *Exploiting Syntactic Structure for Natural Language Modeling*. Ph.D. thesis, Johns Hopkins University.
- Christopher Collins, Bob Carpenter, and Gerald Penn. 2004. Head-driven parsing for word lattices. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL ’04, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Erik D. Demaine, Shay Mozes, Benjamin Rossman, and Oren Weimann. 2009. An optimal decomposition algorithm for tree edit distance. *ACM Trans. Algorithms*, 6(1):2:1–2:19, dec.
- Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yoav Goldberg and Michael Elhadad. 2011. Joint hebrew segmentation and parsing using a pcfg-la lattice parser. In *Proceedings of the 49th Annual Meeting of the ACL*, Stroudsburg, PA, USA.
- Keith Hall and Mark Johnson. 2003. Language modeling using efficient best-first bottom-up parsing. In *In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.
- Keith Hall and Mark Johnson. 2004. Attention shifting for parsing speech. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Beth A. Hockey and Manny Rayner. 2005. Comparison of grammar based and statistical language models trained on the same data. In *Proceedings of the AAAI Workshop on SLU, Pittsburgh, PA*, July.
- Jie Jiang, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way. 2011. Phonetic representation-based speech translation. In *13th Machine Translation Summit*, Xiamen , China.
- Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Selçuk Köprü and Adnan Yazici. 2009. Lattice parsing to integrate speech recognition and rule-based machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 469–477, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Brian Roark, Mary Harper, Eugene Charniak, Bonnie Dorr, Mark Johnson, Jeremy G. Kahn, Yang Liu, Mari Ostendorf, John Hale, Anna Krasnyanskaya, Matthew Lease, Izhak Shafran, Matthew Snover, Robin Stewart, and Lisa Yung. 2006. Sparseval: Evaluation metrics for parsing speech. In *Language Resources and Evaluation (LREC)*, Genoa, Italy.
- Brian Edward Roark. 2001. *Robust probabilistic predictive syntactic processing: motivations, models, and applications*. Ph.D. thesis, Providence, RI, USA.
- Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *International Conference on Spoken Language Processing*, Denver, Colorado.
- Masaru Tomita. 1985. An efficient context-free parsing algorithm for natural languages. In *Proceedings of the 9th international joint conference on Artificial intelligence*, pages 756–764, San Francisco, USA.
- Jonathan Weese, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the thrax grammar extractor. In *6th Workshop on Statistical Machine Translation*, pages 478–484, Edinburgh, Scotland.

Linguistic analyses of the LAST MINUTE corpus

Dietmar Rösner, Manuela Kunze, Mirko Otto

Otto-von-Guericke Universität,
Institut für Wissens-
und Sprachverarbeitung
Fakultät für Informatik
Postfach 4120, D-39016 Magdeburg
roesner@ovgu.de

Jörg Frommer

Otto-von-Guericke-Universität,
Universitätsklinik für
Psychosomatische Medizin
und Psychotherapie
Leipziger Straße 44,D-39120 Magdeburg
joerg.frommer@med.ovgu.de

Abstract

The LAST MINUTE corpus comprises multimodal records from a Wizard of Oz (WoZ) experiment with naturalistic dialogs between users and a simulated companion system. We report about analysing the transcripts of the user companion dialogs and about insights gained so far from this ongoing empirical research.

1 Introduction

"Really natural language processing" (Cowie and Schröder, 2005), i.e. the possibility that human users speak to machines just as they would speak to another person, is a prerequisite for many future applications and devices. It is especially essential for so called companion systems (Wilks, 2010).

Corpora with naturalistic data from either human to human (e.g. (Oertel et al., 2012)) or human-machine interactions (e.g. (Legát et al., 2008), (Webb et al., 2010)) are an essential resource for research in this area. In the following we report about ongoing work in the linguistic analysis of the transcripts from the LAST MINUTE corpus. This corpus comprises multimodal records from a Wizard of Oz (WoZ) experiment with naturalistic dialogs between users and a simulated companion system.

The paper is organized as follows: In section 2 we give a short overview of the WoZ experiments. This is followed by a description of the LAST MINUTE corpus in section 3. In section 4 we report about analyses and empirical inves-

tigations with the transcripts. We sum up with a discussion of ongoing and future work.

2 The WoZ experiments

2.1 Design issues

Our WoZ-scenario is designed in such a way that many aspects of user companion interaction (UCI) that are relevant in mundane situations of planning, re-planning and strategy change (e.g. conflicting goals, time pressure, ...) will be experienced by the subjects (Rösner et al., 2011).

The overall structure of an experiment is divided into a personalisation module, followed by the 'LAST MINUTE' module. These modules serve quite different purposes and are further sub-structured in a different manner (for more details cf. (Rösner et al., 2012b)).

2.1.1 Personalisation module

Throughout the whole personalisation module the dominant mode of interaction is system initiative only, i.e. the system asks a question or gives a prompt. In other words this module is a series of dialog turns (or adjacency pairs (Jurafsky and Martin, 2008)) that are made up by a system question or prompt followed by the user's answer or reaction. In some sense this module thus resembles more an investigative questioning than a symmetric dialog.

2.1.2 The LAST MINUTE module

Selection In the bulk of 'LAST MINUTE' the subject is expected to pack a suitcase for a two week holiday trip by choosing items from an online catalogue with twelve different categories

that are presented in a fixed order. In a simplified view we thus have an iterative structure made up from twelve repetitions of structurally similar subdialogs each for the selection from a single category. The options of each category are given as menu on the subject's screen.

Normal packing sub dialog In a normal packing subdialog we essentially have a series of adjacency pairs made up of a user request for a number of items (more precisely: a user request for a number of instances from an item type) from the current selection menu (e.g. 'ten t-shirts') followed by a confirmation of the system (e.g. 'ten t-shirts have been added').

An example excerpt from an unproblematic segment of a packing dialog (subject 20110404bcm)¹:

```
{07:39} 058 P zwei tops [two tops]
{07:40} 059 (2.46)
{07:43} 060 W zwei tops wurden hinzugefügt (.)
[two tops have been selected]
sie können fortfahren [you can proceed]
{07:46} 061 (1.13)
{07:47} 062 P drei tshirts [three tshirts]
{07:49} 063 (2.42)
{07:51} 064 W drei tshirts wurden hinzugefügt
[three tshirts have been selected]
{07:53} 065 (3.63)
{07:57} 066 P ich möchte zur nächsten rubrik
[i want to go to the next category]
{07:59} 067 (2.73)
{08:01} 068 W sie können jetzt aus der rubrik jacken
[you may now choose from category]
und mäntel auswählen [jackets and coats]
{08:05} 069 (4.6)
{08:09} 070 P eine sommerjacke [a summer jacket]
```

Barriers The normal course of a sequence of repetitive subdialogs is modified for all subjects at specific time points.

These modifications or barriers are:

- after the sixth category, the current contents of the suitcase are listed verbally (listing barrier),
- during the eighth category, the system for the first time refuses to pack selected items because the airline's weight limit for the suitcase is reached (weight limit barrier).
- at the end of the tenth category, the system informs the user that now more detailed information about the target location Waiuku is available (Waiuku barrier).

¹All excerpts from transcript are given - unless otherwise noted - with the GAT 2 minimal coding (cf. below). English glosses added in brackets for convenience.

Additional barriers may occur depending on the course of the dialog. These are typically caused by user errors or limitations of the system or a combination of both.

2.2 Challenges for the subjects

In their initial briefing the subjects have been informed that all interaction shall be based on speech only and that neither keyboard or mouse are therefore available to them. Since the briefing does not comprise any detailed information about the natural language processing and the problem solving capabilities or limitations of the system the subjects are more or less forced to actively explore these aspects during the course of interaction.

The challenge for the subjects is twofold: They have to find out how (i.e. with which actions) they can solve problems that they encounter during interaction and they have to find out what linguistic means are available for them to instruct the system to perform the necessary actions. In other words, in order to be successful they have to build up a model of the capacities and limitations of the system based on their experience from successful or unsuccessful interactions. The user's model of the system will of course strongly influence the behavior of the user and the subsequent course of the interaction.

The discussion in (Edlund et al., 2008) leads to the following rephrasing of this challenge: Which metaphor will the subjects use when interacting with the WoZ simulated system? Will they treat the system more like a tool, i.e. choose the *interface metaphor*, or will they prefer the *human metaphor*, i.e. accept the system as an interlocutor and behave more like they would in human-human dialogs?

One approach to these questions is in the qualitative evaluation of the post-hoc in-depth interviews that a subset of ca. half of our subjects underwent after the experiments. In this paper we follow a complimentary approach: The linguistic behavior of the subjects is analysed under the perspective what conclusions it licenses about user assumptions about the speech-based system they experience.

Table 1: Comparison between corpora with naturalistic human-computer interactions

	SAL	SEMAINE	LAST MINUTE
Participants	4	20	130
Groups	students	students	balanced in age, gender, education
Duration	4:11:00	6:30:41	ca. 57:30:00
Nr of Sensors	2	9	13
Max. Video Bandwidth	352x288; 25Hz	580x780; 50Hz	1388x1038; 25Hz
Audio Bandwidth	20kHz	48kHz	44kHz
Transcripts	yes	yes	yes (GAT 2 minimal)
Biopsychological data	n.a.	n.a.	yes (heart beat, respiration, skin reductance)
Questionnaires	n.a.	n.a.	sociodemographic, psychometric
In depth Interviews	n.a.	n.a.	yes (70 subjects)
Language	English	English	German

3 Data sources

3.1 LAST MINUTE corpus

The LAST MINUTE corpus comprises multimodal recordings from the WoZ experiments with N = 130 participants (audio, video, biopsychological data), the verbatim transcripts and as additional material data from psychological questionnaires and records and transcripts from interviews (for more details cf. (Rösner et al., 2012a)).

In table 1 we summarize various parameters (as reported in (McKeown et al., 2010)) of two widely employed corpora with recordings from naturalistic human-computer dialogs – SAL (Douglas-Cowie et al., 2008) and SEMAINE (McKeown et al., 2010) – and contrast them with the resp. values for the LAST MINUTE corpus.

Sample The LAST MINUTE corpus consists of data sets from 130 subjects. 70 of them are between 18 and 28 years old ('young'; M=23,2; Md=23,0; s=2,9; 35 of them are male, 35 are female) and 60 of them are 60 years old or older ('elderly'; M=68,1; Md=67,0; s=4,8; the oldest subject is 81 years old; 29 of them are male, 31 are female). Within the group of the young 44 subjects have a high school diploma, 26 have none. Within the group of the elderly 35 subjects have a high school diploma or a university degree and 25 subjects have no university degree.

3.2 Wizard logs

The wizards have been trained and their behaviour has been anticipated and prescribed as

nonambiguous as possible in a manual (Frommer et al., 2012) .

All dialog contributions from the system (i.e. wizard) were pronounced by a text-to-speech system (TTS). The input for the TTS either was generated dynamically from the knowledge base (e.g. verbalisations of the current contents of the suitcase) or was chosen by the wizards from menus with prepared stock phrases. As a last option for unforeseen situations wizards could – supported by autocompletion – type in text to be uttered by the TTS. In the course of more than 130 experiments with on average approx. 90 dialog turns each (in sum a total of ca. 11800 turns) only in one single turn – during the very first experiments – the wizards had to resort to this last option.

After a WoZ session all wizard contributions together with their timings are available as additional log file.

Evaluation of the wizard log files already allows to classify the overall interaction of different subjects with respect to a number of aspects (cf. 4.2).

3.3 Transcripts

All experiments and interviews were transcribed by trained personnel following the GAT 2 minimal standard (Selting et al., 2009). This standard captures the spoken text, pauses, breathing and allows to include comments describing other nonlinguistic sounds.

In order to simplify the production of the GAT 2 transcripts, we started from the logged wizard statements which were converted into GAT2 tran-

scripts of the wizard. These transcripts were used as template for full transcripts of the interaction, into which only transcriptions of the utterances of the subject had to be inserted.

All transcripts were made using FOLKER (Schmidt and Schütte, 2010). Some transcripts were created by more than one transcriber, the parts are connected using exmeralda (Schmidt and Schütte, 2010). Own software was employed to support the transcribers in detecting and correcting possible misspellings.

The transcripts try to be as close as possible to the actual pronunciation of the subjects. They therefore include as well nonstandard writings, e.g. for dialect (e.g. 'jejebn' instead of 'gegeben', engl. 'given'). In addition the utterances of the subjects exhibit phenomena that are typical for spontaneous spoken language, e.g. repairs, restarts, incongruencies.

3.3.1 Corpus size

Counting only the user contributions the total number of tokens in the corpus of transcripts sums up to 79611. The number of tokens per transcript ranges from 252 till 1730 with mean value 612,39 (variance 186,34). The total size may seem small when compared to the size of large available corpora, but one should keep in mind that in spite of their differences all 130 dialogs are focussed and thus allow for in depth comparisons and analyses.

3.3.2 Processing of transcripts

For linguistic processing of the Folker based transcripts we employ the UIMA framework.²

The first step is to transform Folker format into UIMA based annotations. After this, we initiate a number of linguistic and dialogue based analyses. For these analyses, we used internal and external tools and resources. For example, we integrated resources of GermaNet³, LIWC (Wolf et al., 2008) and of the project Wortschatz Leipzig⁴.

4 Linguistic analyses

Linguistic analyses of the LAST MINUTE transcripts are an essential prerequisite for an in depth investigation of the dialog and problem solving

behavior of the subjects in the WoZ experiments. A long term goal is to correlate findings from these analyses with sociodemographic and psychometric data from the questionnaires.

4.1 Linguistic structures employed by subjects?

4.1.1 Motivation

First inspections of transcripts revealed that there are many variations in lexicalisation but only a small number of linguistic constructs that subjects employed *during the packing and unpacking subdialogues* of the LAST MINUTE experiments.

For issuing packing (or unpacking) commands these structural options comprise:

- full sentences with a variety of verbs or verb phrases and variations in constituent ordering,
- elliptical structures without verbs in a (properly inflected) form like
`<number> <item(s)>`,
- 'telegrammatic structures' in a (technically sounding and mostly uninflected) form with an inverted order of head and modifier like
`<item> <number>`.

As a first quantitative analysis of the 'LAST MINUTE' phase, the absolute and relative numbers for the usage of these constructs have been calculated from the full set of transcripts.

4.1.2 Results

Based on the analyses of the packing/unpacking phase of $N = 130$ transcripts we get the following figures:

We have a total of 8622 user utterances. If we perform POS tagging (with a slightly modified version of STTS⁵) and then count the varying POS tag patterns, we find 2041 different patterns for the 8622 utterances. The distribution is strongly skewed (cf. table 2): A small number of (regular) POS patterns captures a large fraction of utterances.

In classifying POS tag sequences we distinguish four categories: full sentences and sentence

²uima.apache.org/

³http://www.sfs.uni-tuebingen.de/lst/

⁴wortschatz.uni-leipzig.de/

⁵www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Table 2: Most frequent POS patterns for elliptical structures

class	sem	POS pattern	nr of occs
E	P	ART NN	1020
E	P	CARD NN	657
E	P	NN	537
E	C	ADJ NN	355
E	C	ADJ,ADV	349
E	C	NN	148

like structures (S; with an obligatory verb), elliptical constructs without a verb (E), telegrammatic constructs (T, cf. above) and meaningful pauses, i.e. user utterances, that more or less consist of interjections only (DP).

In descending order of occurrences we have the following counts:

- 5069 user utterances or 58.79 % (realised with 223 patterns) are classified as E,
- 807 user utterances or 9.36 % (realised with 135 patterns) as S,
- 551 user utterances or 6.39 % (realised with 21 patterns) as T, and finally
- 178 user utterances or 2.06 % (realised with 8 patterns) as DP.

At the time of writing 2017 utterances realised in 1654 different patterns can not uniquely be classified. In many cases this is due to the typical phenomena of spontaneous spoken language, e.g. repairs, restarts and the use of interjections.

4.1.3 Discussion and remarks

The use of elliptical structures is a typical aspect of efficient communication in naturally occurring dialogs (Jurafsky and Martin, 2008). Thus the dominance of elliptical structures in the user contributions of the LAST MINUTE corpus can be seen as a clear indicator that most subjects have experienced the dialog with the system in a way that licensed their natural dialog behavior.

The empirical analysis of the structure of user utterances has fed as well into the implementation of an experimental system that allows to replace the wizard with an automated system based on the commercial speech recogniser Nuance.

4.2 Success and failure of dialog turns?

4.2.1 Motivation

Dialog turns are either successful or they may fail for a variety of reasons. We will first discuss failures during the LAST MINUTE phase. Failed turns can easily be detected in the so called wizard logs (cf. 3.2) because in the case of failure no confirmation is given by the system but some different utterance.

Success A dialog turn starting with a user request to pack or to unpack some items is successful when the situation allows to perform the requested action. In the wizard log file this can easily be detected by the respective confirmative response of the system ('... wurd.* hinzugef.* ...', '... wurd.* entfernt ...').

Error messages The least specific 'error message' of the system tells the user that his utterance can not be processed ('ihre aussage kann nicht verarbeitet werden'). There are a number of reasons for using this 'catch all' system response. These include:

- The wizards conjure that the voice quality of the user's utterance is too poor for current automated speech recognition (ASR) technology.
- The content of the user's utterance is beyond the allowed scope of the current subdialog.
- The syntactic or semantic complexity of the user's utterance is judged to be beyond the limits of current NLP technology.

The following system reactions are more specific:

- When a user tries to unpack items that have not been packed into the suitcase he gets the response that these items are not contained in the suitcase ('... nicht im Koffer enthalten').
- When a user reaches the weight limit for the suitcase again then a packing command is responded to by the system with the message that the chosen item(s) can not be packed due to the weight limit ('... k.*nn.* nicht hinzugef.* werden ...').

- When the time for a category is over (local time limit) then the system tells this to the user and enforces a change of the category ('... muss jetzt beendet werden ...').

The excerpt in table 3 illustrates a problematic dialog situation: the simulated system does not accept a collective term ('tauchausrustung', engl. 'diving equipment') employed by subject 20110224awh in an unpacking request.

A global measure In order to compare different dialogs we start with the following coarse global measure: We distinguish turns that are - based on the logged system response – judged as successful from those that are judged as unsuccessful or faulty. We then use the ratio of unsuccessful turns in relation to all turns as measure of the relative faultiness of the dialog as a whole.

Figure 1 visualises different dialog courses of subjects from the experiments (green: successful turn, red: unsuccessful turn).

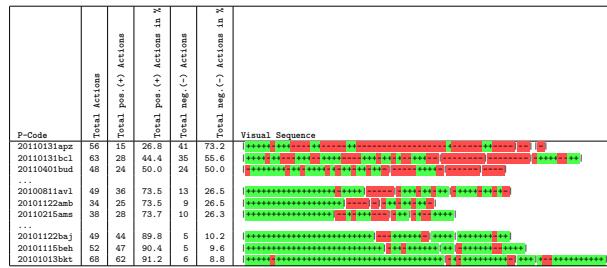


Figure 1: Variations in dialog courses

4.2.2 Questions

Are there correlations of dialog success with sociodemographic variables and with personality traits measured with the psychometric questionnaires?

4.2.3 Results

For a cohort of $N = 130$ subjects the values for this global measure range between 9 % and 73 % with a mean of approximately 26 % and variance 10.

In fig. 2 the result of a contrastive analysis of the dialog courses of all $N = 130$ subjects, divided into the subcohorts of elderly vs. young subjects, is given. The chart illustrates that more than half

of the elderly have significantly more negative dialog turns than the young subjects.

The differences are significant: a t-test yields a t-value of -3.779595 and a p-value of 0.000240.

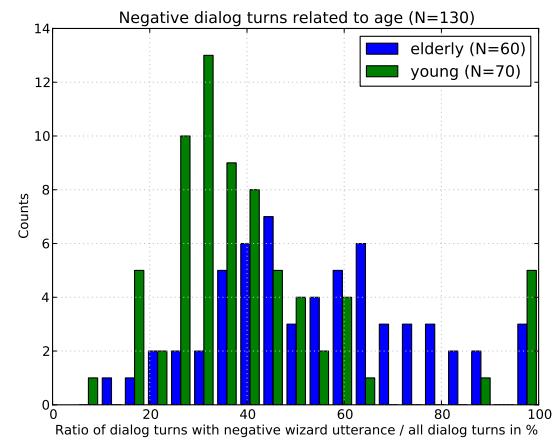


Figure 2: Contrastive evaluation of dialog courses: elderly vs. young subjects

4.3 Do users adapt to the system?

4.3.1 Motivation

The subjects in our WoZ experiments are only informed that they will interact in spoken language with a new type of system that shall be personalised and that therefore will ask some questions and pose some tasks. They do not get explicit instructions about the linguistic constructions and interaction patterns that are possible or not possible to use in the interaction.

How do people with differing technical background interact in such a situation?

One issue to explore is the question of rapport: Do users mirror the language of the system, e.g. on the lexical or syntax level? The system e.g. uses the general and somewhat uncolloquial term 'hinzufügen' (engl. to add) in its feedback for selection operations. Similarly the system always uses 'wurde entfernt' (engl. approx. 'was removed') when confirming unpacking requests. Do users mirror this usage?

4.3.2 Results

Within the total of $N = 130$ transcripts only in $N_1 = 25$ transcripts at least one occurrence of a form of the verb 'hinzufügen' (engl. 'to add')

Table 3: A problematic dialog situation

```
(23:05) 208 P also (1.5) zahnpflegeset [dental care kit]
(23:09) 209 (4.0)
(23:13) 210 W ein zahnpflegeset kann nicht hinzugefügt werden (.) [a dental care kit cannot be added]
andernfalls würde die von der fluggesellschaft vorgeschriebene [otherwise the weight limit prescribed by the airline]
maximale gewichtsgrenze des koffers überschritten werden (.) bevor [will be exceeded (.) before selecting]
weitere artikel ausgewählt werden können (.) müssen sie für genügend [any further items (.) you will have to make room for them]
platz im koffer sorgen (.) hierfür können bereits eingepackte artikel [in your suitcase (.) therefore already selected items]
wieder ausgepackt werden (.) auf nachfrage erhalten sie eine [can be unpacked (.) you can get a list of the selected items]
aufzählung der bereits ausgewählten artikel [upon request]
(23:37) 211 (3.4)
(23:41) 212 P ((fingertrommeln)) was nehm_wer_n da wieder raus (.) [((finger drumming)) well what shall we take out again]
dann nehm_wer_n die tauchausrüstung raus [let us take out the diving equipment]
(23:48) 213 (2.51)
(23:51) 214 W der gewünschte artikel ist nicht im koffer enthalten [the selected item is not contained in the suitcase]
enthalten (1.9) ((fingertrommeln)) [contained (1.9) ((finger drumming))]
(24:02) 217 W ihre aussage kann nicht verarbeitet werden [your statement cannot be processed]
(24:05) 218 (2.69)
(24:08) 219 P so (.) °h °h ° (1.7) ein rasierset (-) brauche ich [so (.) i do need a beard trimmer]
(24:14) 220 (3.5) {24:18} 221 W der artikel rasierset kann nicht [item beard trimmer cannot be added]
hinzugefügt werden (.) andernfalls würde die maximale gewichtsgrenze [otherwise the weight limit]
des koffers überschritten werden (--) ... [of the suitcase will be exceeded (--) ...]
```

could be found in user utterances of the packing/unpacking phase. For these $N_1 = 25$ transcripts we have a range from 1 to maximally 8 occurrences with mean: 2.36, std: 1.85 and median: 2.0.

Within $N_2 = 68$ transcripts at least one occurrence of a form of the verb 'entfernen' (engl. 'to remove') could be found in user utterances of the packing/unpacking phase. For these $N_2 = 68$ transcripts we have a range from 1 to maximally 13 occurrences with mean: 4.22, std: 3.34 and median: 3.0.

In the intersection of both groups, i.e. at least one occurrence each of a form of the verbs 'entfernen' and 'hinzufügen', we have $N_3 = 20$ transcripts. For these $N_3 = 20$ transcripts we have a range from 2 to maximally 19 combined occurrences with mean: 8.85, std: 4.64 and median: 8.0.

4.3.3 Discussion and remarks

That users mirror the lexical items of the system is thus rather the exception than the rule. Nevertheless it seems worth to be explored if - and if so how - the subgroup of subjects that do so differs from those subjects that do not.

4.4 Politeness in user utterances?

4.4.1 Motivation

In all its utterances the system uses the polite version, the German 'Sie' (polite, formal German version of 'you') when addressing the user. In requests the system employs the politeness particle

'bitte' (engl. 'please'). How polite are users in their utterances?

4.4.2 Results

Pronouns The following counts are all (unless otherwise noted) taken from the packing/unpacking phase of the transcripts: Within a total of $N = 130$ transcripts only in $N_1 = 21$ transcripts at least one occurrence of 'sie' als formal personal pronoun in addressing the system is used. Only within $N_2 = 4$ transcripts the informal 'du' (or one of its inflected forms) is used to address the system (other uses of 'du' are within idiomatic versions of swear words like 'ach du lieber gott', engl. 'oh god'). Within $N_3 = 18$ transcripts subjects employ the plural personal pronoun 'wir' (engl. 'we'). Some occurrences of 'wir' in offtalk can be seen as more or less fixed phrasal usages (like 20101115beh 'ach das schaffen wir locker', engl. '... we will make this with ease' or 20110401adh 'wo waren wir', engl. 'where have we been'), but when used in commands (packing, unpacking, ...) then this pronoun can be given an inclusive collective reading as referring to both subject and system as a joint group. Please note: The pronoun 'wir' thus allows users to avoid to explicitly approach the system.

Example of this latter usage:

```
20110307bss:ja dann nehm wir eine jacke raus
[engl.: yeah then we take a jacket off]
20110315agw: dann streichen wir ein hemd
[engl.: then we cancel a shirt]
```

In sum: How users approach the system differs significantly. Most subjects avoid any personal pronouns when addressing the system, some em-

ploy the German 'Sie' (formal German version of 'you') and only very seldom the informal German 'du' is used.

Politeness particles From $N = 130$ subjects $N_1 = 67$ use one of the politeness particles 'bitte' or 'danke' at least once within the packing/unpacking phase. The maximum number of uses is 34, with a mean of 7.57, standard deviation of 7.89 and median of 4.0. If we neglect those subjects at and below the median as only occasional users of these particles we get $N_2 = 32$ subjects that use these particles much more frequent.

Intersecting the group of subjects with at least one occurrence of 'sie' (cf. above) with users of the politeness particles 'bitte' or 'danke' results in a subgroup of $N_3 = 19$ subjects. These have combined numbers of occurrences ranging from 2 till 32 with mean: 8,60, std: 8,38 and median: 4,00. In other words: most user of 'sie' are as well users of the politeness particles.

4.4.3 Discussion and remarks

Politeness is one of a number of indicators of the way how subjects experience the system.

The difference in using personal pronouns and politeness particles is another example that most users do not try to build rapport with the system on the level of lexical choices.

As with other subgroups of subjects (as e.g. detected in 4.3) the following questions have to be further investigated:

Are there differences in the overall dialog success or failure between 'normal' and 'polite' users?

Are there correlations between user politeness and sociodemographic data and personality traits measured with the psychometric questionnaires?

4.5 Conclusion

In the light of the metaphor discussion (cf. 2.2) we can summarize and re-interpret our results as follows: Subjects whose linguistic behavior gives a strong indication for the dominance of one of these metaphors are minorities within our sample. This holds for the minority group of those that prefer technically sounding 'grammatical structures' (cf. 4.1) and thus obviously prefer the interface metaphor. It holds as well - on

the other extreme - for the group of those that heavily employ interpersonal signals such as formal pronouns and politeness particles thus indicating a human metaphor at work. Although further investigations are necessary, the majority of our subjects seems to work with 'a metaphor that lies between a human and machine - the android metaphor' (Edlund et al., 2008).

5 Future work

We report here about on going work. More issues have been or are still investigated with the LAST MINUTE corpus that can - due to limited space - only be mentioned here. Issues to be further explored include:

Effects of reinforcement learning We have found indications that subjects strongly tend to reuse linguistic constructs that have resulted in successful dialog turns (an effect that can be interpreted as a form of reinforcement learning).

Verbosity vs. sparseness of linguistic expression As already noted above, subjects strongly differ in their verbosity. This is of course more obvious in the narratives of the personalisation phase, but it is measurable even in the LAST MINUTE phase.

Detection and analysis of offtalk Linguistic analysis is essential for the detection of offtalk. Many questions arise: How often does offtalk occur? How can offtalk utterances be further classified (e.g. thinking aloud, expressing emotions, ...)? Is there a correlation between the degree and nature of offtalk usage and sociodemographic data and personality traits?

Emotional contents In the experiments reported here we have three sources of utterances with emotional contents: self reporting about past emotions in the personalisation phase for all subjects, self reporting about current emotions in the intervention phase for the randomly chosen subjects with an intervention and spontaneous expression of emotions (e.g. swear words, offtalk, self accusations, etc.) especially at the barriers or when problems occur during the interaction.

A detailed linguistic analysis of these various forms of emotional contents in the LAST MINUTE transcripts is on the agenda.

6 Summary

We have presented the current state of the linguistic analyses of the LAST MINUTE corpus. This corpus of recordings from naturalistic interactions between humans and a WoZ simulated companion system excels available corpora with respect to cohort size, volume and quality of data and comes with accompanying data from psychometric questionnaires and from post hoc in depth interviews with participants. The material is a cornerstone for work in the SFB TRR 62 but is as well available for research in affective computing in general.

Acknowledgments

The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors.

Availability

The LAST MINUTE corpus is available for research purposes upon written request from the authors.

References

- Roddy Cowie and Marc Schröder. 2005. Piecing together the emotion jigsaw. *Machine Learning for Multimodal Interaction*, pages 305–317.
- E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen. 2008. The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. In L. Devillers, J-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France. ELRA.
- Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630 – 645.
- Jörg Frommer, Dietmar Rösner, Matthias Haase, Julia Lange, Rafael Friesen, and Mirko Otto. 2012. *Verhinderung negativer Dialogverläufe – Operator-manual für das Wizard of Oz-Experiment*. Pabst Science Publishers.
- Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2nd (May 26, 2008) edition.
- M. Legát, M. Grüber, and P. Ircing. 2008. Wizard of oz data collection for the czech senior companion dialogue system. In *Fourth International Workshop on Human-Computer Conversation*, pages 1 – 4, University of Sheffield.
- G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084, July.
- Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2012. D64: a corpus of richly recorded conversational interaction. *Journal of Multimodal User Interfaces*, in press.
- Dietmar Rösner, Rafael Friesen, Mirko Otto, Julia Lange, Matthias Haase, and Jörg Frommer. 2011. Intentionality in interacting with companion systems – an empirical approach. In Julie Jacko, editor, *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, volume 6763 of *Lecture Notes in Computer Science*, pages 593–602. Springer Berlin / Heidelberg.
- Dietmar Rösner, Jörg Frommer, Rico Andrich, Rafael Friesen, Matthias Haase, Manuela Kunze, Julia Lange, and Mirko Otto. 2012a. LAST MINUTE: a novel corpus to support emotion, sentiment and social signal processing. In *LREC 2012 Workshop Abstracts*, pages 171–171.
- Dietmar Rösner, Jörg Frommer, Rafael Friesen, Matthias Haase, Julia Lange, and Mirko Otto. 2012b. LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In *LREC 2012 Conference Abstracts*, page 96.
- Thomas Schmidt and Wilfried Schütte. 2010. Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapia, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte,

- Anja Stukenbrock, and Susanne Uhmann, 2009. *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)*. Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion, 10 edition.
- Nick Webb, David Benyon, Jay Bradley, Preben Hansen, and Oli Mival. 2010. Wizard of oz experiments for a companion dialogue system: Eliciting companionable conversation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).
- Y. Wilks. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological and Design issues*. John Benjamins, Amsterdam.
- Markus Wolf, Andrea B. Horn, Matthias R. Mehl, Sev- erin Haug, James W. Pennebraker, and Hans Kordy. 2008. Computergestützte quantitative Textanalyse. In *Diagnostica*, volume Vol. 54, Number 2/2008. Hogrefe, Göttingen.

S-restricted monotone alignments

Steffen Eger

Faculty of Economics

Goethe Universität Frankfurt

Grüneburg-Platz 1, 60323 Frankfurt am Main, Germany

steffen.eger@yahoo.com

Abstract

We present an alignment algorithm for monotone many-to-many alignments, which are relevant e.g. in the field of grapheme-to-phoneme conversion (G2P). Moreover, we specify the size of the search space for monotone many-to-many alignments in G2P, which indicates that exhaustive enumeration is generally possible, so that some limitations of our approach can easily be overcome. Finally, we present a decoding scheme, within the monotone many-to-many alignment paradigm, that relates the decoding problem to restricted integer compositions and that is, putatively, superior to alternatives suggested in the literature.

1 Introduction

Grapheme-to-phoneme conversion (G2P) is the problem of transducing, or converting, a grapheme, or letter, string \mathbf{x} over an alphabet Σ_x into a phoneme string \mathbf{y} over an alphabet Σ_y . An important first step thereby is finding *alignments* between grapheme and phoneme strings in training data. The classical alignment paradigm has presupposed alignments that were

- (i) *one-to-one* or *one-to-zero*; i.e. one grapheme character is mapped to at most one phoneme character; this assumption has probably been a relic of both the traditional assumptions in machine translation (Brown et al. 1990) and in biological sequence alignment (Needleman and Wunsch, 1970). In the field of G2P such alignment models are sometimes

also called ϵ -scattering models (Black et al., 1998).

- (ii) *monotone*, that is, the order between characters in grapheme and phoneme strings is preserved.

It is clear that, despite its benefits, the classical alignment paradigm has a couple of limitations; in particular, it may be unable to explain certain grapheme-phoneme sequence pairs, a.o. those where the length of the phoneme string is greater than the length of the grapheme string such as in

exact igzækt

where \mathbf{x} has length 5 and \mathbf{y} has length 6. In the same context, even if an input pair can be explained, the one-to-one or one-to-zero assumption may lead to alignments that, linguistically, seem nonsensical, such as

p	h	o	e	n	i	x
f	—	i:	n	i	k	s

where the reader may verify that, no matter where the ϵ is inserted, some associations will always appear unmotivated. Moreover, monotonicity appears in some cases violated as well, such as in the following,

centre sentər

where it seems, linguistically, that the letter character r corresponds to phonemic r and graphemic word final e corresponds to \emptyset .

Fortunately, better alignment models have been suggested to overcome these problems. For example, Jiampojamarn et al. (2007) and Jiampojamarn and Kondrak (2010) suggest ‘many-to-many’ alignment models that address issue (i) above. Similar ideas were already present in (Baldwin and Tanaka, 2000), (Galescu and Allen, 2001) and (Taylor, 2005). Bisani and Ney (2008) likewise propose many-to-many alignment models; more precisely, their idea is to *segment* grapheme-phoneme pairs into non-overlapping parts (‘co-segmentation’), calling each segment a *graphone*, as in the following example, consisting of five graphones,

ph	oe	n	i	x
f	i:	n	i	ks

The purpose of the present paper is to introduce a very simple, flexible and general monotone many-to-many alignment algorithm (in Section 3) that competes with the approach suggested in Jiampojamarn et al. (2007). Thereby, our algorithm is an intuitive and straightforward generalization of the classical Needleman-Wunsch algorithm for (biological or linguistic) sequence alignment. Moreover, we explore simple and valuable extensions of the presented framework, likewise in Section 3, which may be useful e.g. to detect latent classes in alignments, similar to what has been done in e.g. Dreyer et al. (2008). We also mention limitations of our procedure, in Section 4, and discuss the naive brute-force approach, exhaustive enumeration, as an alternative; furthermore, by specifying the search space for monotone many-to-many alignments, we indicate that exhaustive enumeration appears generally a feasible option in G2P and related fields. Then, a second contribution of this work is to suggest an alternative decoding procedure when transducing strings \mathbf{x} into strings \mathbf{y} , within the monotone many-to-many alignment paradigm (in Section 6.2). We thereby relate the decoding problem to restricted integer compositions, a field in mathematical combinatorics that has received increased attention in the last few years (cf. (Heubach and Mansour, 2004; Malandro, 2012)). Finally, we demonstrate the superiority of our approach by applying it to several data sets in Section 7.

It must be mentioned, generally, that we take G2P only as an (important) sample application of monotone many-to-many alignments, but that they clearly apply to other fields of natural language processing as well, such as transliteration, morphology/lemmatization, etc. and we thus also incorporate experiments on morphology data. Moreover, as indicated, we do not question the premise of monotonicity in the current work, but take it as a crucial assumption of our approach, leading to efficient algorithms. Still, ‘local non-monotonicities’ as exemplified above can certainly be adequately addressed within our framework, as should become clear from our illustrations below (e.g. with higher-order ‘steps’).

2 S-restricted paths and alignments

Consider the two-dimensional lattice \mathbb{Z}^2 . In \mathbb{Z}^2 , we call an ordered list of pairs $(\alpha_0, \beta_0) = (0, 0), \dots, (\alpha_k, \beta_k) = (m, n)$ a *path* from $(0, 0)$ to (m, n) , and we call $(a_i, b_i) := (\alpha_i, \beta_i) - (\alpha_{i-1}, \beta_{i-1})$, $i = 1, \dots, k$, *steps*. Moreover, we call a path λ in the lattice \mathbb{Z}^2 from $(0, 0)$ to (m, n) *monotone* if all steps (a, b) are non-negative, i.e. $a \geq 0, b \geq 0$, and we call the monotone path λ *S-restricted* for a subset S of \mathbb{N}^2 if all steps lie within S , i.e. $(a, b) \in S$.

Note that *S*-restricted monotone paths define (restricted) co-segmentations, or (a special class of) monotone alignments, between strings \mathbf{x} and \mathbf{y} . For example, the two paths in Figure 1 correspond to the two monotone alignments between $\mathbf{x} = \text{phoenix}$ and $\mathbf{y} = \text{fi:niks}$ illustrated above. Thus, we identify *S*-restricted monotone paths with *S*-restricted monotone alignments in the sequel.

Moreover, note that the set and number of *S*-restricted monotone paths allow simple recursions. To illustrate, the number $T_S(m, n)$ of *S*-restricted monotone paths from $(0, 0)$ to (m, n) satisfies

$$T_S(m, n) = \sum_{(a,b) \in S} T_S(m-a, n-b), \quad (1)$$

with initial condition $T_S(0, 0) = 1$ and $T_S(m, n) = 0$ if $m < 0$ or $n < 0$. As will be seen in the next section, under certain assumptions, *optimal* monotone alignments (or, equiva-

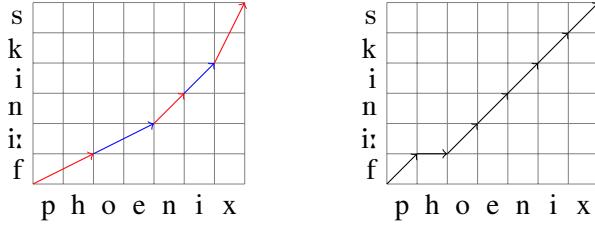


Figure 1: Monotone paths in two-dimensional lattices corresponding to the monotone alignments between $\mathbf{x} = \text{phoenix}$ and $\mathbf{y} = \text{fi:niks}$ given in Section 1. In the left lattice, we have arbitrarily (but suggestively) colored each step in either red or blue.

lently, paths) can be found via a very similar recursion.

3 Algorithm for S -restricted alignments

Let two strings $\mathbf{x} \in \Sigma_x^*$ and $\mathbf{y} \in \Sigma_y^*$ be given. Moreover, assume that a set S of allowable steps is specified together with a real-valued similarity function $\text{sim} : \Sigma_x^* \times \Sigma_y^* \rightarrow \mathbb{R}$ between characters of Σ_x and Σ_y . Finally, assume that the *score* or value of an S -restricted monotone path $\lambda = (\alpha_0, \beta_0), \dots, (\alpha_k, \beta_k)$ is defined additively linear in the similarity of the substrings of \mathbf{x} and \mathbf{y} corresponding to the steps (a, b) taken, i.e.

$$\text{score}(\lambda) = \sum_{i=1}^k \text{sim}(x_{\alpha_{i-1}+1}^{\alpha_i}, y_{\beta_{i-1}+1}^{\beta_i}), \quad (2)$$

where by $x_{\alpha_{i-1}+1}^{\alpha_i}$ we denote the subsequence $x_{\alpha_{i-1}+1} \dots x_{\alpha_i}$ of \mathbf{x} and analogously for \mathbf{y} . Then it is not difficult to see that the problem of finding the path (alignment) with maximal score can be solved efficiently using a very similar (dynamic programming) recursion as in Eq. (1), which we outline in Algorithm 1. Moreover, this algorithm is obviously a straightforward generalization of the classical Needleman-Wunsch algorithm, which specifies S as $\{(0, 1), (1, 0), (1, 1)\}$.

Note, too, that in Algorithm 1 we include two additional quantities, not present in the original sequence alignment approach, namely, firstly, the ‘quality’ q of a step (a, b) , weighted by a factor $\gamma \in \mathbb{R}$. This quantity may be of practical importance in many situations. For example, if we specify sim as log-probability (see below), then Algorithm 1 has a ‘built-in’ tendency to substitute ‘smaller’, individually more likely steps (a, b) by larger, less likely steps because in the

Algorithm 1 Gen. Needleman-Wunsch (GNW)

```

1: procedure GNW( $x_1 \dots x_m, y_1 \dots y_n; S,$ 
  sim, q, L)
2:    $M_{ij} \leftarrow 0$  for all  $(i, j) \in \mathbb{Z}^2$  such that
    $i < 0$  or  $j < 0$ 
3:    $M_{00} \leftarrow 1$ 
4:   for  $i = 0 \dots m$  do
5:     for  $j = 0 \dots n$  do
6:       if  $(i, j) \neq (0, 0)$  then
7:          $M_{ij} \leftarrow \max_{(a,b)^c \in S} \{M_{i-a,j-b} +$ 
           $\text{sim}(x_{i-a+1}^i, y_{j-b+1}^j) + \gamma q(a, b) +$ 
           $\chi L((x_{i-a+1}^i, y_{j-b+1}^j), c)\}$ 
8:       end if
9:     end for
10:   end for
11:   return  $M_{mn}$ 
12: end procedure

```

latter case fewer negative numbers are added; if sim assigns strictly positive values, this relationship is reversed. We can counteract these biases by factoring in the *per se* quality of a given step. Also note that if q is added linearly, as we have specified, then the dynamic programming recursion is not violated.

Secondly, we specify a function $L : (\Sigma_x^* \times \Sigma_y^*) \times \text{colors} \rightarrow \mathbb{R}$, where colors is a finite set of ‘colors’, that encodes the following idea. Assume that each step $(a, b) \in S$ appears in C , $C \in \mathbb{N}$, different ‘colors’, or states. Then, when taking step (a, b) with color $c \in \text{colors}$ (which we denote by the symbol $(a, b)^c$ in Algorithm 1), we assess the ‘goodness’ of this decision by the ‘likelihood’ L that the current subsequences of \mathbf{x} and \mathbf{y} selected by the step (a, b) ‘belong to’/‘are of’ color (or state) c . As will be seen below, this al-

lows to very conveniently identify (or postulate) ‘latent classes’ for character subsequences, while increasing the algorithm’s running time only by a constant factor.

As to the similarity measure sim employed in Algorithm 1, a popular choice is to specify it as the (logarithm of the) joint probability of the pair $(u, v) \in \Sigma_x^* \times \Sigma_y^*$, but a multitude of alternatives is conceivable here such as the χ^2 similarity, pointwise mutual information, etc. (see for instance the overview in Hoang et al. (2009)). Also note that if sim is e.g. defined as joint probability $\Pr(u, v)$ of the string pair (u, v) , then $\Pr(u, v)$ is usually initially unknown but can be iteratively estimated via application of Algorithm 1 and count estimates in an EM-like fashion (Dempster et al., 1977), see Algorithm 2. As concerns q and L , we can likewise estimate them iteratively from data, specifying their abstract forms via any well-defined (goodness) measures. The associated coefficients γ and χ can be optimized on a development set or set exogenously.

Algorithm 2 (Hard) EM Training

```

1: procedure EM( $\{(x_i, y_i) | i = 1, \dots, N\}; S,$ 
    $T, \hat{\text{sim}}_0, \hat{q}_0, \hat{L}_0$ )
2:    $t \leftarrow 0$ 
3:   while  $t < T$  do
4:     for  $i = 1 \dots N$  do
5:        $(x_i^a, y_i^a) \leftarrow$ 
        GNW( $x_i, y_i; S, \hat{\text{sim}}_t, \hat{q}_t, \hat{L}_t$ )
6:     end for
7:      $\hat{\text{sim}}_{t+1}, \hat{q}_{t+1}, \hat{L}_{t+1} \leftarrow f(\{x_i^a, y_i^a | i =$ 
       $1, \dots, N\})$ 
8:      $t \leftarrow t + 1$ 
9:   end while
10: end procedure

```

4 Exhaustive enumeration and alignments

In the last section, we have specified a polynomial time algorithm for solving the monotonic S -restricted string alignment problem, under the following restriction; namely, we defined the score of an alignment additively linear in the similarities of the involved subsequences. This, however, entails an independence assumption between successive aligned substrings that oftentimes does

not seem justified in linguistic applications. If, on the contrary, we specified the score, $\text{score}(\lambda)$, of an alignment λ between strings x and y as e.g.

$$\sum_{i=1}^k \log \Pr((x_{\alpha_{i-1}+1}^{\alpha_i}, y_{\beta_{i-1}+1}^{\beta_i}) | (x_{\alpha_{i-2}+1}^{\alpha_{i-1}}, y_{\beta_{i-2}+1}^{\beta_{i-1}}))$$

(using joint probability as similarity measure) — this would correspond to a ‘bigram scoring model’ — then Algorithm 1 would not apply.

To address this issue, we suggest *exhaustive enumeration* as a possibly noteworthy alternative — enumerate *all* S -restricted monotone alignments between strings x and y , score each of them individually, taking the one with maximal score. This brute-force approach is, despite its simplicity, the most general approach conceivable and works under all specifications of scoring functions. Its practical applicability relies on the sizes of the search spaces for S -restricted monotone alignments and on the lengths of the strings x and y involved.

We note the following here. By Eq. 1, for the choice $S = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1)\}$, a seemingly reasonable specification in the context of G2P (see below), the number $T_S(n, n)$ of S -restricted monotone alignments is given as (for explicit formulae, cf. (Eger, 2012))

$$1, 1, 3, 7, 16, 39, 95, 233, 572, 1406, 3479, 8647$$

for $n = 1, 2, \dots, 12$ and e.g. $T_S(15, 15) = 134,913$. Moreover, for the distribution of letter string and phoneme string lengths we estimate Poisson distributions (Wimmer et al., 1994) with parameters $\mu \in \mathbb{R}$ as listed in Table 1 for the German Celex (Baayen et al., 1996), French Brulex (Content et al., 1990) and English Celex datasets, as used in Section 7. As the table and the above numbers show, there are on average only a few hundred or few thousand possible monotone many-to-many alignments between grapheme and phoneme string pairs, for which exhaustive enumeration appears, thus, quite feasible; moreover, given enough data, it usually does not harm much to exclude a few string pairs, for which alignment numbers are too large.

5 Choice of S

Choice of the set of steps S is a question of *model selection*, cf. (Zucchini, 2000). Several ap-

Dataset	μ_G	μ_P	$P_{[G>15]}$	$P_{[P>15]}$
German-Celex	9.98	8.67	4.80%	1.62%
French-Brulex	8.49	6.71	1.36%	0.15%
English-Celex	8.21	7.39	1.03%	0.40%

Table 1: Avg. grapheme and phoneme string lengths in resp. data set, and probabilities that lengths exceed 15.

proaches are conceivable here. First, for a given domain of application one might specify a possibly ‘large’ set of steps Ω capturing a preferably comprehensive class of alignment phenomena in the domain. This may not be the best option because it may provide Algorithm 1 with too many ‘degrees of freedom’, allowing it to settle in unfavorable local optima. A better, but potentially very costly, alternative is to exhaustively enumerate all possible subsets S of Ω , apply Algorithm 1 and/or Algorithm 2, and evaluate the quality of the resulting alignments with any choice of suitable measures such as alignment entropy (Perrouchine et al., 2009), average log-likelihood, Akaike’s information criterion (Akaike, 1974) or the like. Another possibility would be to use a comprehensive Ω , but to penalize unlikely steps, which could be achieved by setting γ in Algorithm 1 to a ‘large’ real number and then, in subsequent runs, employ the remaining steps $S \subseteq \Omega$; we outline this approach in Section 7.

Sometimes, specific knowledge about a particular domain of application may be helpful, too. For example, in the field of G2P, we would expect most associations in alignments to be of the type M -to-1, i.e. one or several graphemes encode a single phoneme. This is because it seems reasonable to assume that the number of phonetic units used in language communities typically exceeds the number of units in alphabetic writing systems — 26 in the case of the Latin alphabet — so that one or several letters must be employed to represent a single phoneme. There may be 1-to- N or even M -to- N relationships but we would consider these exceptions. In the current work, we choose $S = \{(1, 1), (2, 1), (3, 1), (4, 1), (1, 2)\}$ for G2P data sets, and for the morphology data sets we either adopt from (Eger, 2012) or use a comprehensive Ω with ‘largest’ step $(2, 2)$.

6 Decoding

6.1 Training a string transduction model

We first generate monotone many-to-many alignments between string pairs with one of the procedures outlined in Sections 3 and 4. Then, we train a linear chain conditional random field (CRF; see (Lafferty et al., 2001)) as a graphical model for string transduction on the aligned data. The choice of CRFs is arbitrary; any transduction procedure tr would do, but we decide for CRFs because they generally have good generalization properties. In all cases, we use window sizes of three or four to predict y string elements from x string elements.

6.2 Segmentation

Our overall decoding procedure is as follows. Given an input string x , we exhaustively generate all possible segmentations of x , feeding the segmented strings to the CRF for transduction and evaluate each individual resulting sequence of ‘graphones’ with an n -gram model learned on the aligned data, taking the y string corresponding to the graphone sequence with maximal probability as the most likely transduced string for x . We illustrate in Algorithm 3.

Algorithm 3 Decoding

```

1: procedure DECODE( $x = x_1 \dots x_m; k^*, a, b,$ 
    $\text{tr}$ )
2:    $Z \leftarrow \emptyset$ 
3:   for  $s \in \mathcal{C}(m, k^*, a, b)$  do ▷
    $\mathcal{C}(m, k^*, a, b)$  : the set of all integer compositions of  $m$  with  $k^*$  parts, each between  $a$  and  $b$ 
4:      $\hat{y} \leftarrow \text{tr}(s)$ 
5:      $z_{\hat{y}} \leftarrow \text{ngramScore}(x, \hat{y})$ 
6:      $Z \leftarrow Z \cup \{z_{\hat{y}}\}$ 
7:   end for
8:    $z_{\hat{y}^*} \leftarrow \max_{z_{\hat{y}}} Z$ 
9:   return  $\hat{y}^*$ 
10: end procedure

```

As to the size of the search space that this procedure entails, note that any segmentation of a string x of length n with k parts uniquely corresponds to an *integer composition* (a way of writing n as a sum of non-negative integers) of the integer n with k parts, as illustrated below,

$$7 = \begin{array}{ccccccc} & \text{ph} & & \text{oe} & & \text{n} & & \text{i} & & \text{x} \\ & 2 & + & 2 & + & 1 & + & 1 & + & 1 \end{array}$$

It is a simple exercise to show that there are $\binom{n-1}{k-1}$ integer compositions of n with k parts, where by $\binom{n}{k}$ we denote the respective binomial coefficient. Furthermore, if we put restrictions on the maximal size of parts — e.g. in G2P a reasonable upper bound l on the size of parts would probably be 4 — we have that there are $\binom{k}{n-k}_l$ integer compositions of n with k parts, each between 1 and l , where by $\binom{k}{n-l+1}$ we denote the respective l -nomial or polynomial coefficient (Comtet, 1974). To avoid having to enumerate segmentations for all possible numbers k of segment parts of a given input string x of length n — these would range between 1 and n , entailing $\sum_{k=1}^n \binom{n-1}{k-1} = 2^{n-1}$ possible segmentations in total in the case without upper bound¹ — we additionally train a ‘number of parts’ prediction model with which to estimate k ; we call this in short *predictor model*.

To illustrate the number of possible segmentations with a concrete example, if x has length $n = 15$, a rather large string size given the values in Table 1, there are

2472, 2598, 1902, 990, 364, 91, 14, 1

possible segmentations of x with $k = 8, 9, 10, 11, 12, 13, 14, 15$ parts, each between 1 and 4.

7 Experiments

We conduct our experiments on three G2P data sets, the German Celex (G-Celex) and French Brulex data set (F-Brulex) taken from the Pascal challenge (van den Bosch et al., 2006), and the English Celex dataset (E-Celex). Furthermore, we apply our algorithms to the four German morphology data sets discussed in Dreyer et al. (2008), which we refer to, in accordance with the named authors, as rP, 2PKE, 13SIA and 2PIE, respectively. Both for the G2P and the morphology data, we hold monotonicity, by and large, a

¹In the case of upper bounds, Malandro (2012) provides asymptotics for the number of restricted integer compositions, which are beyond the scope of the present work, however.

2PKE	abrech et , entgegentret et , zuziehet
rP	z. abzubrechen, entgegenzutreten, zuzuziehen
2PKE	rP. redet, reib t , treibt, verbindet
13SIA	pA. geredet, gerieben, getrieben, verbunden
2PIE	

Table 2: String pairs in morphology data sets 2PKE and rP (omitting 2PIE and 13SIA for space reasons) discussed by (Dreyer et al., 2008). Changes from one form to the other are in bold (information not given in training). Adapted from Dreyer et al. (2008).

E-Celex	$\{(1, 1), (2, 1), (3, 1), (4, 1), (1, 2)\}$
rP	$\{(0, 2), (1, 1), (1, 2), (2, 1), (2, 2)\}$
2PKE	$\{(0, 2), (1, 1), (2, 1), (2, 2)\}$
13SIA	$\{(1, 1), (1, 2), (2, 1), (2, 2)\}$
2PIE	$\{(1, 1), (1, 2)\}$

Table 3: Data set and choice of S . Note that for all three G2P data sets, we select the same S , exemplarily shown for E-Celex. The choice of S for rP and 2PKE is taken from Eger (2012). For 13SIA and 2PIE we use comprehensive Ω ’s with largest step $(2, 2)$ but the algorithm ends up using just the outlined set of steps.

legitimate assumption so that our approach would appear justified. As to the morphology data sets, we illustrate in Table 2 a few string pair relationships that they contain, as indicated by Dreyer et al. (2008).

7.1 Alignments

We generate alignments for our data sets using Algorithms 1 and 2 and, as a comparison, we implement an exhaustive search bigram scoring model as indicated in Section 4 in an EM-like fashion similar as in Algorithm 2, employing the CMU SLM toolkit (Clarkson and Rosenfeld, 1997) with Witten-Bell smoothing as n -gram model. For Algorithm 1, which we also refer to as unigram model in the following, we choose steps S as shown in Table 3. As similarity measure sim, we use log prob with Good-Turing smoothing and for q we likewise use log prob; we outline the choice of L below. Initially, we set γ and χ to zero. As an alignment quality measure we consider conditional entropy $H(L | P)$ (or $H(P | L)$) as suggested by Pervouchine et al. (2009). Conditional entropy measures the average uncertainty of a (grapheme) substring L given a (phoneme) substring P ; apparently, the smaller $H(L | P)$ the better is the alignment because it

	Perplexity	$H(L P)$
2PKE-Uni	7.002 ± 0.04	0.094 ± 0.001
2PKE-Bi	6.865 ± 0.02	0.141 ± 0.003
rP-Uni	9.848 ± 0.09	0.092 ± 0.003
rP-Bi	9.796 ± 0.05	0.107 ± 0.006
Brulex-Uni	22.488 ± 0.35	0.706 ± 0.002
Brulex-Bi	22.215 ± 0.21	0.725 ± 0.003

Table 4: Conditional entropy vs. n -gram perplexity ($n = 2$) of alignments for different data sets. In bold: Statistically best results. $K = 300$ throughout.

produces more consistent associations.

In the following, all results are averages over several runs, 5 in the case of the unigram model and 2 in the case of the bigram model. Both for the bigram model and the unigram model, we select K , where $K \in \{50, 100, 300, 500\}$, training samples randomly in each EM iteration for alignment and from which to update probability estimates.

In Figure 2, we show learning curves over EM iterations in the case of the unigram and bigram models, and over training set sizes. We see that performance, as measured by conditional entropy, increases over iterations both for the bigram model and the unigram model (in Figure 2), but apparently alignment quality decreases again when too large training set sizes K are considered in the case of the bigram model (omitted for space reasons); similar outcomes have been observed when similarity measures other than log prob are employed in Algorithm 1 for the unigram model, e.g. the χ^2 similarity measure (Eger, 2012). To explain this, we hypothesize that the bigram model (and likewise for specific similarity measures) is more susceptible to overfitting when it is trained on too large training sets so that it is more reluctant to escape ‘non-optimal’ local minima. We also see that, apparently, the unigram model performs frequently better than the bigram model.

The latter results may be partly misleading, however. Conditional entropy, the way Pervouchine et al. (2009) have specified it, is a ‘unigram’ assessment model itself and may therefore be incapable of accounting for certain ‘contextual’ phenomena. For example, in the 2PKE and

rP data, we find position dependent alignments of the following type,

–	g	e	b	t	g	e	–	b	t
ge	g	e	b	en	g	e	ge	b	en

where we list the linguistically ‘correct’, due to the prefixal character of *ge* in German, alignment on the left and the ‘incorrect’ alignment on the right. By its specification, Algorithm 1 *must* assign both these alignments the same score and can hence not distinguish between them; *the same holds true for the conditional entropy measure*. To address this issue, we evaluate alignments by a second method as follows. From the aligned data, we extract a random sample of size 1000 and train an n -gram grapheme model (that can account for ‘positional associations’) on the residual, assessing its perplexity on the held-out set of size 1000. Results are shown in Table 4. We see that, in agreement with our visual impression at least for the morphology data, the alignments produced by the bigram model seem to be slightly more consistent in that they reduce perplexity of the n -gram grapheme model, whereas conditional entropy proclaims the opposite ranking.

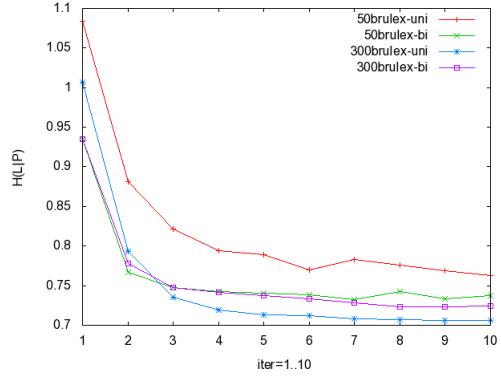


Figure 2: Learning curves over iterations for F-Brulex data, $K = 50$ and $K = 300$, for unigram and bigram models.

7.1.1 Quality q of steps

In Table 5 we report results when experimenting with the coefficient γ of the quality of steps measure q . Overall, we do not find that increasing γ would generally lead to a performance increase, as measured by e.g. $H(L | P)$. On the contrary, when choosing as set of steps a compe-

hensive Ω as in Table 5, where we choose $\Omega = \{(a, b) \mid a \leq 4, b \leq 4\} \setminus \{(0, 0)\}$, for $\gamma = 0$, we find values of 0.278, 0.546, 0.662 for $H(L \mid P)$ for G-Celex, F-Brulex and E-Celex, respectively, while corresponding values for $\gamma = 10$ are 0.351, 0.833, 1.401. Contrarily, $H(P \mid L)$, the putatively more indicative measure for transduction from \mathbf{x} to \mathbf{y} , has 0.499, 0.417, 0.598 for $\gamma = 0$ and 0.378, 0.401, 1.113 for $\gamma = 10$, so that, except for the E-Celex data, $\gamma = 10$ apparently leads to improved $H(P \mid L)$ values in this situation, while $\gamma = 0$ seems to lead to better $H(L \mid P)$ values.

In any case, from a model complexity perspective,² increasing γ may certainly be beneficial. For example, Table 5 shows that with $\gamma = 0$, Algorithm 1 will select up to 15 different steps for the given choice Ω , most of which seem linguistically questionable. On the contrary, with a large γ , Algorithm 1 employs only four resp. five different steps for the G2P data; most importantly, among these are (1, 1), (2, 1) and (3, 1), all of which are in accordance with linguistic reasoning as e.g. outlined in Section 5.

7.1.2 Colors

We shortly discuss here a possibility to detect latent classes via the concept of colored paths. Assume that a corpus of colored alignments is available and let each color be represented by the contexts (graphemes to the left and right) of its members; moreover, define the ‘likelihood’ L that the pair $p_{x,y} := (x_{\alpha_i-1+1}^{\alpha_i}, y_{\beta_i-1+1}^{\beta_i})$ is of color c as the (document) similarity (in an information retrieval sense) of $p_{x,y}$ ’s contexts with color c , which we can e.g. implement via the cosine similarity of the context vectors associated with $p_{x,y}$ and c . For number of colors $C = 2$, we then find, under this specification, the following kinds of alignments when running Algorithms 1 and 2 with $\gamma = 0$ and $\chi = 1$,

a	nn	u	al	ph	o	n	e	me
&	n	jU	1	f	@U	n	i	m
1	0	1	1	0	1	0	1	0

where we arbitrarily denote colors by 0 and 1, and use original E-Celex notation for phonemic

²Taking into account model complexity is, for example, in accordance with Occam’s razor or Akaike’s information criterion.

characters. It is clear that the algorithm has detected some kind of consonant/vowel distinction on a phonemic level here. We find similar kinds of latent classes for the other G2P data sets, and for the morphology data, the algorithm learns (less interestingly) to detect word endings and starts, under this specification.

7.2 Transductions

We report results of experiments on transducing \mathbf{x} strings to \mathbf{y} strings for the G2P data and the morphology data sets. We exclude E-Celex because training the CRF with our parametrizations (e.g. all features in window size of four) did regularly not terminate, due to the large size of the data set ($> 60,000$ string pairs). Likewise for computing resources reasons,³ we do not use ten-fold cross-validation but, as in Jiampojamarn et al. (2008), train on the first 9 folds given by the Pascal challenge, testing on the last. Moreover, for the G2P data, we use an ϵ -scattering model with steps $S = \{(1, 0), (1, 1)\}$ as a predictor model from which to infer the number of parts k^* for decoding and then apply Algorithm 3.⁴ For alignments, we use in all cases Algorithms 1 and 2. As reference for the G2P data, we give word accuracy rates as announced by Bisani and Ney (2008), Jiampojamarn et al. (2007), and Rama et al. (2009), who gives the Moses ‘baseline’ (Koehn et al., 2007).

For the morphology data we use exactly the same training/test data splits as in Dreyer et al. (2008). Moreover, because Dreyer et al. (2008) report all results in terms of window sizes of 3, we do likewise for this data. For decoding we do not use a (complex) predictor model here but rely on simple statistics; e.g. we find that for the class 13SIA, k^* is always in $\{m - 2, m - 1, m\}$, where m is the length of \mathbf{x} , so we apply Algorithm 3 three times and select the best scoring $\hat{\mathbf{y}}$ string. To avoid zeros in the decoding process (see discussion in Section 6.2), we replace the (0, 2) steps used in the rP and 2PKE data sets by a step (1, 3).

Results are shown in Table 6. Note that, for the G2P data, our approach always outperforms the

³E.g. a single run of the CRF on the G-Celex data takes longer than 24 hours on a standard PC.

⁴We train the ϵ -scattering model on data where all multi-character phonemes such as ks are merged to a single character, as obtained from the alignments as given by Algorithms 1 and 2.

	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(1, 2)	(1, 0)	(2, 3)	(3, 2)	(3, 3)	(4, 2)	(4, 3)	(4, 4)	(2, 2)	(0, 1)	(1, 3)
G-Celex	86.50	11.61	1.77	-	0.10	-	-	-	-	-	-	-	-	-	-
	86.14	8.17	1.63	0.02	0.00	2.56	0.10	0.04	0.01	0.09	0.91	0.28	-	-	-
F-Brulex	78.85	15.08	5.85	-	-	-	0.20	-	-	-	-	-	-	-	-
	75.64	13.80	2.52	0.36	0.07	5.07	0.29	0.10	0.02	0.38	1.01	0.68	-	-	-
E-Celex	88.87	6.58	3.05	-	-	-	-	-	-	-	-	-	1.29	0.18	-
	75.54	8.45	0.75	0.04	1.48	4.57	0.41	0.03	0.16	0.44	2.03	3.03	0.00	2.87	0.12

Table 5: Steps and their frequency masses in percent for different data sets for $\gamma = 10$ (top rows) and $\gamma = 0$ (bottom rows), averaged over two runs. We include only steps whose average occurrence exceeds 10.

best reported results for pipeline approaches (see below), while we are significantly below the results reported by Dreyer et al. (2008) for the morphology data in two out of four cases. On the contrary, when ‘pure’ alignments are taken into consideration — note that Dreyer et al. (2008) learn very complex latent classes with which to enrich alignments — our results are considerably better throughout. In almost all cases, we significantly beat the Moses ‘baseline’.

8 Conclusion

We have presented a simple and general framework for generating monotone many-to-many alignments that competes with Jiampojamarn et al. (2007)’s alignment procedure. Moreover, we have discussed crucial independence assumptions and, thus, limitations of this algorithm and shown that exhaustive enumeration (among other methods) can overcome these problems — in particular, due to the relatively small search space — in the field of monotone alignments. Additionally, we have discussed problems of standard alignment quality measures such as conditional entropy and have suggested an alternative decoding procedure for string transduction that addresses the limitations of the procedures suggested by Jiampojamarn et al. (2007) and Jiampojamarn et al. (2008).

References

- H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.
- H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. *The CELEX2 lexical database*. LDC96L14.
- T. Baldwin and H. Tanaka. 1999. Automated Japanese grapheme-phoneme alignment. In Proc. of the International Conference on Cognitive Science, 349–354.
- A.W. Black, K. Lenzo, and V. Pagel. 1998. *Issues in Building General Letter to Sound Rules*. In The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. ISCA.
- M. Bisani, and H. Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5): 434–451.
- P.F. Brown, J. Cocke, J., S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79–85.
- P.R. Clarkson and R. Rosenfeld. 1997. *Statistical Language Modeling Using the CMU-Cambridge Toolkit*. Proceedings ESCA Eurospeech.
- L. Comtet. 1974. *Advanced Combinatorics*. D. Reidel Publishing Company.
- A. Content, P. Mousty, and M. Radeau. 1990. Une base de données lexicales informatisée pour le français écrit et parlé. *L’Année Psychologique*, 551–566.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38.
- M. Dreyer, J. Smith, and J. Eisner. 2008. *Latent-Variable Modeling of String Transductions With Finite-State Methods*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, Hawaii.
- S. Eger. 2012. Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics. Submitted.
- L. Galescu, and J.F. Allen. 2001. *Bi-directional Conversion between Graphemes and Phonemes using a joint n-gram model*. Proc. 4th ISCA Tutorial and ResearchWorkshop on Speech Synthesis. Perthshire, Scotland.
- S. Heubach and T. Mansour. 2004. Compositions of n with parts in a set. *Congressus Numerantium*, 164: 127–143.
- H.H. Hoang, S.N. Kim, and M.-Y. Kan. 2009. *A Re-examination of Lexical Association Measures*. MWE ’09 Proceedings of the Workshop on Mul-

	CRF-3	CRF-4	CRF-4*	DSE-F	DSE-FL	Moses3	Moses15	M-M+HMM	BN	MeR+A*
F-Brulex		93.7	94.6					90.9	93.7	86.7
G-Celex		91.1	92.6					89.8		90.2
2PKE	79.8	80.9		74.7	87.4	67.1	<u>82.8</u>			
rP	74.1	<u>77.2</u>		69.9	84.9	67.6	70.8			
13SIA	85.6	<u>86.5</u>		82.8	87.5	73.9	85.3			
2PIE	94.6	94.2		88.7	93.4	92.0	94.0			

Table 6: Data sets and word accuracy rates in percent. **DSE-F**: Dreyer et al. (2008) using ‘pure’ alignments and features. **DSE-FL**: Dreyer et al. (2008) using alignments, features and latent classes. **Moses3**, **Moses15**: Moses system with window sizes of 3 and 15, respectively, as reported by Dreyer et al. (2008). **M-M+HMM**: Many-to-many aligner with HMM and instance-based segmenter for decoding as reported by Jiampojamarn et al. (2007). **BN**: Bisani and Ney (2008) using a machine translation motivated approach to many-to-many alignments. **MeR+A***: Results of Moses system on G2P data as reported by Rama et al. (2009). **CRF-3**: Our approach with window size of 3 and 3-gram ngram scoring model (see Algorithm 3). **CRF-4**: Our approach with window size of 4 and 3-gram scoring model. **CRF-4***: Our approach with window size of 4 and 4-gram scoring model and 2-best lists. In bold: Best results (no statistical tests). Underlined: best results using ‘pure’ alignments.

- tiword Expressions: Identification, Interpretation, Disambiguation and Applications.
- S. Jiampojamarn, G. Kondrak, and T. Sherif. 2007. *Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion*. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007), Rochester, NY, April 2007, 372–379.
- S. Jiampojamarn, C. Cherry and G. Kondrak. 2008. *Joint Processing and Discriminative Training for Letter-to-Phoneme Conversion*. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), 905–913, Columbus, OH, June 2008.
- S. Jiampojamarn, and G. Kondrak. 2010. *Letter-Phoneme Alignment: An Exploration*. The 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 780–788, Uppsala, Sweden. July 2010.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
- J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Proc. 18th International Conf. on Machine Learning, 282–289.
- M.E. Malandro. 2012. Asymptotics for restricted integer compositions. Preprint available at <http://arxiv.org/pdf/1108.0337v1>.
- M. Mohri. 2002. Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3): 321–350.
- S.B. Needleman and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48: 443–453.
- V. Pervouchine, H. Li, and B. Lin. 2009. *Transliteration alignment*. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore. Association for Computational Linguistics, 136–144.
- T. Rama, A.S. Kumar, and S. Kolachina. 2009. *Modeling Letter to Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training*. NAACL HLT 2009 Student Research Workshop, Colorado, USA.
- P. Taylor. 2005. *Hidden Markov Models for grapheme to phoneme conversion*. Proceedings of the 9th European Conference on Speech Communication and Technology 2005.
- J.D. Updyke. 2008. A unified approach to algorithms generating unrestricted and restricted integer compositions and integer partitions. *Journal of Mathematical Modelling and Algorithms*.
- A. van den Bosch, S.F. Chen, W. Daelemans, R.I. Damper, R.I., K. Gustafson, Y. Marchand, and F. Yvon. 2006. *Pascal letter-to-phoneme conversion challenge*. <http://www.pascalnetwork.org/Challenges/PRONALSYL>.
- G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann. 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1:98–106.
- W. Zucchini. 2000. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61.

Use of linguistic features for improving English-Persian SMT

Zakieh Shakeri Computer Eng. Dept. Alzahra University Tehran, Iran z.shakeri@ student.alzahra.ac.ir	Neda Noormohammadi HLT Lab. AmirKabir Uni. of Tech. Tehran, Iran noor@ce.sharif.edu	Shahram Khadivi HLT Lab. AmirKabir Uni. of Tech. Tehran, Iran khadivi@aut.ac.ir	Noushin Riahi Computer Eng. Dept. Alzahra University Tehran, Iran nriahi@ alzahra.ac.ir
---	--	--	---

Abstract

In this paper, we investigate the effects of using linguistic information for improvement of statistical machine translation for English-Persian language pair. We choose POS tags as helping linguistic feature. A monolingual Persian corpus with POS tags is prepared and variety of tags is chosen to be small. Using the POS tagger trained on this corpus, we apply a factored translation model. We also create manual reordering rules that try to harmonize the order of words in Persian and English languages.

In the experiments, factored translation model shows better performance compared to unfactored model. Also using the manual rules, which just contain few local reordering rules, increases the BLEU score compared to monotone distortion model.

1 Introduction

Machine translation is considered as a hard task because of differences between languages, referred to as translation divergences. Translation divergence could be structural, like differences in morphology, argument structure, and word order or it could be lexical like homonymous, many-to-many translation mappings and lexical gaps (Jurafsky et al., 2010). For the language pairs with less divergence, translation process is easier and output sentences have better quality, but the other language pairs suffer from this issue. In order to decrease the divergence effects, it is possible to incorporate linguistic information such as: words lemma, part-of-speech tags and morphological information in the translation process.

First we chose factored translation model introduced in (Koehn and Hoang , 2007) as the approach, and because of recent researches (Hoang , 2011) indicates that influence of POS tag is more than other features, the POS tag was chosen to be the helping feature. We needed accurate and suitable POS taggers for our purpose. For English language there was plenty of choices, but for Persian language we couldn't find a suitable tagger. Experiments showed that using basic POS tags rather than detailed ones result in better performance for our purpose. Bijankhan corpus (Oroumchian et al., 2006) was our only available option for training the tagger. Investigations showed that the corpus can not be used for our purpose without some modifications. First step was reducing the number of tags defined for Persian words and second step was correcting some effective mistakes in POS tags assigned to words. The job was done using the FLEXICON database of SCICT (SCICT , 2010).Stanford POS tagger was trained on the modified Bijankhan corpus. The result was an accurate POS tagger which served well for our purpose. Using the Moses SMT toolkit (Koehn et al., 2007), factored translation model was trained for the English-Persian language pair. The results showed improvement in BLEU (Papineni et al., 2002) measure.

As mentioned, one of the major problems in SMT is different word orders in source and target languages. So we also focused on improving the word reordering for statistical machine translation which involves Persian. An appropriate method for word reordering is using the pre-processing step before training (Popović and Ney,

2006; Matusov and Köprü , 2010). In this study, we propose manual reordering rules which rely on POS tags. Applying manual rules to the source sentences in the preprocessing step, leads to improved translation quality.

In the next section we have a brief introduction of Persian language structure and in Section 3 we will have a glance at factored translation model. we introduce our method on preparing a suitable Persian POS tagger on Section 4, and in Section 5 we suggest some manual rules to apply in pre-processing step and Section 6 shows the results of the related experiments.

2 Aspects of Persian syntax

Persian is a Subject Object Verb language, i.e. the type of language in which the subject, object, and verb of a sentence appear (usually) in that order. Like English, it is a member of the Indo-European language and has many common properties with other languages of this family like morphology, syntax, phonology, and lexicon. Persian is closer to Hindi and Urdu than English. Although the Persian Alphabet is like Arabic alphabet, but the language family of Arabic and Persian is different (Ghayoomi , 2004). There are many points about Persian language structure but in this article we only discuss about the main and key properties of its grammar.

1. Persian inflectional morphology (Megredoomian , 2000): Persian is an affixal system consisting mainly of suffixes and a few prefixes. Like Turkish and German languages, Persian has a rich morphological structure. This means that this language has a complete verbal inflectional system, which can be obtained by the combination of prefixes, stems inflections and auxiliaries. This feature is considered as a problem in MT because of large number of vocabularies.
2. Free word order (Mahootian, 1997; Megredoomian, 2004):Although SOV structure is said to be used by Persian language, Persian can have relatively free word order and the sentential constituents may occur in various positions in the clause. The main reason is that parts of speech are unambiguous. So

Persian has high degree of flexibility for verification. NLP field and text mining suffer from this parameter.

3. Nouns (Mahootian, 1997; Megredoomian, 2004): Nouns in Persian have no grammatical gender. They belong to an open class of words. The noun could be a common noun, a proper noun, or a pronoun. If this noun is not a proper noun or a pronoun, some elements can come before it and some after it. In addition, there are no overt markers, such as case morphology, to indicate the function of a noun phrase or its boundary; in Persian, only specific direct objects receive an overt marker.
4. Pronouns (Megredoomian , 2000) : Persian is a null-subject language, so personal pronouns is not mandatory.
5. Ezafe construction (Larson and Yamakido , 2005; Ghayoomi , 2004) : Basically, the Ezafe construction is composed of two or more words related to each other within the noun phrase. The Ezafe vowel é (or -ye) appears in between, suffixed to the noun. Although vowel é is pronounced in speech, it is not written which obscures the Persian syntax for NLP. Ezafe (Ez) appears on (Kahne-muyipour, 2000) a noun before another noun (attributive), a noun before an adjective, a noun before a possessor (noun or pronoun), an adjective before another adjective, a pronoun before an adjective, first names before last names or a combination of the above. Note that Ezafe only appears on a noun when it is modified. In other words, it does not appear on a bare noun (e.g. ‘كتاب’/ketāb/ ‘book’).
6. Verb (Roberts , 2003) : The verb functions primarily as the predicate in the clause. Finite and infinite forms are clearly distinguished. Finite verbs inflect for both tense and agreement with the subject while also taking negative, subjunctive and imperfective prefixes. Infinite forms do not inflect for tense and agreement with the subject, nor do they take the subjunctive or imperfective prefixes. Persian is a verb-final language,

whereas English has a verb-initial structure. This difference is not favorable for a MT task in which Persian and English are involved.

3 Factored translation model

Phrase-based models proposed by (Zens et al., 2002; Koehn , 2004; Koehn et al., 2007) translate contiguous sequences of words in the source sentence to contiguous words in the target. The term phrase in this case just means contiguous words, rather than any syntactic phrasal category.

The current phrase-based model, is limited to the mapping of small text chunks without any explicit use of linguistic information, may it be morphological, syntactic, or semantic. Such additional information has been demonstrated to be valuable by integrating it in preprocessing or postprocessing steps. However, a tighter integration of linguistic information into the translation model is desirable (Koehn and Hoang , 2007).

Factored model is an extension of phrase-based approach to statistical translation which tightly integrates linguistic information. This approach allows additional annotation at the word level. A word in this approach is not only a token, but a vector of factors that represent different levels of annotation. These factors can be surface form, lemma, part-of-speech, morphological features such as gender, count and case, automatic word classes, true case forms of words, shallow syntactic tags, as well as dedicated factors.

Factored translation model introduced in (Koehn and Hoang , 2007) follows strictly the phrase-based approach, with the additional decomposition of phrase translation into a sequence of mapping steps that either translate input factors into output factors, or generate additional output factors from existing output factors. All mapping steps operate on the same phrase segmentation of the input and output sentence into phrase pairs, so called synchronous factored models.

4 Preparing the suitable Persian POS tagger

Since we chose POS tags as the helping linguistic feature, we needed a bilingual tagged corpus. For English language there are suitable taggers but such taggers do not exist for Persian, so we

defined our first step as creating an accurate suitable POS tagger for Persian.

We chose POS tagger of Stanford University, which uses maximum entropy model, as our tagger.

To prepare a suitable tagged corpus, we used Bijankhan corpus which contains about 88,000 sentences, 2,600,000 manually tagged words with a tag set containing 40 POS labels. But that corpus did not particularly fit for our purposes as it had some influential incorrect tags for some words and variety of its tags was large. This was first noted when the bilingual corpus was prepared using the Stanford POS tagger trained on the unmodified Bijankhan corpus. Application of factored model on this version of bilingual corpus caused the BLEU measure to drop compared to the baseline.

As a result we tried to both limit the variety of tags and to correct the mistakes. Based on investigating the process of translating a few sentences, we deduced that POS tags basically helped with finding a word's suitable position in target language. For example a noun-adjective composite in English language is reversed in Persian language and the detailed type of the adjective or noun won't affect this.

Because of this we defined a new set of basic tags for Persian words and we trained the Stanford tagger with this version of Bijankhan (original words but with new defined tags). Table 1 shows the list of our new tags, the corresponding Bijankhan tags and also corresponding FLEXICON tags.

To correct the corpus, FLEXICON database of SCICT - which is an accurate lexicon of 66,000 words with extra information about them - was utilized. List of possible tags for the word form was created using FLEXICON. If the list contained the Bijankhan assigned tag, it was considered to be correct and vice versa. The algorithm can be found in 1.

To get possible tags for a word form, it was checked if it's non-Persian or numeric(containing only digits) and if it was so, the list was populated accordingly. For Persian forms, the word was looked for in the FLEXICON and all possible tags for its form were populated into the result list. For Persian forms that did not exist in FLEXICON,

Our new tag group	Bijankhan tag group	FLEXICON tag group
ADJ	ADJ_SUP, ADJ_SIM, ADJ_INO, ADJ_CMPR, ADJ_ORD, ADJ, DET, QUA	A0, A1, A2
N	N_SING, N_PL, SPEC	N1, N2, N3, N4
PO	P, PP	PO, PR, PR1
PP	PP	PR1
V_PR	V_PRS, V_SUB, V_AUX, V_IMP, V_PRE	V1, V3, V4, V5
V_PA	V_PA	V2, V3, V4, V5
ADV	ADV_NEGG, ADV_NI, ADV_EXM, ADV_TIME, ADV_I, ADV, MQUA	AD
CON	CON	C0, C1
PRO	PRO	N5, N6, N7, N8, NA
NO	NN	NO, NU
INT	INT, OH, OHH	INTJ
EXP	PS	AD, EXP
EN	MS	-

Table 1: Different Tag Groups

Algorithm 1 Correct the corpus

```

for all (word, tag)  $\in$  BKh corpus do
    pt  $\leftarrow$  GETPOSSIBLETAGS(word)
    if tag  $\in$  pt then
        label  $\leftarrow$  "Correct"
    else
        label  $\leftarrow$  "Incorrect"
    end if
end for

```

composite(derivative word) forms were checked using affix data. If after all, no possible tags were found for the word it would be considered as proper noun. Detailed algorithm can be found in 2.

The process of looking into derivatives can be found in Algorithm 3.

From the words that were recognized to be incorrectly tagged, those with only one acceptable tag in their list were labeled as correctable.

Finally correctable sentences were corrected by assigning their incorrect words, their known acceptable tag and a final modified version of Bijankhan corpus was ready. During the above mentioned process it was found that from 2,597,937 words in the corpus, 18,238 words were incorrectly tagged. We deleted them and their corresponding sentences from corpus which led to ex-

Algorithm 2 Get possible tags

```

function GETPOSSIBLETAGS(word)
    if word is English then return {EN}
    else if word is num. then return {NO}
    else if word  $\in$  FLEXICON then
        return FLEXICON tag list
    else
        result  $\leftarrow$  {}
        for all (affix, preTag, postTag)  $\in$ 
        Persian do
            if word contains affix then
                rem  $\leftarrow$  word - affix
                if EXISTS(rem, preTag) then
                    INSERT(result, postTag)
                end if
            end if
        end for
    end if
end function

```

traction of 70,470 usable sentences from 88,145 sentences. Although this process is not flawless, it at least makes the corpus more accurate.

From those 70,470 sentences, 60,470 sentences were randomly chosen to be used for training the Stanford POS tagger and 10,000 to be used for testing the resulting taggers accuracy.

Table 2 shows the result of training Stanford

	Accuracy of Correct tags	Accuracy of correct sentences	Accuracy of correct Unknown words
Original Bijankhan	97.50%	58.43%	82.81%
Modified Bijankhan: in tags	97.74%	61.25%	84.55%
Modified Bijankhan: in words and tags	99.36%	85.73%	86.59%

Table 2: result of training Stanford tagger with different versions of Bijankhan

Algorithm 3 Check if a word with given POS tag exists

```

function EXISTS(word, tag)
  if word ∈ FLEXICON then
    if tag ∈ FLEXICON tag list then
      return true
    end if
  end if
  for all (affix, preTag, postTag) ∈
  Persian do
    if word contains affix and tag =
    postTag then
      rem ← word – affix
      if EXISTS(rem, preTag) then
        return true
      end if
    end if
  end for
  return false
end function

```

tagger with different versions of Bijankhan.

5 Using manual rules in the preprocessing step

Persian is considered to be a verb-final language, but this language is not believed to follow a regular word order. This means the sentential constituents may occur in various positions in the clause. This feature makes the MT task more difficult and is a serious problem for MT. In the studies, it is supposed that word order of Persian is not free and follows a specified structure.

One of the ways suggested for reordering source sentence is to mimic the word order in target language. It can be performed by both statistical methods and syntactical methods. In this study we use syntactic rules utilizing POS tags.

POS tag-based reordering rules try to arrange the source word order according to the target word order. To this end, reordering rules are applied to the source sentences, and then translation step is performed monotonically.

In regards to the Persian syntax, the following items could be considered as needed reorderings for translation from and into Persian:

- Local reordering (appropriate for Ezafe construction)
- Long-range reordering (appropriate for verb reordering)

In this work, we propose rules for local reordering which occur widely in Persian. Although Persian verb reordering might be useful, there are some challenges for it. In what follows, we will attempt to argue why we do not use this reordering in the experiments.

Long-range verb reordering requires that the verb is moved towards the end of the clause where Persian is target language. So in the first place, it is needed that the boundary of clauses is detected, but there are no overt markers to indicate the boundary of clauses in written form of Persian (Iranpour and Minaei et al., 2009). However in some researches (Matusov and Köprü , 2010), punctuation marks and conjunctions have been used to achieve this goal. Since the accuracy of these markers is low and there are plenty of exceptions, using these hints does not lead to improved translation quality (Matusov and Köprü , 2010). The second point to take in to consideration is that Persian widely uses from compound verbs. So where Persian is source language, moving words one by one towards the beginning of clause may break the integrity of clause. Then we should be required to mark compound verbs in a

preprocessing step (Matusov and Köprü , 2010), but this task is ambiguous and includes lots of exceptions. So this preprocessing does not improve the quality of translation, too (Matusov and Köprü , 2010).

5.1 Local reordering

With respect to Persian syntax mentioned in the second section, we propose the following local reordering rules for translation tasks into Persian. The rules 1 and 2 are considered as the most important local reorderings in Persian. In what follows, we have supposed English is source language. Note that in each of the following rules, the Ezafe construction occurs.

1. Adjective group before a noun: Unlike English, adjectives using the Ezafe construction mostly follow the corresponding noun, whereas this order is the other way round in English. This reordering rule seems helpful. An example of this reordering rule is shown in Table 1. This rule can be expressed as follows:

$$JJ[JJ||CCJJ||, JJ]^*[NN||NNS] \rightarrow [NN||NNS]JJ([(JJ||CCJJ||, JJ)])^1$$
2. A noun before another noun: When the relationship between some nouns is needed to be shown (which Ezafe occurs), noun phrase is used. It is possible to use the preposition of which in this state, the order between words is compatible with Persian. For example (the) door of class. But this example could be expressed in other form: class door. For matching with word order in Persian, the word order of the ‘class door’ (dar-e kelas)’ phrase should be changed as ‘door class’ (it can be seen on Table 4). We represent this reordering as the following rule:

$$[((NN||NNS))]_1[((NN||NNS))]_2 \dots ((NN||NNS))_n \rightarrow (((NN||NNS))_n \dots (NN||NNS))_2 \dots ((NN||NNS))_1$$
3. Pronoun before a possessor: for example, ‘your offer/شما پیشنهاد شد (pishnahaad(-e) shoma)’ is changed to ‘offer your’.

¹JJ, CC, NN and NNS labels show the adjective, conjunction, noun and plural noun Parts Of Speech.

4. A noun before a possessor: the order between possessor and its related noun is changed. For example ‘Ali’s bag/کیف علی (kif-e Ali)’ is converted to ‘bag Ali’.

Phrase	Reordered phrase
Strong regularity and political control	control political and regularity Strong

Table 3: reordering of adjective group and corresponding noun (rule 1)

Phrase	Reordered phrase
Energy research programs	Programs research energy

Table 4: reordering of noun before another noun in Ezafe construction (rule 2)

6 Experiments

6.1 Experimental setup

In application of factored model of Moses, the experiments are in two directions from both English to Persian and Persian to English and have been done on two corpora, an open domain corpora and a limited domain one. For using manual rules, we did experiments in English to Persian direction and only on open domain corpora.

The open domain corpora is an in-house data for both training and test and were gathered from the following sources: news websites, articles, books available in two Persian and English languages, comparable texts like Wikipedia dump from which parallel texts were extracted, English texts which were translated to Persian by linguistics in-house, etc. To prepare parallel corpora, a particular method was used for each of mentioned sources, such as: Microsoft aligner, hunalign, in some cases document aligner and etc. The corpora statistics are shown in Table 5.

The limited domain corpora is an English-Persian version of Verbmobil parallel corpora. The statistics are shown in Table 6.

For the English side of corpora, we used the last version of Stanford POS tagger (Toutanova et al. , 2003), which is a maximum entropy based tagger. It models the sequence of words in a sentence as a

	English	Persian
Train: Sentences	305,000	305,000
Running words	6,884,823	7,666,405
Singleton	72,820	63,912
Tune: Sentences	1,000	1,000
Running words	25,567	22,582
Singleton	2,976	3,644
Test: Sentences	1,000	1,000
Running words	23,078	26,343
Singleton	3,824	3,046

Table 5: Statistics of open domain corpora in training, tuning and test data

bidirectional dependency network, which considers the lexical and tag context on both the sides to tag the current word. For Persian, POS tags were generated using the tagger which was prepared in course of this research. We additionally used the Moses toolkit as translation system. Also, the evaluation was performed using the automatic measure BLEU. We did the evaluation of the factored model with some further metrics.

6.2 Application of factored model

	English	Persian
Train: Sentences	23,145	23,145
Running words	249,355	216,577
Singleton	1,038	2,415
Tune: Sentences	276	276
Test: Sentences	526	526

Table 6: Statistics of limited domain corpora in the training, tuning and test data

Moses toolkit was utilized for training the factored model. Usage of POS tags for different phases of the translation process was experimented also different configuration options for training factored model was examined. Results showed that only application of source tags improved the quality so we continued our experiments on this configuration.

In Tables 7 and 8 we compare the results of training factored model with source language tag option on Verbmobil tagged corpus derived from different versions of Bijankhan.

As we can see in Table 7, applying factored

model on corpus which has been tagged with tagger trained on original Bijankhan, drops the Precision, Recall and also BLEU about 1.05% compared to baseline; But with using tagger trained on the Bijankhan where only its tags are our defined tags without any modification on the words, the BLEU increases about 0.47% compared to baseline, also Precision and Recall increase. And finally using tagger trained on Modified Bijankhan which the tags are our defined tags and its mistakes are removed, the BLEU increases about 0.74% compared to baseline, also Precision and Recall increase and TER decreases compared to other systems which was the best result.

En-Fr	BLEU
Baseline (phrase-based SMT)	16.87
Factored model	17.33(+0.46)

Table 8: result of application of factored model with English(source language) POS tags on limited domain corpora

As it shows, application of factored model improves the translation quality in English to Persian direction with a 0.46% increase in BLEU measure.

And in Tables 9 and 10 we compare the results of training factored model with source language tag option on tagged corpus derived from different versions of Bijankhan on the open domain corpora.

En-Fr	BLEU
Baseline (phrase-based SMT)	16.38
Factored model	16.52(+0.14)

Table 10: result of application of factored model with English(source language) POS tags on open domain corpora

6.3 Manual reordering rules

Table 11 presents the experimental results for the English to Persian MT system. As mentioned in Section 5, rules 1(adjective-noun rule) and 2(noun-noun rule) are the most important local reordering rules. For studying the influence of each

Fr-En	BLEU	Precision	Recall	TER
Baseline (phrase-based SMT)	25.18	73.16	65.81	50.86
FM with original Bkh	24.13(-1.05)	72.10	65.11	52.42
FM with Bkh modified in tags	25.65(+0.47)	75.03	64.77	50.48
FM with Bkh modified in tags & words	25.92(+0.74)	74.83	66.33	49.92

Table 7: results of application of factored model with different versions of Bijankhan on limited domain corpora

Fr-En	BLEU	Precision	Recall	TER
Baseline (phrase-based SMT)	18.00	65.00	60.95	65.54
FM with original Bkh	17.97(-0.03)	64.82	62.19	66.62
FM with Bkh modified in tags	18.10(+0.10)	64.88	62.03	66.60
FM with Bkh modified in tags & words	18.29(+0.29)	65.67	62.23	65.94

Table 9: results of application of factored model with different versions of Bijankhan on open domain corpora

of these two rules on the quality of output, we build two systems: the first system consisting of all rules stated in Section 5 except noun-noun rule and the second system including the same rules as the first system plus noun-noun rule. The monotonic translation and the distance-based translation are used as baselines for comparison. The results show unlike the second system that improves from 15.08% to 15.68% as compared to monotone model, the improvements are less significant for the latter system. This means that in Persian, noun-noun rule is used wider than adjective-noun rule. The second point to take in to consideration is that although the second system does not consist of long-range reorderings (which is very important for Persian), it shows the same result as distance-based model. This result demonstrates the accuracy and significance of these rules for Persian.

7 Conclusion

In case of languages for which accurate and large corpora are not readily available, simple statistical machine translation does not perform very

Reordering model	BLEU on dev set	BLEU on test set
Monotone	19.1	15.08
Distance-base	21.88	15.70
All manual rules minus noun-noun rule	19.95	15.12
All manual rules	20.22	15.68

Table 11: results of application of manual rules on development and test sets of open domain corpora

well. In such languages which include Persian language, linguistic features and language grammar can help the SMT to bring about better and acceptable results. To be able to utilize linguistic features, existence of tools for extracting accurate and suitable information is of vital importance.

In the course of our research, we tried to prepare one sample of such tools i.e. the POS tagger. We customized it for our purpose by defining a new tag set. The result of this effort seems to have positive effect on translation quality. Also there are plenty of approaches to using this data

to compensate for sparsity of parallel phrases that hasn't yet been examined.

In addition, we proposed reordering rules for harmonizing word order of source sentences with the structure of target language of which adjective-noun and noun-noun rules were the most significant. Experiments showed that noun-noun rule for translation into Persian is more effective.

References

- M. Ghayoomi, B. Guillaume. 2009 Interaction Grammar for the Persian Language: Noun and Adjectival Phrases. *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP* pages 107–114.
- H. Hoang. 2011 Improving Statistical Machine Translation with Linguistic Information. *PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics University of Edinburgh*.
- M. Iranpour Mobarakeh, B. Minaei-Bidgoli 2009 Verb Detection in Persian Corpora. *International Journal of Digital Content Technology and its Applications Volume 3, Number 1*.
- D. Jurafsky, J. Martin, A. Kehler, K. Vander Linden, and N. Ward. 2010. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, vol. 163.
- A. Kahnemuyipour. 2000 Persian Ezafe construction revisited: Evidence for modifier phrase. *Anual Conf. of the Canadian Linguistic Association*.
- P. Koehn, F. J. Och and D. Marcu. 2003 Statistical phrase based translation. *In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*. pp. 868-876.
- P. Koehn. 2004 Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *In AMTA*.
- P. Koehn, H. Hoang 2007 Factored Translation Models. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 868-876.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007 Moses: Open source toolkit for statistical machine translation. *in Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic*.
- R. Larson, H. Yamakido 2005 Ezafe and the Deep Position of Nominal Modifiers. *Barcelona workshop on adjectives and adverbs*.
- Sh. Mahootian 1997 Persian. *London: Routledge* p. 190.
- E. Matusov and S. Köprü. 2010 Improving Reordering in Statistical Machine Translation from Farsi. *in AMTA The Ninth Conference of the Association for Machine Translation in the Americas, Denver, Colorado, USA*.
- K. Megredoomian 2000 Unification-Based Persian Morphology. *In Proceedings of CICLing 2000. Alexander Gelbukh (ed.). Centro de Investigacion en Computacion-IPN, Mexico*, pages 311–318, Morristown, NJ, USA.
- K. Megredoomian 2003 Text Mining, Corpora Building and Testing. *In A Handbook for Language Engineers; edited by Ali Farghaly*, CSLI publications: Stanford, CA.
- K. Megredoomian 2004 Developing a Persian Part-of-Speech Tagger. *In Proceedings of the First Workshop on Persian Language and Computers*. Tehran University, Iran.
- K. Megredoomian 2004 A Semantic Template for Light Verb Constructions. *In Proceedings of the First Workshop on Persian Language and Computers* Tehran University, Iran.
- F. Oroumchian, S. Tasharofi, H. Amiri, H. Hojjat and F. Raja 2006 Creating a feasible corpus for Persian POS tagging. *Technical Report TR3/06, University of Wollongong in Dubai*.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002 BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics(ACL). Morristown, NJ, USA* pp. 311–318.
- M. Popović and H. Ney. 2006 POS-based Word Reorderings for Statistical Machine Translation. *5th International Conference on Language Resources and Evaluation (LREC)* pages 1278-1283, Genoa, Italy.
- J. Roberts 2003 Persian Grammar Sketch(book).
- SCICT FLEXICON database 2010. "http://prosody.ir/attachments/046_FLEXICON.rar"
- K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the AAAI'94 Workshop on CaseBased Reasoning*, pages 252-259
- R. Zens, F. J. Och, and H. Ney. 2002 Phrase-based statistical machine translation. *In Proceedings of the German Conference on Artificial Intelligence*.

A Semantic Similarity Measure Based on Lexico-Syntactic Patterns

Alexander Panchenko, Olga Morozova and Hubert Naets

Center for Natural Language Processing (CENTAL)

Université catholique de Louvain – Belgium

{Firstname.Lastname}@uclouvain.be

Abstract

This paper presents a novel semantic similarity measure based on lexico-syntactic patterns such as those proposed by Hearst (1992). The measure achieves a correlation with human judgements up to 0.739. Additionally, we evaluate it on the tasks of semantic relation ranking and extraction. Our results show that the measure provides results comparable to the baselines without the need for any fine-grained semantic resource such as WordNet.

1 Introduction

Semantic similarity measures are valuable for various NLP applications, such as relation extraction, query expansion, and short text similarity. Three well-established approaches to semantic similarity are based on WordNet (Miller, 1995), dictionaries and corpora. WordNet-based measures such as *WuPalmer* (1994), *LeacockChodorow* (1998) and *Resnik* (1995) achieve high precision, but suffer from limited coverage. Dictionary-based methods such as *ExtendedLesk* (Banerjee and Pedersen, 2003), *GlossVectors* (Patwardhan and Pedersen, 2006) and *WiktionaryOverlap* (Zesch et al., 2008) have just about the same properties as they rely on manually-crafted semantic resources. Corpus-based measures such as *ContextWindow* (Van de Cruys, 2010), *SyntacticContext* (Lin, 1998) or *LSA* (Landauer et al., 1998) provide decent recall as they can derive similarity scores directly from a corpus. However, these methods suffer from lower precision as most of them rely on a simple representation based on

the vector space model. *WikiRelate* (Strube and Ponzetto, 2006) relies on texts and/or categories of Wikipedia to achieve a good lexical coverage.

To overcome coverage issues of the resource-based techniques while maintaining their precision, we adapt an approach to semantic similarity, based on lexico-syntactic patterns. Bollegala et al. (2007) proposed to compute semantic similarity with automatically harvested patterns. In our approach, we rather rely on explicit relation extraction rules such as those proposed by Hearst (1992).

Contributions of the paper are two-fold. First, we present a novel corpus-based semantic similarity (relatedness) measure *PatternSim* based on lexico-syntactic patterns. The measure performs comparably to the baseline measures, but requires no semantic resources such as WordNet or dictionaries. Second, we release an Open Source implementation of the proposed approach.

2 Lexico-Syntactic Patterns

We extended a set of the 6 classical Hearst (1992) patterns (1-6) with 12 further patterns (7-18), which aim at extracting hypernymic and synonymous relations. The patterns are encoded in finite-state transducers (FSTs) with the help of the corpus processing tool UNITEX¹:

1. such NP as NP, NP[,] and/or NP;
2. NP such as NP, NP[,] and/or NP;
3. NP, NP [,] or other NP;
4. NP, NP [,] and other NP;
5. NP, including NP, NP [,] and/or NP;
6. NP, especially NP, NP [,] and/or NP;

¹<http://igm.univ-mlv.fr/~unitex/>

Name	# Documents	# Tokens	# Lemmas	Size
WaCypedia	2.694.815	$2.026 \cdot 10^9$	3.368.147	5.88 Gb
ukWaC	2.694.643	$0.889 \cdot 10^9$	5.469.313	11.76 Gb
WaCypedia + ukWaC	5.387.431	$2.915 \cdot 10^9$	7.585.989	17.64 Gb

Table 1: Corpora used by the *PatternSim* measure.

7. NP: NP, [NP,] and/or NP;
8. NP is DET ADJ.Superl NP;
9. NP, e. g., NP, NP[,] and/or NP;
10. NP, for example, NP, NP[,] and/or NP;
11. NP, i. e.[,] NP;
12. NP (or NP);
13. NP means the same as NP;
14. NP, in other words[,] NP;
15. NP, also known as NP;
16. NP, also called NP;
17. NP alias NP;
18. NP aka NP.

Patterns are based on linguistic knowledge and thus provide a more precise representation than co-occurrences or bag-of-word models. UNITER makes it possible to build negative and positive contexts, to exclude meaningless adjectives, and so on. Above we presented the key features of the patterns. However, they are more complex as they take into account variation of natural language expressions. Thus, FST-based patterns can achieve higher recall than the string-based patterns such as those used by Bollegala et al. (2007).

3 Semantic Similarity Measures

The outline of the similarity measure *PatternSim* is provided in Algorithm 1. The method takes as input a set of terms of interest C . Semantic similarities between these terms are returned in a $C \times C$ sparse similarity matrix S . An element of this matrix s_{ij} is a real number within the interval $[0; 1]$ which represents the strength of semantic similarity. The algorithm also takes as input a text corpus D .

As a first step, lexico-syntactic patterns are applied to the input corpus D (line 1). In our experiments we used three corpora: WACYPEDIA, UKWAC and the combination of both (see Table 1). Applying a cascade of FSTs to a corpus is a memory and CPU consuming operation. To make processing of these huge corpora feasible, we split the entire corpus into blocks of 250 Mb. Processing such a block took around one

hour on an Intel i5 M520@2.40GHz with 4 Gb of RAM. This is the most computationally heavy operation of Algorithm 1. The method retrieves all the concordances matching the 18 patterns. Each concordance is marked up in a specific way:

- such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}[PATTERN=1]
- {[traditional[food]}, such as {[sandwich]}, {[burger]}, and {[fry]}[PATTERN=2]

Figure brackets mark the noun phrases, which are in the semantic relation; nouns and compound nouns stand between the square brackets. We extracted 1.196.468 concordances K of this type from WACYPEDIA corpus and 2.227.025 concordances from UKWAC – 3.423.493 in total.

For the next step (line 2), the nouns in the square brackets are lemmatized with the DELA dictionary², which consists of around 300.000 simple and 130.000 compound words. The concordances which contain at least two terms from the input vocabulary C are selected (line 3).

Subsequently, the similarity matrix S is filled with frequencies of pairwise extractions (line 4). At this stage, a semantic similarity score s_{ij} is equal to the number of co-occurrences of terms in the square brackets within the same concordance e_{ij} . Finally, the word pairs are re-ranked with one of the methods described below (line 5):

Algorithm 1: Similarity measure *PatternSim*.

Input: Terms C , Corpus D
Output: Similarity matrix, S [$C \times C$]
 1 $K \leftarrow extract_concord(D)$;
 2 $K_{lem} \leftarrow lemmatize_concord(K)$;
 3 $K_C \leftarrow filter_concord(K_{lem}, C)$;
 4 $S \leftarrow get_extraction_freq(C, K)$;
 5 $S \leftarrow rerank(S, C, D)$;
 6 $S \leftarrow normalize(S)$;
 7 **return** S ;

Efreq (no re-ranking). Semantic similarity s_{ij} between c_i and c_j is equal to the frequency of extractions e_{ij} between the terms $c_i, c_j \in C$ in a set of concordances K .

Efreq-Rfreq. This formula penalizes terms that are strongly related to many words. In this case, semantic similarity of terms equals: $s_{ij} = \frac{2 \cdot \alpha \cdot e_{ij}}{e_{i*} + e_{*j}}$, where $e_{i*} = \sum_{j=1}^{|C|} e_{ij}$ is a number of

²Available at <http://infolingu.univ-mlv.fr/>

concordances containing word c_i and α is an expected number of semantically related words per term ($\alpha = 20$). Similarly, $e_{*j} = \sum_{i=1}^{|C|} e_{ij}$.

Efreq-Rnum. This formula also reduces the weight of terms which have many relations to other words. Here we rely on the number of extractions b_{i*} with a frequency superior to β : $b_{i*} = \sum_{j:e_{ij} \geq \beta} 1$. Semantic ranking is calculated in this case as follows: $s_{ij} = \frac{2 \cdot \mu_b \cdot e_{ij}}{b_{i*} + b_{*j}}$, where $\mu_b = \frac{1}{|C|} \sum_{i=1}^{|C|} b_{i*}$ – is an average number of related words per term and $b_{*j} = \sum_{i:e_{ij} \geq \beta} 1$. We experiment with values of $\beta \in \{1, 2, 5, 10\}$.

Efreq-Cfreq. This formula penalizes relations to general words, such as “item”. According to this formula, similarity equals: $s_{ij} = \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$, where $P(c_i, c_j) = \frac{e_{ij}}{\sum_{ij} e_{ij}}$ is the extraction probability of the pair $\langle c_i, c_j \rangle$, $P(c_i) = \frac{f_i}{\sum_i f_i}$ is the probability of the word c_i , and f_i is the frequency of c_i in the corpus. We use the original corpus D and the corpus of concordances K to derive f_i .

Efreq-Rnum-Cfreq. This formula combines the two previous ones: $s_{ij} = \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$.

Efreq-Rnum-Cfreq-Pnum. This formula integrates information to the previous one about the number of patterns $p_{ij} = \overline{1, 18}$ extracted given pair of terms $\langle c_i, c_j \rangle$. The patterns, especially (5) and (7), are prone to errors. The pairs extracted independently by several patterns are more robust than those extracted only by a single pattern. The similarity of terms equals in this case: $s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$.

Once the reranking is done, the similarity scores are mapped to the interval $[0; 1]$ as follows (line 6): $\hat{S} = \frac{S - \min(S)}{\max(S)}$. The method described above is implemented in an Open Source system *PatternSim*³ (LGPLv3).

4 Evaluation and Results

We evaluated the similarity measures proposed above on three tasks – correlations with human judgements about semantic similarity, ranking of word pairs and extraction of semantic relations.⁴

³<https://github.com/cental/PatternSim>

⁴Evaluation scripts and the results: <http://cental.fltr.ucl.ac.be/team/panchenko/sim-eval>

4.1 Correlation with Human Judgements

We use three standard human judgement datasets – MC (Miller and Charles, 1991), RG (Rubenstein and Goodenough, 1965) and WordSim353 (Finkelstein et al., 2001), composed of 30, 65, and 353 pairs of terms respectively. The quality of a measure is assessed with Spearman’s correlation between vectors of scores.

The first three columns of Table 2 present the correlations. The first part of the table reports on scores of 12 baseline similarity measures: three WordNet-based (*WuPalmer*, *Lecock-Chodorow*, and *Resnik*), three corpus-based (*ContextWindow*, *SyntacticContext*, and *LSA*), three definition-based (*WiktionaryOverlap*, *GlossVectors*, and *ExtendedLesk*), and three *WikiRelate* measures. The second part of the table presents various modifications of our measure based on lexico-syntactic patterns. The first two are based on WACKY and UKWAC corpora, respectively. All the remaining *PatternSim* measures are based on both corpora (WACKY+UKWAC) as, according to our experiments, they provide better results. Correlations of measures based on patterns are comparable to those of the baselines. In particular, *PatternSim* performs similarly to the measures based on WordNet and dictionary glosses, but requires no hand-crafted resources. Furthermore, the proposed measures outperform most of the baselines on the WordSim353 dataset achieving a correlation of 0.520.

4.2 Semantic Relation Ranking

In this task, a similarity measure is used to rank pairs of terms. Each “target” term has roughly the same number of meaningful and random “relatums”. A measure should rank semantically similar pairs higher than the random ones. We use two datasets: BLESS (Baroni and Lenci, 2011) and SN (Panchenko and Morozova, 2012). BLESS relates 200 target nouns to 8625 relatums with 26.554 semantic relations (14.440 are meaningful and 12.154 are random) of the following types: hypernymy, co-hypernymy, meronymy, attribute, event, or random. SN relates 462 target nouns to 5.910 relatum with 14.682 semantic relations (7.341 are meaningful and 7.341 are random) of the following types: synonymy, hypernymy, co-hypernymy, and random. Let R be a set of cor-

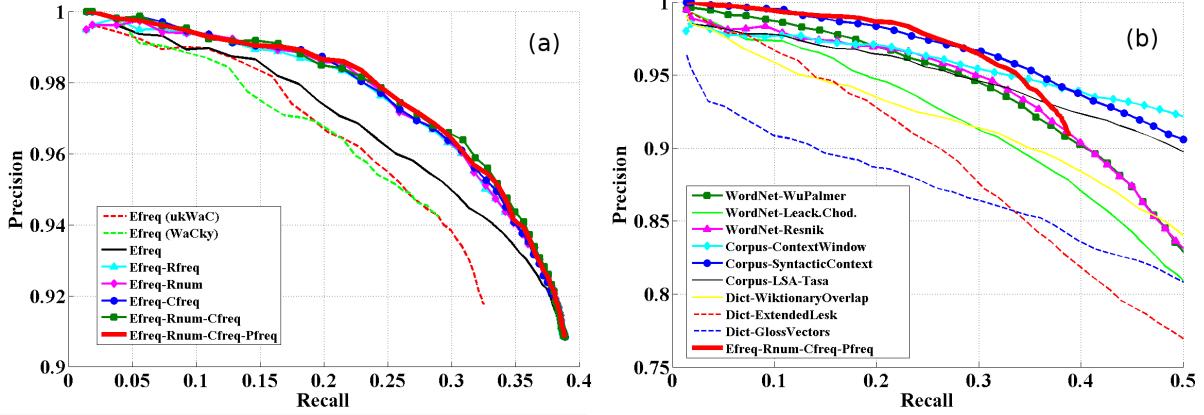


Figure 1: Precision-Recall graphs calculated on the BLESS (hypo, cohypo, mero, attri, event) dataset: (a) variations of the *PatternSim* measure; (b) the best *PatternSim* measure as compared to the baseline similarity measures.

rect relations and \hat{R}_k be a set of semantic relations among the top $k\%$ nearest neighbors of target terms. Then precision and recall at k are defined as follows: $P(k) = \frac{|R \cap \hat{R}_k|}{|\hat{R}_k|}$, $R(k) = \frac{|R \cap \hat{R}_k|}{|R|}$. The quality of a measure is assessed with $P(10)$, $P(20)$, $P(50)$, and $R(50)$.

Table 2 and Figure 1 present the performance of baseline and pattern-based measures on these datasets. Precision of the similarity scores learned from the WACKY corpus is higher than that obtained from the UKWAC, but recall of UKWAC is better since this corpus is bigger (see Figure 1 (a)). Thus, in accordance with the previous evaluation, the biggest corpus WACKY+UKWAC provides better results than the WACKY or the UKWAC alone. Ranking relations with extraction frequencies (*Efreq*) provides results that are significantly worse than any re-ranking strategies. On the other hand, the difference between various re-ranking formulas is small with a slight advantage for *Efreq-Rnum-Cfreq-Pnum*.

The performance of the *Efreq-Rnum-Cfreq-Pnum* measure is comparable to the baselines (see Figure 1 (b)). Furthermore, in terms of precision, it outperforms the 9 baselines, including syntactic distributional analysis (*Corpus-SyntacticContext*). However, its recall is seriously lower than the baselines because of the sparsity of the pattern-based approach. The similarity of terms can only be calculated if they co-occur in the corpus within an extraction pattern. Contrastingly, *PatternSim* achieves both high recall and precision on BLESS dataset containing only hy-

ponyms and co-hyponyms (see Table 2).

4.3 Semantic Relation Extraction

We evaluated relations extracted with the *Efreq* and the *Efreq-Rnum-Cfreq-Pnum* measures for 49 words (vocabulary of the RG dataset). Three annotators indicated whether the terms were semantically related or not. We calculated for each of 49 words extraction precision at $k = \{1, 5, 10, 20, 50\}$. Figure 2 shows the results of this evaluation. For the *Efreq* measure, average precision indicated by white squares varies between 0.792 (the top relation) and 0.594 (the 20 top relations), whereas it goes from 0.736 (the top relation) to 0.599 (the 20 top relations) for the *Efreq-Rnum-Cfreq-Pnum* measure. The interrater agreement (Fleiss's kappa) is substantial (0.61-0.80) or moderate (0.41-0.60).

5 Conclusion

In this work, we presented a similarity measure based on manually-crafted lexico-syntactic patterns. The measure was evaluated on five ground truth datasets (MC, RG, WordSim353, BLESS, SN) and on the task of semantic relation extraction. Our results have shown that the measure provides results comparable to the baseline WordNet-, dictionary-, and corpus-based measures and does not require semantic resources.

In future work, we are going to use a logistic regression to choose parameter values (α and β) and to combine different factors (e_{ij} , e_{i*} , $P(c_i)$, $P(c_i, c_j)$, p_{ij} , etc.) in one model.

Similarity Measure	MC		RG		WS		BLESS (hypo,cohypo,mero,attri,event)				SN (syn, hypo, cohypo)				BLESS (hypo, cohypo)				
	ρ	P	ρ	P	ρ	P	$P(10)$	$P(20)$	$P(50)$	$R(50)$	$P(10)$	$P(20)$	$P(50)$	$R(50)$	$P(10)$	$P(20)$	$P(50)$	$R(50)$	
Random	0.056	-0.047	-0.122	0.546	0.542	0.544	0.522	0.504	0.502	0.499	0.498	0.271	0.279	0.286	0.502				
WordNet-WuPalmer	0.742	0.775	0.331	0.974	0.929	0.702	0.674	0.982	0.959	0.766	0.763	0.977	0.932	0.547	0.968				
WordNet-Leack.Chod.	0.724	0.789	0.295	0.953	0.901	0.702	0.648	0.984	0.953	0.757	0.755	0.951	0.897	0.542	0.957				
WordNet-Resnik	0.784	0.757	0.331	0.970	0.933	0.700	0.647	0.948	0.908	0.724	0.722	0.968	0.938	0.542	0.956				
Corpus-ContextWindow	0.693	0.782	0.466	0.971	0.947	0.836	0.772	0.974	0.932	0.742	0.740	0.908	0.828	0.502	0.886				
Corpus-SynContext	0.790	0.786	0.491	0.985	0.953	0.811	0.749	0.978	0.945	0.751	0.743	0.979	0.921	0.536	0.947				
Corpus-LSA-Tasa	0.694	0.605	0.566	0.968	0.937	0.802	0.740	0.903	0.846	0.641	0.609	0.877	0.775	0.467	0.824				
Dict-WiktionaryOverlap	0.759	0.754	0.521	0.943	0.905	0.750	0.679	0.922	0.887	0.725	0.656	0.837	0.769	0.518	0.739				
Dict-GlossVectors	0.653	0.738	0.322	0.894	0.860	0.742	0.686	0.932	0.899	0.722	0.709	0.777	0.702	0.449	0.793				
Dict-ExtendedLesk	0.792	0.718	0.409	0.937	0.866	0.711	0.657	0.952	0.873	0.655	0.654	0.873	0.751	0.464	0.820				
WikiRelate-Gloss	0.460	0.460	0.200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WikiRelate-Leack.Chod.	0.410	0.500	0.480	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WikiRelate-SVM	-	-	0.590	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Efreq (WaCky)	0.522	0.574	0.405	0.971	0.950	0.942	0.289	0.930	0.912	0.897	0.306	0.976	0.937	0.923	0.626				
Efreq (ukWaC)	0.384	0.562	0.411	0.974	0.944	0.918	0.325	0.922	0.905	0.869	0.329	0.971	0.926	0.884	0.653				
Efreq	0.486	0.632	0.429	0.980	0.945	0.909	0.389	0.938	0.915	0.866	0.400	0.976	0.929	0.865	0.739				
Efreq-Rfreq	0.666	0.739	0.508	0.987	0.955	0.909	0.389	0.951	0.922	0.867	0.400	0.983	0.940	0.865	0.739				
Efreq-Rnum	0.647	0.720	0.499	0.989	0.955	0.909	0.389	0.951	0.922	0.867	0.400	0.983	0.940	0.865	0.739				
Efreq-Cfreq	0.600	0.709	0.493	0.989	0.956	0.909	0.389	0.949	0.920	0.867	0.400	0.986	0.948	0.865	0.739				
Efreq-Cfreq (concord.)	0.666	0.739	0.508	0.986	0.954	0.909	0.389	0.952	0.921	0.867	0.400	0.984	0.944	0.865	0.739				
Efreq-Rnum-Cfreq	0.647	0.737	0.513	0.988	0.959	0.909	0.389	0.953	0.924	0.867	0.400	0.987	0.947	0.865	0.739				
Efreq-Rnum-Cfreq-Pnum	0.647	0.737	0.520	0.989	0.957	0.909	0.389	0.952	0.924	0.867	0.400	0.985	0.947	0.865	0.739				

Table 2: Performance of the baseline similarity measures as compared to various modifications of the *PatternSim* measure on human judgements datasets (MC, RG, WS) and semantic relation datasets (BLESS and SN).

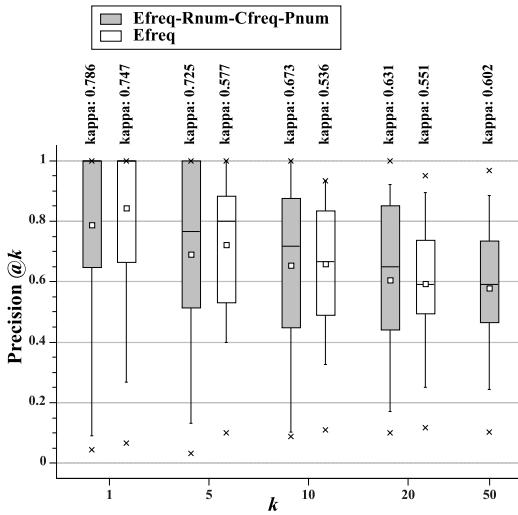


Figure 2: Semantic relation extraction: precision at k .

References

- S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 18, pages 805–810.
- M. Baroni and A. Lenci. 2011. How we blessed distributional semantic evaluation. In *GEMS (EMNLP), 2011*, pages 1–11.
- D. Bollegala, Y. Matsuo, and M. Ishizuka. 2007. Measuring semantic similarity between words using web search engines. In *WWW*, volume 766.
- L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: The concept revisited. In *WWW 2001*, pages 406–414.
- M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL*, pages 539–545.
- T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.
- C. Leacock and M. Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet*, pages 265–283.
- D. Lin. 1998. Automatic retrieval and clustering of similar words. In *ACL*, pages 768–774.
- G. A. Miller. 1995. Wordnet: a lexical database for english. *Communications of ACM*, 38(11):39–41.
- A. Panchenko and O. Morozova. 2012. A study of hybrid similarity measures for semantic relation extraction. *Hybrid Approaches to the Processing of Textual Data (EACL)*, pages 10–18.
- S. Patwardhan and T. Pedersen. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 1–12.
- P. Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI*, volume 1, pages 448–453.
- M. Strube and S. P. Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 21, pages 14–19.
- T. Van de Cruys. 2010. *Mining for Meaning: The Extraction of Lexico-Semantic Knowledge from Text*. Ph.D. thesis, University of Groningen.
- Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *ACL'1994*, pages 133–138.
- T. Zesch, C. Müller, and I. Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC'08*, pages 1646–1652.

A Multi-level Annotation Model for Fine-grained Opinion Detection in German Blog Comments

Bianka Trevisan

Textlinguistics/ Technical
Communication
RWTH Aachen University
b.trevisan@tk.rwth-
aachen.de

Melanie Neunerdt

Institute for Theoretical
Information Technology
RWTH Aachen University
neunerdt@ti.rwth-
aachen.de

Eva-Maria Jakobs

Textlinguistics/ Technical
Communication
RWTH Aachen University
e.m.jakobs@tk.rwth-
aachen.de

Abstract

Subject of this paper is a fine-grained multi-level annotation model to enhance opinion detection in German blog comments. Up to now, only little research deals with the fine-grained analysis of evaluative expressions in German blog comments. Therefore, we suggest a multi-level annotation model where different linguistic means as well as linguistic peculiarities of users' formulation and evaluation styles in blog comments are considered. The model is intended as a basic scheme for the annotation of evaluative expressions in blog data. This annotation provides suitable features for implementing methods to automatically detect user opinions in blog comments.

1 Introduction

Evaluations are complex linguistic acts that are realized through the use of linguistic means. In social media applications and blogs respectively, evaluations are, moreover, expressed by colloquial or grapho-stylistic means such as emoticons, multiple punctuations or Internet lingo (Schlobinski and Siever, 2005; Trevisan and Jakobs, 2010, 2012; Neunerdt et al., 2012). So far these text type-specific evaluative means are not considered in annotation models as most approaches focusing on strongly structured data such as newspaper texts where colloquial expressions are not common, generally. Therefore, we suggest a fine-grained multi-level annotation model, which consists of different annotation levels with various purposes of annotation, e.g., annotation of polarity vs. annotation of syntactical function. The model serves for the description of user-, context-, topic- and blog comment-related characteristics of opinion-indicating ex-

pressions in blog comments. The aim of the model is to determine how different users express themselves judgmental on a specific topic in blog comments, i.e., which linguistic means (typographic, rhetoric, syntactical etc.) they use in what combination and for what purpose, e.g., interaction signs at the end of a sentence to mark ironic expressions. We refer to this phenomenon as *evaluation pattern*. In this paper, the annotation model is presented using the example of emoticons.

The approach generates additional values for different disciplines. In computer sciences, for instance, the multi-level annotation model can serve as instrument for automatic opinion detection, where information on each annotation level serves as separate feature for classification. From the perspective of communication science, the fine-grained annotation model can be used for sentiment analysis, e.g., analysis of the grammatical function of an emoticon vs. its stylistic function. Up to now, the model is exemplarily used in acceptance research for the identification and analysis of opinion-indicating statements about *mobile communication systems (MCS)*.

The paper is structured as follows. Section 2 outlines characteristics of evaluative expressions and formulation styles in German blog comments. In section 3, related work on sentiment analysis and linguistic annotation are presented. Subject of Section 4 is the methodological approach and the developed multi-level annotation model; the different annotation levels and the related annotation schemes are thoroughly described. Results focusing on the validity and the identification of evaluation patterns are shown in

section 5. Finally, a summary and an outlook on future work are given in section 6.

2 Linguistic Means of Evaluation

In evaluating, writers use specific evaluative means that can be positive or negative connotated lexemes, specific syntactic structures (e.g. interrogative sentences), idioms, rhetorical means (e.g. metaphor) or non-verbal signs (e.g. emoticons) (Schlobinski and Siever, 2005). As we know from other contexts, which evaluative mean is chosen, depends on the function and purpose of the evaluation, group membership, the evaluating individual, the respective situation and the used text type (Sandig, 2003).

Moreover, the language in blogs has a low standardization degree, i.e. bloggers have specific modes of expression or formulation styles that are not characterized by recurring and recognizable structural features. This specific expression style can lead to processing errors, and, thus, impede the analysis accuracy of evaluative expressions in blog comments (Neunerdt et al., 2012).

Up to now, there is little research that deals with formulation and evaluation styles or patterns in blog comments. Rather, formulation styles of blog texts in general are described. First and foremost, it is evident that the language or vocabulary of bloggers is ordinary or colloquial (e.g. geil “cool”) and, thereby, bears analogy to spoken language (Schlobinski and Siever, 2005; Thorlechter et al., 2010). Typically, interaction words (e.g. IMO “in my opinion”) or inflective expressions (e.g. seufz “sigh”) are used as means of emphasis. Syntactic characteristics are simple word order and unclear sentence boundaries, e.g., because of missing sentence-terminating punctuations (Missen et al., 2009).

Regarding grapho-stylistic means, emoticons are characteristic for blog texts. They perform the function of conversation particles, interjections, prosody, facial expressions and gestures and have, in conclusion, an expressive and evaluative function (Mishne, 2005). Asterisks play a similar role; they are used to mark emotive and evaluative speech acts, e.g., in ironic expressions (e.g. *grins* “*smile*”). Further typographical means of blog communication are word-internal capitalizations (e.g. CrazyChicks), iteration of letters (e.g. jaaaaaa “yeeeees”) and italic (e.g. Das ist eine *große* Lüge. “That's a *big*

lie.”). Regarding stylistic means, most likely redemptions, e.g., of word endings occur (e.g. ‘nen “einen/one”) (Schlobinski and Siever, 2005). Other stylistic features are dialectic expressions (e.g. Mädl “girl”), foreign-language-switch phenomena (e.g. Killerapplikation “killer application”), increased use of rhetorical phrases and idioms (e.g. Alle Wege führen nach Rom. “All roads lead to Rome.”) as well as ellipsis.

In total, on all linguistic levels transfer phenomena from written to spoken language are evident, which also impact the processing of evaluative expressions. Thus, an annotation scheme by which evaluative means and expressions can be analyzed holistically must consider the different evaluative means and relate relevant linguistic information to each other.

3 Related Work

The exploitation of opinions from unstructured content such as blog comments is a challenging task, because evaluative expressions are vague, have a high degree of semantic variability (Balahur and Montoyo, 2010), and, as shown in section 2, text type-specific formulation styles occur. How well users evaluations are captured, thus, depends crucially on the analysis approach. Up to date, there is little research done regarding the analysis of evaluative expressions in German blog comments. In German research, mainly annotation schemes for diachronic corpora such as forum and chat data are developed (Storrer, 2007; Luckhardt, 2009; Beißwenger, 2008, 2009, 2010; Beißwenger et al., 2012). The following section focuses established approaches in opinion detection and annotation of complex tasks.

3.1 Opinion Annotation

Depending on corpus, text type (written vs. spoken language), language, annotation task and annotation goal, the extent and kind of annotation or coding models varies. In our case, we annotate evaluative expressions in German blog comments, with the aim of identifying recurring evaluation patterns.

Annotation approaches and models for opinion detection and sentiment analysis can be distinguished according to their degree of granularity. Using *coarse-grained annotation schemes*, texts are annotated on the document

and sentence level with the aim of making general statements about the polarity, e.g., as it is of interest for product reviews. This annotation type is normally used in marketing research using shallow text-processing tools (e.g. PASW Modeler, evaluated in Trevisan and Jakobs, 2012). However, coarse-grained approaches possess the advantage that they provide a higher tagging or annotation accuracy compared to fine-grained annotations. Giesbrecht et al. (2009) have already shown this for part-of-speech (POS) tagging of German mixed corpora, Gimbel et al. (2011) reached a better annotator agreement for the tagging of English tweets.

So far, several coarse-grained approaches have been developed. Aman and Szpakowicz (2007) proposed a coarse-grained annotation scheme for English blog posts. According to the approach, emotional expressions are being identified in blog posts and assigned to the six basic emotions happiness, sadness, anger, disgust, surprise and fear by Ekman (1993). Every sentence, in which one or more related keyword occurs, is manually assigned to a basic emotion and classified by polarity. Strapparava and Mihalcea (2007) and Balahur and Montoyo (2010) developed a similar approach. However, according to Stoyanov and Cardie (2008) with a coarse-grained annotation, only a distinction between subjective and objective statements is possible, but they cannot be related to the object of evaluation.

In contrast, *fine-grained annotation schemes* serve for the annotation on the phrase and clause level or below. Fine-grained annotations are especially required for scientific purposes. In acceptance research, for instance, fine-grained annotations are essential for identifying evaluated components and properties of large-scale technologies such as *mobile communication systems (MCS)* (Trevisan and Jakobs, 2010, 2012).

In recent years, a number of fine-grained annotation approaches has been developed. Initially, Wiebe et al. (2005) and Wilson (2008) provided fine-grained annotation approaches. In the further course, Stoyanov and Cardie (2008) proposed a topic annotation model that serves for the identification of opinion topics in text corpora. Annotated are six evaluation components: opinion expression, source (opinion holder), polarity, topic, topic span and target span. The inter-annotator results show that the

fine-grained annotation system provides reliable results. Remus und Häning (2011) present the *Polarity Composition Model (PCM)*, which is a two-level structure, where evaluations are analyzed on the word- and phrase-level. The authors draw the conclusion that evaluations are manifested through negations, positive and negative reinforces and the formal construct within the phrase; the word-level polarity analysis is carried out with recourse to the German-language polarity dictionary *SentiWS* (see also Remus et al., 2010). Lately, Fink et al. (2011) published their fine-grained annotation approach for sentiment analysis. Thereby, they identify sentiment-bearing sentences to spot sentiment targets and their valence.

Fine-grained and, especially, coarse-grained annotations schemes are used for qualitative and automatic text analysis methods. However, these annotation approaches do not allow the investigation of specific linguistic analysis purposes and furthermore do not provide a well-separated feature space for automatic opinion detection methods as multi-level annotation schemes do.

3.2 Annotation Models

Each token fulfills different grammatical functions, which are also relevant for the constitution of evaluations. In usual annotation schemes, these different information are not separated from each other. Queries and studies related to one type of linguistic mean are therefore difficult to perform. In multi-level annotations systems, several information such as different linguistic or evaluative means can be assigned independently to a single token or sequence of tokens, e.g., the morpho-syntactic function of a token vs. its semantic meaning (Lüdeling et al., 2005). Such multi-level annotations systems are commonly used for the annotation of learner corpora and transcripts (Lüdeling, 2011). Regarding the annotation of evaluative expressions in blog comments, multi-level annotation systems provide several advantages compared to flat or tabular annotation models, such as in TreeTagger used (Lüdeling et al., 2005; Dipper, 2005):

- Level-specific annotation standards can be applied.
- The number and categories of annotation levels is expandable and modifiable dur-

ing the editing process, e.g., adding or deleting of an annotation level.

- Token columns are mergeable and separable again, e.g., in case of multi-word expressions.
- Characteristic features of the text data can be considered, e.g., in case of written texts providing a linear mapping.

Adapting this approach, evaluative expressions in blog comments can be analyzed holistically. However, standards for processing evaluative expressions in German blog comments are mainly missing, such as a multi-level annotation scheme and text type-specific annotation guidelines for multi-level architectures¹ (Balahur and Montoyo, 2010; Lüdeling, 2011; Clematide et al., 2012), which are part of the proposed model.

4 Approach

The aim is to develop a multi-level annotation scheme for the identification of evaluative expressions in blog comments. In the following, the process of model development is described.

4.1 Corpus

As an exemplary corpus, a topic-specific German blog dealing with mobile communication systems (MCS) is selected and blog comments from two years semi-automatically collected ($t=2008, 2009$) (Trevisan and Jakobs, 2012). In total, the corpus contains 12,888,453 tokens and 160,034 blog comments². The collected blog comment corpus was bowdlerized: Enclosed website elements, e.g., menu and navigation elements (*anchor texts*), which do not belong to the blog comment content, have been deleted. Blog comments and their contextual metadata, such as the bloggers name and the date and time of publication, are stored in a database for further analysis.

4.2 Tokenization

Tokenization is a fundamental initializing step to divide the input text into coherent units (tokens),

in which each token is either a word or something else, e.g., a punctuation mark. The resulting tokens serve as basic information for designing the annotation scheme. Annotation labels must be determined according to all feasible tokens. Blog comments pose a special challenge to the task of tokenization due to the usage of non-standard language, including emoticons (e.g. :-), ;)) and multiple punctuation (e.g. ???, ??!). Such irregularities are not considered in standard tokenizers, provided with POS taggers, e.g. TreeTagger and Standford Tagger, which are developed on well-structured data such as newspaper texts. Hence, tokenization and POS tagging results based on blog comments suffer from high error rates (Neunerdt et al., 2012).

POS tagging is the initial level in multi-level annotation; therefore we develop a tokenizer particularly referring to this annotation level. A rule-based tokenizer is developed, which is adapted to the language in blog comments. By means of regular expressions text type-specific expressions, e.g., URLs, multiple punctuations and emoticons are detected as coherent tokens. Furthermore, the tokenizer treats text type-specific writing styles, like short forms, e.g., *geht's* “it works”, *gibts* “there's”, filenames, e.g. *test.jpg*, interaction words, e.g., *lol*, numbers, e.g., 3. November and so on. We design the tokenization rules with respect to the desired annotation scheme, e.g., *geht's* is separated into two tokens *geht* and 's. After successful tokenization the text can be annotated on all annotation levels.

4.3 POS Tagging

Initially, evaluation patterns emerge on the POS level, e.g., word orders of noun phrases [ADJA NN – *useful application*] or comparatives [ADJD KOKOM – *better than*]³. Therefore, blog comments of the selected corpus are automatically labeled with POS tags by means of the TreeTagger. Instead of the provided tokenizer, we use the developed blog comment tokenizer.

TreeTagger is a statistical tool for the annotation of text data with POS tags, according to the Stuttgart-Tübingen Tagset (STTS), and lemma information, according to a special lexicon

¹ This point requires special consideration to ensure sustainability and reusability of annotated data.

² While there is currently no firm legislation for the use and disclosure of Internet-based corpora, we can not share the corpus to the scientific community. We are working to solve this problem.

³ The speech tags are taken from the STTS-Tagset, <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>.

(Schmid, 1995; Schiller et al., 1999). In total, the tagset consists of 54 tags for part of speech, morpho-syntactic functions and non-speech. TreeTagger is trained on newspaper articles, which are grammatically well structured. Hence, the annotation of blog comments results in a high number of unknown words, thus annotation errors. Therefore we manually correct the annotation in a second step. Resulting data serve as training data for later development of an automatic POS tagger for blog comments.

Basically, syntactic information is given on the POS annotation level. Hence, text type-specific expressions such as emoticons and interaction words (*netspeak jargon*) as well as topic-specific terms such as URLs and file names are annotated according to their syntactical function, which has considerable advantages. First, there is no need to extend the existing STTS-tagset for blog comment annotation. Second, existing tools developed for texts with given STTS-annotation can still be applied.

For instance, emoticons are tagged according to their position as sentence-internal or sentence-final token. Much more difficult is the morpho-syntactic annotation of interaction words. Taken literally, interaction words are acronyms of multi-word expressions, e.g., *lol* for *Laughing out loud*. Thus, they cannot be classified as a part of speech. Rather, interaction words are similar to interjections, which are defined as single words or fixed phrases, which are invariable in their form and, moreover, they are seen as syntactically unconnected sentence-valent expressions. Table 1 shows which STTS-tags are assigned for exemplarily text type-specific expressions in blog comments.

Tag	Description	Example
\$.	Emoticons	:-(*)_* o.O
NE	File names, Internet address	test.jpg www.rwth-aachen.de
ITJ	Interaction words, inflectives	lol seufz
\$(Special characters	# * @ ^

Table 1: Morpho-syntactic annotation of text type-specific expressions.

As a result, we receive POS annotated blog comments with lemmas, which provide additional information for higher annotation levels.

4.4 Multi-level Annotation Model

In our understanding, an evaluation consists of different components, which are the *evaluated topic*, e.g., MCS, the *source of evaluation* (author or blogger), the *expressed evaluation* itself and the *textual context*, in which the evaluation is embedded. Therefore, in our model four types of levels are distinguished: *User-related levels*, *context-related levels*, *topic-related levels* and *blog comment-related levels*.

User-related levels. Blogger use in blogs typically nicknames such as *Andy2002* or *Triamos81*. The information contains suppositions about the gender of the blogger, his age as well as the accession date of his membership in the blog. The annotation of this data is most useful in identifying user profiles and user types due to the distribution by gender and age.

Context-related levels. For each blog comment, contextual metadata is supplied that are bloggers name, comment title, date and time of comment submission. The contextual metadata provides information about the user's blogging behavior, e.g., periods in which the user post comments vs. frequency of commentary.

Topic-related levels. According to the discussion topic, different terminologies are important. Particularly in the case of MCS, many topic-specific terms occur in blog comments, e.g., Sendemast "transmitter mast" vs. Nokia 808 PureView. These terms are typically not part of common lexicons respectively taggers are not trained on these terms which at worst leads to tagging errors. Thus, topic-specific terms must be detected in the corpus and classified (noun vs. proper noun, topic-specific vs. non-specific). The collected terms become lexical entries and are used for the development of the blog comment tagger. Moreover, the use and distribution of topic-specific terms in the corpus can also draw conclusions about topic preferences, e.g., system-related topics vs. device-related topics.

Blog comment-related levels. Annotations on these levels provide information about linguistic means, which form an evaluative expression. Actually, there are five different kinds of levels distinguished according to the grammatical fields: the graphemic level, the morphological level, the syntactic level, semantic level and the pragmatic level. At the *graphemic level*, expressions at the text surface as well as grapho-

TOK	Die	haben	es	doch	begriffen	,	die	liefern	einfach	immer	weniger	:)
TRANS	<i>They</i>	<i>have</i>	<i>it</i>	<i>yet</i>	<i>realized</i>	,	<i>they</i>	<i>provide</i>	<i>simply</i>	<i>increasingly</i>	<i>less</i>	:)
POS	PIS	VAFIN	PPER	ADV	VVPP	\$,	PDS	VVFIN	ADV	ADV	PIS	\$.
LEM	d	haben	es	doch	begreifen	,	d	liefern	einfach	immer	weniger	:)
TYPO												EMO
POL										^	%	
ESA												IRONIZE

Table 2: Example of a multi-level annotation: comment token (TOK) part of speech (POS), lemma (LEM), typography (TYPO), polarity (POL), evaluative speech act (ESA). The line translation (TRANS) is actually not part of the annotation system, but has been added here for reasons of comprehensibility.

stilistic features that show special notational styles, e.g., emoticons. The annotation at the *morphological level* focuses on word formation, inflection and formation of abbreviated words, e.g., topic-specific new word creations. The designs of the tag sets at both levels are inspired by Gimpel et al. (2011). At the *syntactical level*, syntactic structures are assigned as pointed out in Stoyanov and Cardie (2008). At the *semantic level*, semantic characteristics at the word (lexical semantics) and sentence level (sentence semantics) are recorded, e.g., polarities; tags are partially taken from Clematide et al. (2012). Finally at the *pramatic level*, information is given about the evaluative substance of a speech act, e.g., someone has the intention to BLAME, PRAISE, CRITICIZE something (Austin, 1972; Sandig, 2006).

Thus, each token and each evaluation-indicating token sequence is, in addition to the automatic annotation by the POS tagger, enriched with information regarding its various grammatical functions. Table 2 shows an example of a multi-level annotated text passage, which shows that annotations can cover one or more tokens, depending on the annotation level. Reading the annotations vertically allows for recognizing tag sequences of evaluative expressions in blog comments. These evaluation patterns can be useful for the purpose of automatic opinion detection.

5 Model Application

The aim of the model application was to validate the provided annotation model in terms of its reliability, exemplarily, and, to demonstrate its ability for pattern recognition. The annotation is performed in the score editor EXMARaLDA⁴.

Pragmatic	Semantic	Graphemic	Typography	1 st Level	2 nd Level	Tag	Description
				Polarity			
Evaluative speech act				AKR			interaction words
				EMO			emoticons
				ITER			iterations
				MAJ			capital letters
				MARK			highlighting
				MAT			maths symbols
				+			positive
				-			negative
				~			shifter
				^			intensifier
				%			diminisher
				ANGER			be upset about sth.
				BLAME			express disapproval
				CLAIM			accuse so., insist on sth.
				COMPARE			oppose sth. to each other
				CONCLUDE			sum up judgments
				CRITICIZE			judge by standards
				ESTIMATE			speak valuing about sth.
				IRONIZE			draw sth. ridiculous
				NEGATE			consider as non existant
				OVERSTATE			pose sth. overly positive
				PRAISE			express appreciation
				SUSPECT			raise concerns about sth.
				UNDERSTATE			pose sth. overly negative

Table 3: Excerpt of the level-specific tagsets.

5.1 Inter-annotator Agreement Study

To determine the reliability of the model, a two-stage inter-annotator agreement study is carried out. The test corpus contains comments of bloggers with 20 posts in the respective blog over two years. In total, the corpus comprises 50 comments and 5,362 token.

⁴ EXMARaLDA is a freely available tool and can be downloaded under <http://www.exmaralda.org/downloads.html>.

	Tag	Typography					Polarity					
		AKR	MARK	ITER	EMO	MAJ	MAT	+	-	~	^	%
EI	A1	23	31	20	5	6	3	298	172	120	31	10
	A2	22	34	2	4	4	4	295	185	124	50	23
	D	1	3	18	1	1	1	3	13	4	19	13
	%	4.31	8.8	90	20	33.3	23.1	0.99	7.4	2.9	37.9	56.5
	T	20.7					50.7					
EII	A1	29	27	3	5	3	3	190	80	121	24	35
	A2	41	29	4	5	3	3	215	99	111	25	36
	D	25	7	1	0	0	3	25	19	10	1	1
	%	29.1	6.5	23.1	0	0	0	23.7	19.4	8.3	1.4	2.9
	T	17.4					7.4					

Table 4: Results of inter-annotator agreement study: first evaluation (EI), second evaluation (EII), annotator 1 (A1), annotator 2 (A2), allocation difference (D) in percent (%), allocation difference in total/percent per evaluation (T).

The model is tested on three blog comment-related sub-levels, exemplarily: (i) grapho-stylistic means, (ii) polarity and (iii) evaluative speech act. The annotation process is guided by a stylebook that contains information about (i) the level-specific tagsets and (ii) the annotation guidelines. Before each of the evaluation stages, both annotators had to read the respective stylebook intensively and to ask questions, where appropriate. Requests during the annotation process were not allowed. The annotators worked on the same corpus using the identical tagset, separately. Table 3 shows an excerpt of the tag set⁵.

The inter-annotator agreement is calculated for each level manually. To precise, the evaluation focused two objectives: (i) analyzing the difference in the allocation of tags on the levels typography and polarity and (ii) identifying variations in the annotation scope on the level evaluative speech act. Table 4 shows the annotation differences for the selected MCS-corpus per level and per annotator (objective (i)).

As it is evident from the results of the first evaluation (EI), the error rates for tag allocation are relatively high. Particularly, the annotation of polarities appears to be problematic ($T=50.7\%$), especially for the annotation of diminishers (56.5%). At the level of typography, an enormous error rate (90%) for the allocation of iterations (ITER) is recognizable, particularly. However, the overall rating for tag allocation delivers a better result ($T=20.7\%$).

⁵ Due to page limitations, the entire tagset could not be presented.

Related to evaluation objective (ii), we have tested in how many cases the annotation scope (number of tokens that have been attributed to a tag) varies. Results show that at no text passage the annotation scope is identical for both annotators. Examples A1 (annotator 1) and A2 (annotator 2) show differences in the annotation of the evaluative speech act IRONIZE; the respective annotated text passages are marked in bold.

(A1) Pralerei oder **sind die Taschen zu klein?**

(A2) **Pralerei oder sind die Taschen zu klein?**

Thus, the provided annotation guidelines of the first evaluation (EI) seem to be too shallow. Comparing the annotated data of A1 and A2 for objective (ii) shows that annotation differences had to be diminished through guideline modifications. Therefore, for each tag it was re-defined

- i. with which *part of speech* it can appear, e.g., diminishers (%) occur only in combination with the POS tags PIAT (attributive indefinite pronoun) and ADV (adverb);
- ii. where an annotation starts and ends (*scope*), i.e., which feature terminates an annotation over multiple tokens, e.g., punctuation marks as terminating features.
- iii. whether *special characters* and interaction signs are annotated within an evaluative speech act, e.g., emoticons at the end of a sentence.

For testing the modified guidelines, a second evaluation (EII) was carried out (see table 4). The results show that a significant improvement is recognizable in the tag allocation (objective (i)), which is due to modification i. Nevertheless,

the allocation rate of several tags has not improved.

An excellent result, however, provides the comparison in terms of objective (ii). Compared to the results of the first evaluation (EI), the error rate of the second evaluation (EII) is 20.6%. The improved result is probably due to the modifications ii and iii.

5.2 Pattern Recognition Study

The aim of the study is to show that the model can be used for the recognition of evaluative patterns. For this purpose, a selected corpus is created. To ensure that different formulation and evaluation styles of bloggers are considered and represented (e.g. use vs. non-use of emoticons, use of specific emoticon types), the corpus is formed with comments from different user groups (selection criteria: average number of comments in the entire corpus). From each group, 50 comments are taken.

A single annotator semi-automatically annotated the corpus. First, the automatically tokenized and POS tagged data are checked and corrected, manually. Second, blog comment data is annotated also manually on the levels polarity, evaluative speech act and graphemic using the provided tagset. The investigation focuses on the identification of evaluative speech acts that occur in combination with emoticons and their representation by regular expressions. Results of the frequency analysis are summarized in Table 5.

Blogger / #Comments	Token	Emoticons			Σ
		+	-	0	
1	2,897	4	2	3	9
10	3,083	15	1	0	16
20	5,362	5	0	0	5
max	4,264	1	1	0	2
Σ	15,606	25	4	3	32

Table 5: Distribution of emoticons over user groups.

In the analyzed corpus, a total of 32 emoticons occur, of which 25 (78.125 %) were annotated as positive and 4 (12.5 %) as negative; 3 (9.4 %) got no polarity attribution. Overall, most commonly positive connoted emoticons were used for the marking of ironic speech acts (IRONIZE), e.g., ;-, :), ;), ,) and ^^. Negatively connoted emoticons, e.g., :(, :-(and :/-, never occurred in combination with an ironic speech act. Furthermore, in the user group with ten

comments per blogger, positive emoticons are used most often. The weighting is even more apparent, when the number of emoticons per user group is compared to the number of tokens per user group corpus (see table 5).

Regarding the morpho-syntactic function of emoticons, the result show that emoticons are in most cases used as sentence boundary signs (78.1 %), what means they are set instead of punctuation marks such as ! ?. Thereby, emoticons were set at the end of the blog comments in 44.1 % of the occurrence, i.e., behind the emoticon no more token followed. Then, on the POS level, the left neighbor of the emoticon is an internal character (\$), a noun (NN), an adverbial adjective (ADJD) or a substituting indefinite pronoun (PIS). Moreover, emoticons take over the function of internal characters such as – “ #.

Finally related to the selected user corpus, the following pattern is identified as stereotypical for the evaluative speech act IRONIZE:

[EMO₊[^ v ;) v :)] = \$.] + [\$. v Ø]

The term implies that three types of positive emoticons (^ v ;) :)) are typically used to mark the speech act IRONIZE. These emoticons usually take over the morpho-syntactic function of a sentence boundary sign. Habitually, a further sentence boundary sign follows the emoticon or no more token occurs.

6 Summary and Future Work

In this paper, we presented a fine-grained annotation model for the analysis of evaluative expressions in blog comments, exemplarily. The results of the evaluation studies show that the model is reliable. Moreover, we have demonstrated that the model serves for the identification of evaluation patterns. Future work will focus on the further improvement of the annotation guidelines and the identification of additional evaluation patterns, such as the presented pattern for ironic speech acts.

Acknowledgments

We owe gratitude to the Excellence Initiative of the German Federal and State Government as well as Eva Reimer, Julia Ninnemann, Tariq Maqbool and Simon Rüppel for their support in data annotation and tool modification.

References

- Saima Aman and Stan Szpakowicz. 2007. Identifying Expressions of Emotions in Text. In *Proceedings of TSD'07*. Berlin:196-205.
- John Langshaw Austin. 1972. *Zur Theorie der Sprachakte*. Stuttgart.
- Alexandra Balahur and Andrés Montoyo. 2010. Semantic Approaches to Fine and Coarse-Grained Feature-Based Opinion Mining. *Natural Language Processing and Information Systems*, (5723):142-153.
- Michael Beißwenger and Angelika Storrer. 2008. Corpora of Computer-Mediated Communication. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 1. de Gruyter, Berlin, New York:292-308.
- Michael Beißwenger. 2009. Chattern unter die Finger geschaut: Formulieren und Revidieren bei der schriftlichen Verbalisierung in synchroner internetbasierter Kommunikation. In Vilmos Ágel and Mathilde Hennig, editors, *Nähe und Distanz im Kontext variationslinguistischer Forschung*. de Gruyter, Berlin, New York.
- Michael Beißwenger. 2010. Empirische Untersuchungen zur Produktion von Chat-Beiträgen. In Tilmann Sutter and Alexander Mehler, editors, *Medienwandel als Wandel von Interaktionsformen*. Verlag für Sozialwissenschaften, Wiesbaden:47-81.
- Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012/in press. A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative* (TEI).
- Simon Clematide and others. 2012. MLSA — A Multi-layered Reference Corpus for German Sentiment Analysis. In *Proceedings of the 14th LREC*, Istanbul, Turkey.
- Stefanie Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of BXML'05*. Berlin:39-50.
- EXMARaLDA,
<http://www.exmaralda.org/downloads.html>, cited on 03.02.2012.
- Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4):384–392.
- Clayton R. Fink, Danielle S. Chou, Jonathon J. Kopcke, and Ashley J. Llorens. 2011. Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. *Johns Hopkins APL technocal Digest* 30(1):22-30.
- Eugenie Giesbrecht and Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop*. San Sebastian, Spain.
- Kevin Gimpel and others. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features and Experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Portland. ACL.
- Kristin Luckhardt. 2009. *Stilanalysen zur Chat-Kommunikation. Eine korpusgestützte Untersuchung am Beispiel eines medialen Chats*. Diss., TU Dortmund. Digitale Ressource: <http://hdl.handle.net/2003/26055>
- Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*, Birmingham:105-115.
- Anke Lüdeling. 2011. Corpora in Linguistics: Sampling and Annotation. In *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. Stockholm: 220-243.
- Malik M.S. Missen, Mohand Boughanem, and Guillaume Cabanac. 2009. Challenges for Sentence Level Opinion Detection in Blogs. In *Proceedings of ICIS'09*. Shanghai:347-351.
- Gilad Mishne. 2005. Experiments with Mood Classification in Blog Posts. In *Style2005: The 1st Workshop on Stylistic Analysis of Text for Information Access*. SIGIR '05. Salvador, Brasil.
- Melanie Neunerdt, Bianka Trevisan, Rudolf Mathar, and Eva-Maria Jakobs. 2012/in press. Detecting Irregularities in Blog Comment Language Affecting POS Tagging Accuracy. In *Proceedings of 13th CICLING*. Springer, New Delhi.
- Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a Publicly Available Germanlanguage Resource for Sentiment Analysis. In *Proceedings of the 7th LREC*. Malta:1168–1171.
- Robert Remus and Christian Häning. 2011. Towards Well-Grounded Phrase-Level Polarity Analysis. In *Proceedings of 11th CICLING*. Tokyo:380-392.
- Barbara Sandig. 2003. Formen des Bewertens. *Anabasis*. In *Festschrift für Krystyna Pisarkowa*, Lexis, Kraków: 279-287.
- Sandig, Barbara. 2006. *Textstilistik des Deutschen*, volume 2. de Gruyter, Berlin.
- Anne Schiller, Simone Teufel, and Christine Stöckert. 1999. *Guidelines für das Tagging deutscher Textkorpora mit STTS. Technischer Bericht*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart und Seminar für Sprachwissenschaft, Universität Tübingen.
- Peter Schlobinski and Torsten Siever. 2005. Sprachliche und textuelle Aspekte in Weblogs. Ein internationales Projekt. *Networx* (46).
- Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of SIGDAT'95*. Dublin. ACL.
- Angelika Storrer. 2009. Rhetorisch-stilistische Eigenschaften der Sprache des Internets. In Ulla Fix,

- Andreas Gardt, and Joachim Knape, editors, *Rhetorik und Stilistik – Rhetorics and Stilistics*. de Gruyter, Berlin, New York.
- Veselin Stoyanov and Claire Cardie. 2008. Annotating Topics of Opinions. *Proceedings of the 6th LREC*, Marrakech, Morocco.
- Carlo Strapparava and Rada Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the SAC'08*. Fortaleza, Brasil:1556-1560. ACM.
- Dirk Thorlechter, Dirk van den Poel, and Anita Prinzie. 2010. Extracting Consumers Needs for New Products - A Web Mining Approach. In *Proceedings of WKDD '10*. Phuket:440-443.
- TreeTagger <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>, cited on 01.02.2012.
- Bianka Trevisan and Eva-Maria Jakobs. 2010. Talking about Mobile Communication Systems. Verbal Comments in the Web as a Source for Acceptance Research in Large-scale Technologies. In *Proceedings of the IPCC 2010*. University of Twente (NL):93-100.
- Bianka Trevisan and Eva-Maria Jakobs. 2012. Probleme und Herausforderungen bei der Identifikation von Bewertungen in großen Textkorpora. Am Beispiel Mobilfunk. In Sabrina Braukmeier, Julia Burkhardt, and Fleur Pfeifer, editors, *Wege in den Sprachraum*. Lang, Frankfurt/Main:189-209.
- Janyce Wiebe, Teresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165-210.
- Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. University of Pittsburgh.

Robust processing of noisy web-collected data

Jelke Bloem

University of Groningen

j.bloem.3@student.rug.nl

Michaela Regneri

Saarland University

regneri@coli.uni-saarland.de

Stefan Thater

Saarland University

stth@coli.uni-saarland.de

Abstract

Crowdsourcing has become an important means for collecting linguistic data. However, the output of web-based experiments is often challenging in terms of spelling, grammar and out-of-dictionary words, and is therefore hard to process with standard NLP tools. Instead of the common practice of discarding data outliers that seem unsuitable for further processing, we introduce an approach that tunes NLP tools such that they can reliably clean and process noisy data collected for a narrow but unknown domain. We demonstrate this by modifying a spell-checker and building a coreference resolution tool to process data for paraphrasing and script learning, and we reach state-of-the-art performance where the original state-of-the-art tools fail.

1 Introduction

Web experiments in general and crowdsourcing in particular have become increasingly popular as sources for linguistic data and their annotations. For annotation purposes, crowdsourcing seems to deliver high-quality results (Snow et al., 2008) even for complex tasks like resolving anaphora (Chamberlain et al., 2009) or semantic relations (von Ahn et al., 2006). Collecting textual raw data via crowdsourcing seems very appealing at first sight, because it gives access to large amounts of commonsense knowledge which can usually not be extracted from text (Regneri et al., 2010; Singh et al., 2002).

However, processing the output of crowdsourcing or any web-based experiment (like corpora with emails, blog posts or Twitter messages) poses major challenges for natural language processing (NLP) tasks that take such data as their in-

put: The data usually contains a lot of spelling and grammar mistakes and many idiosyncratic words. Most domain-independent standard tools such as part of speech taggers, parsers or coreference resolution systems are not robust enough to handle such unpredictable idiosyncrasies. It is possible to use manual annotation instead of NLP tools to process web-collected datasets (Vertanen and Kristensson, 2011; Regneri et al., 2010), however, this does not scale well to larger experiments and is thus not always an option.

Many web-based approaches rely on collecting multiple instances for a narrow task or domain, and thus validate any text by comparing it to other examples from the same experiment and filtering them out if they seem too idiosyncratic (Law and von Ahn, 2009). We consider this approach too restrictive, because it may discard instances that are actually good but simply rare, or discard too much data when many instances are noisy. We instead make use of similar instances in crowd-sourced data at a later stage in the processing pipeline: We show that we can enhance the output of standard NLP systems rather than immediately discarding the original texts. Instead of restricting the input set to a much smaller, more homogeneous resource that is easier to process, we want to keep as much variety as possible in the data, by putting more knowledge into the actual systems. Such data parallelism can also be found in paraphrasing data or comparable corpora, but we are not aware of tool adaptation in this area.

We show how to implement that paradigm by example of two NLP tools, namely spell checking and coreference resolution. This paper is structured as follows: We first introduce a highly parallel dataset of web-collected event sequence descriptions for several common-sense tasks in the

kitchen domain (Sec. 2). The main focus of the remainder is the preprocessing of the data, using tools modified as described above, such that it can be put into a previously introduced paraphrasing pipeline. We explain and evaluate our modification of an open-source spell-checker (Sec. 3), and we demonstrate that a coreference resolution technique that solely relies on the parallel dataset can outperform a state-of-the-art system or a system trained on a general corpus (Sec. 4).

2 A highly parallel data set

Scripts (Schank and Abelson, 1977) describe a certain scenario (e.g. “eating in a restaurant”) with temporally ordered events (*the patron enters restaurant, he takes a seat, he reads the menu...*) and participants (*patron, waiter, food, menu,...*). Written event sequences for a scenario have been collected by Regneri et al. (2010). A similar collection is used by Rohrbach et al. (2012), with basic kitchen tasks from *Jamie Oliver’s Home Cooking Skills*¹. They collect natural language sequences using Mechanical Turk. For each scenario, subjects were asked to provide tutorial-like sequential instructions for executing the respective kitchen task. Overall, the corpus contains data for 52 cooking tasks, with 30-50 sequences per task. Fig. 1 shows three examples for the scenario *preparing garlic*.

Making Use of Redundancy

We mainly use the kitchen data set collected by Rohrbach et al. (2012). The data was gathered for script mining, which involves finding “event paraphrases” within the text, e.g. determine that *chop them finely* and *chip the garlic up until its small* denote the same event. In order to match event paraphrases with many variants in wording, spelling correction (for all kinds of processing) and pronoun resolution (to compare the phrase adjuncts) are essential (e.g. to match *chop them* with the misspelled phrase *chip the garlic up*).

Regneri et al. (2011) have already shown how to re-train a parser to gain robustness on the bullet-point style sentences by modifying big training corpora so as to replicate the nonstandard syntax. We want to show how the smaller data set

¹<http://www.jamieshomecookingskills.com/>

itself can serve to modify or filter standard applications in order to gain robustness and process the noisy data reliably. Given the example in Fig. 1, it is clear that big parts of the content word inventory (*cloves, skin, garlic*) are shared between the three sequences. Under this prerequisite, we propose two different application modifications for preprocessing:

- 1. Spelling Correction:** We use the data to amend the lexicon of a standard spell-checker. The kitchen data set contains many domain specific words, e.g. *microplane grater*, which the standard spell-checker corrects to *micro plane grater*. We also use the vocabulary of the input texts to rank correction proposals of actual misspellings.
- 2. Coreference Resolution:** Coreference resolution is a hard unsolved problem, even for standard text. Our fragmentary discourses that consist of incomplete sentences were not processable by any state-of-the-art tool: On top of the problematic syntax, the systems fail because the most salient referent is sometimes not even present in the text itself. E.g. under the headline *Preparing garlic*, one can give a complete instruction without using the word *garlic* by referring to it with pronouns only. Given the very restricted set of possible noun antecedents in our data, selectional preferences can be used to fully resolve most pronominal references.

3 Spelling correction task

Orthographically correct text is much more reliable as input for standard applications or dictionaries, thus we need to spell-check the noisy web texts. We use the widely used spelling correction system GNU Aspell². Aspell’s corrections are purely based on the edit- and phonological distance between suggestions and the misspelled token, and many kitchen-domain specific words are not in Aspell’s dictionary. This leads to implausible choices that could easily be avoided by considering the domain context — due to the parallel nature of our data set, correct words tend to occur in more than one sequence. The first objective is to gain more precision for the *identification*

²<http://www.aspell.net/>

1. first strip of the papery skin of the bulb	1. peel the skin of a clove of garlic	1. remove the papery skin of the garlic clove
2. ease out as many intact cloves as possible	2. cut the ends of the clove	2. mash the glove into pulp
3. chop them finely if you want a stronger taste	3. chip the garlic up until its small	3. this can be done using a ziploc bag and a meat tenderizer
4. chop them coarsely if you want a weaker taste	4. use in your favorite dish	4. use the garlic pulp in all sorts of meals
5. crushed garlic is the strongest taste		

Figure 1: Three example sequences for the scenario *preparing garlic*

of misspellings. We indirectly expand Aspell’s dictionary: If a word occurs in at least n other sequences in the same spelling, the system accepts it as correct. (In our experiment, $n = 3$ turned out to be a reliable value.) As a second enhancement, we improve *correction* by guiding the selection of corrected words. We verify their plausibility by taking the first suggested word that occurs in at least one other sequence. This excludes out-of-domain words that are accidentally too similar to the misspelled word compared to the correct candidate. Similar work on spelling correction has been done by Schierle et al. (2008), who trained their system on a corpus from the same domain as their source text to gain context. We take this idea a step further by training on the noisy (but redundant) dataset itself, which makes definition and expansion of the domain superfluous.

Evaluation

We compare our system’s performance to a baseline of standard Aspell, always picking the first correction suggestion. Since error detection is handled by standard Aspell in both cases, we have not evaluated recall performance. Previous work with manual spelling correction showed low recall (Regneri et al., 2010), so Aspell-based checking is unlikely to perform worse. Fig. 3 shows a comparison of both methods using various measures of precision, based on manual judgement of the corrections made by the checkers. Since Aspell’s spelling error detection is not perfect, some of the detected errors were not actually errors, as shown in the manually counted “False Positives” column. For this reason, we also included “true precision”, which is calculated only over actual spelling errors. Another measure relevant for the script learning task is “semantic precision”, a more loosely defined precision measure in which any correction that results in any inflection of the desired lemma is considered correct, ignoring grammaticality. The

numbers show that we gain 15% overall precision, and even 22% by considering genuine misspellings only. If we relax the measure and take every correct (and thus processable) lemma form as correct, we gain an overall precision of 18%.

4 Pronoun resolution task

The pronoun resolution task involves finding pronouns, looking for candidate antecedents that occur before the pronoun in the text, and then selecting the best one (Mitkov, 1999). While our dataset adds complications for standard pronoun resolution systems, the domain simplifies the task as well. Due to the bullet point writing style, first person and second person pronouns always refer to the subject performing a task. This leaves only third person pronouns, of which possessives are uninteresting for the script learning task — with just one person active in the dataset, there should be no ambiguity attributable to personal possessive pronouns (Regneri et al., 2011). There is also a relatively restricted space of possible antecedents due to the short texts.

Our system uses a POS-tagged and dependency-parsed version of the dataset, created using the Stanford parser (De Marneffe et al., 2006). The last few noun phrases before the pronoun within the same sequence are considered candidate antecedents, as well as the main theme of the descriptive scenario title. It then uses a model of selectional preference (Clark and Weir, 2001) to choose the most likely antecedent, based on verb associations found in the rest of the data. For example, in *Grab a knife and slice it*, the fact that *it* is the object of *slice* can be used to estimate that *knife* is an unlikely antecedent, given that such a verb-object combination does not occur elsewhere in the data.

Our approach only includes the noisy dataset in its selectional preference model, rather than some external corpus. For the phrase *chop it*, the statist-

Method	Precision	False Positives	True Precision	Sem. Precision	Corrections
Aspell	0.43	0.28	0.57	0.58	162
<i>Enhanced Aspell</i>	0.58	0.29	0.79	0.76	150

Figure 2: Evaluation of the baseline and improved spell-checker on 10 scenarios (25.000 words).

ical association strength of candidate antecedents with the head *chop* is checked, as computed on our dataset. The candidate that is most strongly associated with *chop* is then proposed as the preferred antecedent. Association values are computed with the marginal odds ratio measure on a dictionary of all head-subject and head-object relations occurring in our data. While this model is computationally much more simple, it takes advantage of our parallel data.

Evaluation

We compare the models to two baselines: One is a state-of-the-art general, openly available pronoun resolver based on Expectation Maximization (Charniak and Elsner, 2009). This system relies on grammatical features to model pronouns. Therefore it fails to model the pronouns in our parsed noisy data in many cases (recall 0.262). For the other simpler models, it is trivial to achieve full recall on 3rd person pronouns in our parsed sequences, so we won't further evaluate recall. We instead evaluate the models in terms of correctness compared to human annotation. As a second baseline, we use a state-of-the-art vector space model (Thater et al., 2011) to compute selectional preferences instead of tuning the preferences on our dataset. For each word or phrase, this model computes a vector representing its meaning, based on a large general corpus. For each candidate antecedent of phrases containing a pronoun like *chop it*, we compute the model's vector for the original phrase in which the pronoun is instantiated with the antecedent (*chop garlic, chop fingers, ...*). The candidate vector that is closest to the vector of the original verb, computed as the scalar product, is taken as the preferred antecedent.

The results show that our approach outperforms both baselines, despite its simpler heuristics. Compared to the vector space model approach we observe a 16% correctness gain, with

Model	Correct
Vector space model	0.544
EM	0.175
<i>Odds ratio</i>	0.631

Figure 3: Evaluation of different selectional preference models for pronoun resolution, on two scenarios (103 pronouns total).

a much larger gain over the EM system due to its recall issue.

5 Conclusion and Future work

We have demonstrated a new approach to processing the generally noisy output of crowdsourcing experiments in a scalable way. By tuning NLP tools on the dataset they will process, we have shown performance gains on two NLP tasks - spell-checking and coreference resolution. Despite using relatively simple methods and heuristics compared to the state of the art, the parallel nature of our dataset allowed us to achieve state-of-the-art performance. Our approach to processing preserves more crowdsourced information than the common practise of discarding outlier data. Results could be improved further by refining the methods, e.g. weighting candidate antecedents by their distance to the pronoun in the pronoun resolution tool, or improving detection of common spelling mistakes by varying how easily unknown words are accepted based on edit distance. In future research, applying this approach to other datasets and other processing tasks would be interesting, as well as extending the process to languages that lack state-of-the-art NLP tools.

Acknowledgements

We thank Manfred Pinkal, Gosse Bouma and the anonymous reviewers for their helpful comments on the paper.

References

- Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2009. A demonstration of human computation using the phrase detectives annotation game. In *KDD Workshop on Human Computation*.
- E. Charniak and M. Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of EACL*, pages 148–156. Association for Computational Linguistics.
- S. Clark and D. Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.
- M.C. De Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.
- Edith Law and Luis von Ahn. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI ’09, pages 1197–1206, New York, NY, USA. ACM.
- R. Mitkov. 1999. Anaphora resolution: The state of the art. *Unpublished Manuscript*.
- M. Regneri, A. Koller, and M. Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of ACL-10*.
- Michaela Regneri, Alexander Koller, Josef Ruppenhofer, and Manfred Pinkal. 2011. Learning script participants from unlabeled data. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.
- Marcus Rohrbach, Michaela Regneri, Micha Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *Computer Vision - ECCV 2012 : 12th European Conference on Computer Vision*, volume 2012 of *Lecture Notes in Computer Science*, Firenze, Italy, October. Springer, Springer.
- Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.
- M. Schierle, S. Schulz, and M. Ackermann. 2008. From spelling correction to text cleaning—using context information. *Data Analysis, Machine Learning and Applications*, pages 397–404.
- Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan L. Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems - DOA, CoopIS and ODBASE 2002*, London, UK. Springer-Verlag.
- R. Snow, B. O’Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263. Association for Computational Linguistics.
- Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proc. of IJCNLP 2011*.
- K. Vertanen and P.O. Kristensson. 2011. The imagination of crowds: conversational aac language modeling using crowdsourcing and large data sources. In *Proceedings of EMNLP*.
- Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *CHI ’06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, New York, NY, USA. ACM.

Navigating Sense-Aligned Lexical-Semantic Resources: THE WEB INTERFACE TO UBY

Iryna Gurevych^{1,2}, Michael Matuschek¹, Tri-Duc Nghiem¹,
Judith Eckle-Kohler¹, Silvana Hartmann¹, Christian M. Meyer¹

¹Ubiquitous Knowledge Processing Lab (UKP-TUDA)

Department of Computer Science, Technische Universität Darmstadt

²Ubiquitous Knowledge Processing Lab (UKP-DIPF)

German Institute for Educational Research and Educational Information

<http://www.ukp.tu-darmstadt.de>

Abstract

In this paper, we present the Web interface to UBY, a large-scale lexical resource based on the Lexical Markup Framework (LMF). UBY contains interoperable versions of nine resources in two languages. The interface allows to conveniently examine and navigate the encoded information in UBY across resource boundaries. Its main contributions are twofold: 1) The visual view allows to examine the sense clusters for a lemma induced by alignments between different resources at the level of word senses. 2) The textual view uniformly presents senses from different resources in detail and offers the possibility to directly compare them in a parallel view. The Web interface is freely available on our website¹.

1 Introduction

Lexical-semantic resources (LSRs) are the foundation of many Natural Language Processing (NLP) tasks. Recently, the limited coverage of LSRs has led to a number of independent efforts to align existing LSRs at the word sense level.

However, it is very inconvenient to explore the resulting sense-aligned LSRs, because there are no APIs or user interfaces (UIs) and the data formats are heterogeneous. Yet, easy access to sense-aligned LSRs would be crucial for their acceptance and use in NLP, as researchers face the problem of determining the added value of sense-aligned LSRs for particular tasks.

In this paper, we address these issues by presenting UBY-UI, an easy-to-use Web-based UI

to the large sense-aligned LSR UBY (Gurevych et al., 2012). UBY is represented in compliance with the ISO standard LMF (Francopoulo et al., 2006) and currently contains interoperable versions of nine heterogeneous LSRs in two languages, as well as pairwise sense alignments for a subset of them: English WordNet (WN), Wiktionary (WKT-en), Wikipedia (WP-en), FrameNet (FN), and VerbNet (VN); German Wiktionary (WKT-de), Wikipedia (WP-de), and GermaNet (GN), and the English and German entries of OmegaWiki (OW-en/de).

The novel aspects of our interface can be summarized as 1) *A graph-based visualization of sense alignments between the LSRs integrated in UBY*. Different senses of the same lemma which are aligned across LSRs are grouped. This allows intuitively exploring and assessing the individual senses across resource boundaries. 2) *A textual view for uniformly examining lexical information in detail*. For a given lemma, all senses available in UBY can be retrieved and the information attached to them can be inspected in detail. Additionally, this view offers to compare any two senses in a detailed contrasting view.

2 Related Work

Single Resource Interfaces. Web interfaces have been traditionally used for electronic dictionaries, such as the Oxford Dictionary of English. Lew (2011) reviews the interfaces of the most prominent English dictionaries. These interfaces have also largely influenced the development of Web interfaces for LSRs, such as the ones for WN, FN, WKT, or the recently presented DANTE (Kil-

¹<https://uby.ukp.informatik.tu-darmstadt.de>

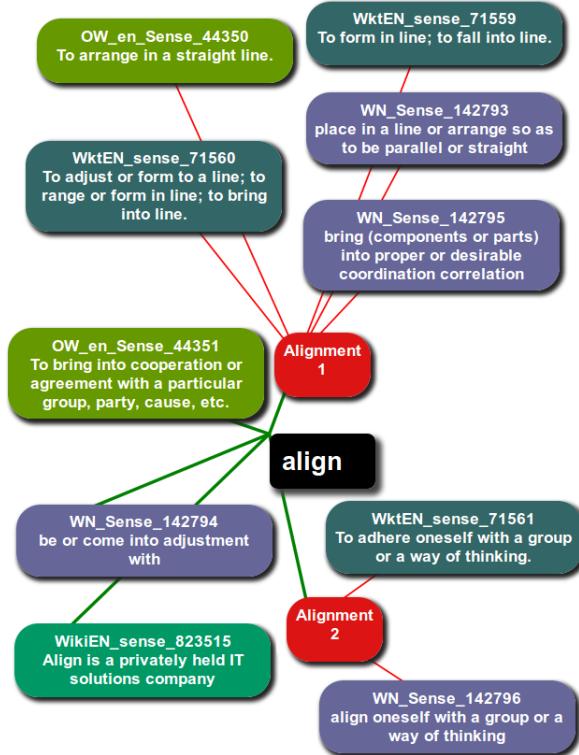


Figure 1: Search result for the verb *align* in the visual view. The aligned senses are connected by sense alignment nodes. Nodes are coloured by resource.

gariff, 2010) which directly adapted the dictionary interface models. All of these Web interfaces have been designed in strict adherence to a specific, single LSR. The UBY-UI is, in contrast, designed for multiple heterogeneous LSRs.

Multi Resource Interfaces. Only a few other Web interfaces are able to display information from multiple LSRs. The majority of them is limited to show preformatted lexical entries one after another without interconnecting them. Popular examples are Dictionary.com² and TheFreeDictionary³. Similarly, the DWDS interface (Klein and Geyken, 2010) displays its entries in small rearrangeable boxes. The Wörterbuchnetz (Burch and Rapp, 2007) is an example of a Web interface that connects its entries by hyperlinks – however, only at the level of lemmas and not word senses.

In contrast, UBY-UI provides hyperlinks to navigate between different *word senses*, as UBY provides mono- and cross-lingual sense align-

²<http://www.dictionary.com>

³<http://www.thefreedictionary.com>

ments between its LSRs. Additionally, UBY-UI supports the direct comparison of two arbitrary word senses present in UBY. In the other multi resource interfaces, this is only possible for whole lexical entries.

Graph-based Interfaces. Two examples for visualizing WN are Visuwords⁴ and WordNet explorer⁵ that allow browsing the WN synset structure. An example for a cross-lingual graph-based interface is VisualThesaurus⁶ which shows related words in six different languages. UBY-UI provides a similar graph-based interface, but combines the information from multiple types of LSRs interlinked by means of sense alignments.

3 UBY – A Sense-Aligned LSR

LMF Model. The large-scale multilingual resource UBY holds standardized and hence interoperable versions of the nine LSRs previously listed. UBY is represented according to the UBY-LMF lexicon model (Eckle-Kohler et al., 2012), an instantiation and extension of the meta lexicon model defined by LMF. Developing a lexicon model such as UBY-LMF involves first selecting appropriate classes (e.g. LexicalEntry) from the LMF packages, second defining attributes for these classes (e.g. part of speech), and third linking the attributes and other linguistic terms (such as attribute values) to ISOCat.⁷ UBY-LMF is capable of representing a wide range of information types from heterogeneous LSRs, including both expert-constructed resources and collaboratively constructed resources. Representing them according to the class structure of UBY-LMF makes them structurally interoperable. The linking of linguistic terms with their meaning as defined in ISOCat contributes to semantic interoperability.

Sense Alignments. UBY-LMF models a LexicalResource as consisting of one Lexicon per integrated resource. These Lexicon instances can be aligned at the sense level by linking pairs of senses or synsets using

⁴<http://www.visuwords.com>

⁵<http://faculty.uoit.ca/collins/research/wnVis.html>

⁶<http://www.visualthesaurus.com>

⁷<http://www.isocat.org/>, the implementation of the ISO 12620 Data Category Registry (Broeder et al., 2010).

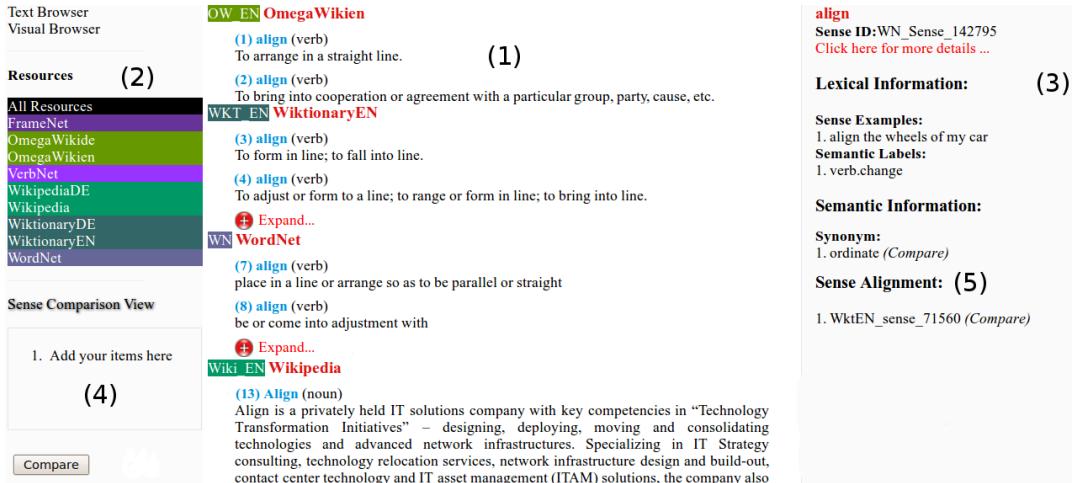


Figure 2: The textual view: (1) Senses of *align*, grouped by resource. (2) Area for selecting resources. (3) Detail view for a selected sense. (4) Drag & drop area for sense comparison. (5) Links to other senses.

instances of the `SenseAxis` class. The resource UBY features pairwise sense alignments between a subset of LSRs. Both monolingual and cross-lingual alignments are present in UBY: WN–WP-en (Niemann and Gurevych, 2011), WN–WKT-en (Meyer and Gurevych, 2011), VN–WN (Kipper et al., 2006), VN–FN (Palmer, 2009), OW–de–WN (Gurevych et al., 2012) and OW–WP, OW–en–OW–de, WP–en–WP–de which are part of the original resources.

The WN–WP-en, WN–WKT-en and OW–de–WN alignments have been automatically created. Please refer to the papers mentioned above for details on the alignment algorithm and detailed statistics.

UBY 1.0 UBY currently contains more than 4.5 million lexical entries, 4.9 million senses, 5.4 million semantic relations between senses and more than 700,000 alignments between senses. There are 890,000 unique German and 3.1 million unique English lemma-POS combinations⁸.

4 UBY Web Interface

Technical Basis. UBY is deployed in an SQL database via hibernate, which is also the foundation of the UBY-API. This allows to easily query all information entities within UBY. More details on the UBY-API can be found on the UBY

⁸Note that for homonyms there may be more than one LexicalEntry for a lemma-POS combination.

website⁹. The frontend of the Web application is based on Apache Wicket¹⁰.

Visual View. The natural entry point to the visual view is the search box for a lemma¹¹, and the result is a graph, with the query lemma as the central node and the retrieved senses as nodes attached to it (see figure 1). The sense nodes are coloured according to the source LSRs. To keep the view compact, the definition is only shown when a node is clicked.

The sense alignments between LSRs available in UBY are represented by *alignment nodes*, which are displayed as hubs connecting aligned senses. For generating the alignment nodes, we cluster senses based on their pairwise alignments and include all senses which are directly or transitively aligned. Thus, the visual view provides a visualization of which and how many senses from different LSRs are aligned in UBY. In Figure 1, we show the grouping of senses for the verb *align*. If a user wants to inspect a specific sense in more detail, a click on the link within a sense node opens the textual view described below.

Textual View. While the query mechanism for the textual view is the same as for the visual view, in this case the interface returns a list of senses (see (1) in Figure 2), including definitions, available for this lemma either in all LSRs, or only

⁹<http://www.ukp.tu-darmstadt.de/uby>

¹⁰<http://wicket.apache.org/>

¹¹Filtering by POS is to be included in a future release.

align (verb) WN_Sense_142793		align (verb) WktEN_sense_71560	
place in a line or arrange so as to be parallel or straight (1)	(1)	To adjust or form to a line; to range or form in line; to bring into line.	
Lexical Information: (2)		Lexical Information:	
Sense Examples: 1. align the car with the curb 2. align the sheets of paper on the table		Semantic Labels: 1. transitive	
Semantic Labels: 1. verb.change			
Semantic Information: (3)		Sense Alignment: (4)	
Synonym: 1. aline (<i>Compare</i>) 2. adjust (<i>Compare</i>) 3. line up (<i>Compare</i>)		1. WN_Sense_130408 (<i>Compare</i>) 2. WN_Sense_142793 (<i>Compare</i>) 3. WN_Sense_47418 (<i>Compare</i>) 4. WN_Sense_74413 (<i>Compare</i>) 5. WN_Sense_140939 (<i>Compare</i>) 6. WN_Sense_142795 (<i>Compare</i>) 7. WN_Sense_49861 (<i>Compare</i>)	
Sense Relation: 1. Relation Name: antonym Destination Sense: WN_Sense_109405			
Sense Alignment: (4)			
1. WktEN_sense_71560 (<i>Compare</i>)			

Figure 3: In the sense comparison view, detailed information for two arbitrary senses can be inspected. Below the definition for each sense (1), lexical (2) and semantic (3) information is listed if available. Note the alignment sections (4) which contain links to the aligned senses, as well as links to compare two senses immediately.

those selected by the user (2). Additionally, the LSRs are colour-coded like in the visual view.

For further exploring the information attached to a single sense, clicking on it opens an expanded view on the right-hand side (3) showing more detailed information (e.g. sense examples). Optionally, a full screen view can be opened which allows the user to explore even more information. In the detailed view of a sense, it is also possible to navigate to other senses by following the hyperlinks, e.g. for following sense alignments across LSRs (5).

For comparing the information attached to two senses in parallel, we integrated the option to open a comparison view. For this, the user can directly drag and drop two senses to a designated area of the UI to compare them (4), or click the *Compare* link in the sense detail view (5).

The advantage of the comparison view is illustrated in Figure 3: As the information is presented in a uniform way (due to the standard-compliant representation of UBY), a user can easily compare the information available from different LSRs without having to use different tools, terminologies, and UIs. In particular, for senses that are aligned across LSRs, the user can immediately detect complementary information, e.g., if a WKT sense does not have sense examples but the aligned WN sense does, this additional information becomes directly accessible. To our knowledge, such a contrasting view of two *word senses* has not been offered by any resource

or UI so far.

5 Conclusions and Future Work

In this paper, we presented a Web interface to UBY, a large-scale sense-aligned LSR integrating knowledge from heterogeneous sources in two languages. The interface combines a novel, intuitively understandable graph view for sense clusters with a textual browser that allows to explore the offered information in greater detail, with the option of comparing senses from different resources.

In future work, we plan to extend the UI to allow editing of the alignment information by the users. The rationale behind this are the errors resulting from automatic alignment. A convenient editing interface will thus help to improve the underlying resource UBY. Another goal is to enhance the display of alignments across multiple resources. Right now, we use pairwise alignments between resources to create sense clusters, but as we plan to add more sense alignments to UBY in the future, the appropriate resolution of invalid alignments will become necessary.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Richard Eckart de Castilho and Zijad Maksuti for their contributions to this work.

References

- Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta.
- Thomas Burch and Andrea Rapp. 2007. Das Wörterbuch-Netz: Verfahren - Methoden - Perspektiven. In *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung hist 2006*, Historisches Forum 10, Teilband I, pages 607–627. Berlin: Humboldt-Universität zu Berlin.
- Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 275–282, Istanbul, Turkey.
- Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 580–590.
- Adam Kilgarriff. 2010. A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of the NAACL-HLT 2010 Demonstration Session*, pages 21–24, Los Angeles, CA, USA.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1027–1032, Genoa, Italy.
- Wolfgang Klein and Alexander Geyken. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica*, 26:79–96.
- Robert Lew. 2011. Online dictionaries of English. In Pedro A. Fuertes-Olivera and Henning Bergenholz, editors, *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, pages 230–250. London/New York: Continuum.
- Christian M. Meyer and Iryna Gurevych. 2011. What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892, Chiang Mai, Thailand.
- Elisabeth Niemann and Iryna Gurevych. 2011. The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 205–214, Oxford, UK.
- Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex)*, pages 9–15, Pisa, Italy.

Using Information Retrieval Technology for a Corpus Analysis Platform

Carsten Schnöber

Institut für Deutsche Sprache

R5 6-13

D-68161 Mannheim

Germany

schnoer@ids-mannheim.de

Abstract

This paper describes a practical approach to use the information retrieval engine Lucene for the corpus analysis platform KorAP, currently being developed at the *Institut für Deutsche Sprache* (IDS Mannheim). It presents a method to use Lucene's indexing technique and to exploit it for linguistically annotated data, allowing full flexibility to handle multiple annotation layers. It uses multiple indexes and MapReduce techniques in order to keep KorAP scalable.

1 Introduction

In this paper, the Lucene information retrieval (IR) framework¹ is taken out of its native compound, and ported to the field of corpus linguistics, to investigate its applicability as an engine for KorAP² (*Korpusanalyseplattform der nächsten Generation*, “Next Generation Corpus Analysis Platform”) that is conducted at the *Institut für Deutsche Sprache* (IDS Mannheim). The aim of this project is to develop a modern, state-of-the-art corpus-analysis platform, capable of handling very large corpora and opening the perspectives for innovative linguistic research (Bański et al., 2012).

The KorAP engine is designed to be scalable and flexible enough to store and analyse the fast-growing amounts of linguistic resources that have become available to corpus linguists. The DeReKo corpus (*Deutsches Referenzkorpus*,

“German Reference Corpus”) serves as an immediate use case, being the largest German text corpus in the world with 5.4 billion words and three concurrent annotation layers (IDS, 2011), currently accessible through Cosmas II (Bodmer, 2005). The corpus is expected to keep growing as rapidly as it has in the past two decades of its existence in which DeReKo at least doubled its size every five years³. Although the actual growth cannot be extrapolated from such figures, KorAP aims to be ready for corpora with 50 billion tokens. However, the KorAP platform is not designed exclusively for DeReKo but to process any corpora, independent of size, writing system, language or other specific characteristics.

1.1 Linguistic Research vs. Information Retrieval

In IR, the primary goal is to find documents containing the information the user needs. In text documents, the surface text is the vehicle that provides and often hides this information, possibly expressed in many different ways. From an IR point of view, “language is an obstacle on the way to resolving a problem” while in corpus linguistics, the language is the research object (Perkuhn et al., 2012, p. 19). A linguistic search may query complex data structures and relationships, such as multiple metrics and levels while the user demands can pose challenging inquiries (e.g. relations, quantifiers, regular expressions (Bański et al., 2012).

¹Lucene homepage: <http://lucene.apache.org>

²KorAP: <http://korap.ids-mannheim.de>

³DeReKo outlines (in German): <http://www.ids-mannheim.de/k1/projekte/korpora/archiv.html#Umfang>

An ideal IR engine would thus be able to fully interpret language and to abstract its meaning. For that purpose, IR techniques typically aim to remove elements that do not bear much meaning, e.g. function words and morphological affixes. Linguistic researchers, on the other hand, are typically interested in finding linguistic phenomena of all kinds, often involving tokens that seem semantically less relevant.

1.2 Lucene

Lucene is an open-source framework for searching large amounts of text and is considered to be the most widely used information retrieval library (McCandless et al., 2010, p. 3). It essentially provides a software library for creating inverted indexes (Section 3).

The Lucene project also provides a ready-to-run search application, Solr, that implements text search with typical information retrieval functionality. However, Solr is neither designed for a linguistic application like KorAP nor easily adaptable for that purpose.

1.3 KorAP Outlines

The KorAP project aims to build a generic solution to numerous problems that have arisen with the increasing size of corpora that are subject to linguistic research. An essential specification for the KorAP engine is that it must comply with scientific requirements and therefore, its results must be falsifiable and traceable, making KorAP a scientific tool while rendering insufficient substitutes like Google Search unnecessary; cf. Kilgarriff (2007).

On the other hand, KorAP does not want to re-invent the wheel and there are numerous projects tackling similar problems, approaching from both the linguistic and the computational side. In that respect, one part of KorAP development is to investigate existing solutions and finding trade-offs between re-use and new development. Other projects taken into account for (partial) adaptation for KorAP include Annis (Zeldes et al., 2009), DDC/DWDS⁴ (Sokirko, 2003), and Poliqarp (Janus and Przepiórkowski, 2007).

⁴Das Wörterbuch der Deutschen Sprache (“The German Language Dictionary”): <http://retro.dwds.de/>

Another central goal for KorAP is to reach a level of generalisation that makes the platform flexible enough to process any kind of data, not only text but also multimodal resources like recorded and transcribed speech; the DGD speech corpora (*Datenbank Gesprochenes Deutsch*, “German Speech Database”) (Fiehler and Wagener, 2005) will serve as a use case. Neither segmentations nor annotations are constrained by the KorAP engine and it should be flexible enough to be prepared for annotation types that might not even be thought of today.

That implies that KorAP does not predefine the scale of a token; instead, any sequence of characters can be defined as the atomic building stones to which other elements can refer. KorAP reads token boundaries from external files and builds an index based on these, but does not tokenize on its own (see Section 3). Annotation tools can freely assign tags or labels to character sequences without being limited by a pre-defined tokenization algorithm or another segmentation logic.

In order to allow adding annotations dynamically at any point and independently of existing indexes, KorAP uses standoff annotations (Thompson and McKelvie, 1997) to allow a clear separation amongst annotations and between annotations and the primary text. This allows for multi-level annotations of different types, including discontinuous spans and structures with internal references like trees and dependency grammars.

2 Previous Work

Lucene, as a fast and well-established text indexing engine, has been applied for linguistic research several times in the past years. Despite the differences between IR and corpus linguistics, there is a close relation. “The ability to interrogate large collections of parsed text [...] opens the way to a new kind of information retrieval (IR) that is sensitive to syntactic information, permitting users to do more focussed search” (Ghodke and Bird, 2010).

Lucene competes with relational and XML database systems for solving the problem of efficiently querying linguistically annotated texts. “Many existing systems load the entire corpus into memory and check a user-supplied query

against every tree. Others avoid the memory limitation, and use relational or XML database systems. Although these have built-in support for indexes, they do not scale up either” (Ghodke and Bird, 2010; Ghodke and Bird, 2008; Zhang et al., 2001). However, there do exist approaches in which relational database management systems (RDBMS) are applied in large scale corpus indexing applications (Schneider, 2012).

The approach presented by Ghodke and Bird (2010) is not directly transferable to KorAP because it does not focus on the same abstraction level. While KorAP strictly uses standoff annotations, Ghodke and Bird (2010) store annotations inline, attached to the primary text. Also, they create Lucene documents on the base of single sentences, which makes it elegantly easy to search for multiple words occurring in a single sentence and reduces the search space radically, but at the same time gives up information about how sentences are related to each other.

Even though this approach is certainly useful for many real-world use cases in corpus-driven linguistic research, assuming the sentence to be a naturally self-sufficient linguistic unit, this forms a restriction that, once accepted, could not be abandoned when more flexibility is needed. For instance, a query might search for co-references or repetitive utterances in subsequent sentences; especially in discourse analysis, the sentence is not always sufficient as highest research unit; cf. Volk (2002) and Sinclair (2004).

3 Inverted Indexes

Inverted indexes (Brin and Page, 1998; Knuth, 1997, Chapter 6.5) are a technique to provide fast querying for large text documents. A naïve algorithm would iterate over the full text to find occurrences of a search term; an inverted index reduces the search space by creating a dictionary that lists all the distinct terms contained in a collection of documents and all the occurrences of each term: “This is the inverse of the natural relationship, in which documents list terms” (Lucene, 2012).

In order to create an inverted index, a precise definition of a term has to be specified, corresponding to a token without implying any further (linguistic) meaning. Such tokens are the minimum units to work with because an inverted index

provides direct access only to those terms that are listed in its term dictionary.

Lucene stores the term dictionary in a file that lists the distinct tokens in alphabetic order. Apart from saving disk space, this allows using common prefixes to perform Wildcard searches, for instance, efficiently. In order to make random access to the dictionary fast, an additional term info index file stores a copy of every n -th entry from the term dictionary with a delta value that defines “the difference between the position of this term’s entry [...] and the position of the previous term’s entry”. The term index is designed to reside entirely in memory (Lucene, 2012).

A Lucene index can be segmented into multiple parts; they are listed in a dedicated file in the index directory, but fully independent of each other. A new segment is added during indexing when the previous one’s size reaches a configurable threshold. New segments can be added at any later point, for instance when new documents are indexed, and multiple segments can be merged.

4 Linguistic Research with Lucene

4.1 A Lucene Analyzer for KorAP Data

The Lucene indexing process happens in three phases: read the text, analyse it, build an index. The first step may simply be reading plain text files, but can as well include text extraction from various formats like XML files, PDFs or other document types. In the second phase, the text is split into tokens that form the key terms in the inverted index.

In IR, filters are typically applied in the text analysis phase that aim to reduce the number of distinct tokens, for instance by stopword filtering and stemming. However, this is not obligatory and the analysis process is fully customisable. Lucene provides various built-in analyzers, but the choice can be broadened by own developments. After the text has been processed, the index is finally built on the basis of the tokens produced during analysis.

KorAP corpora are stored in XML files⁵ where each document comprises one directory, each

⁵See <http://korap.ids-mannheim.de/2012/03/data-set-released/> for a sample KorAP XML data set.

containing a file for the text (Figure 1) and one for the metadata (Figure 2). The annotations are organised in so-called foundries, realised as dedicated directories in which one or multiple annotation levels congregate. “We define a foundry as a collection of annotation layers that have something in common: they may have been simply produced by the same tool, or at least they elaborate on the same theoretical view of the data (in the case of foundries containing hierarchical annotation layers building upon one another)” (Bański et al., 2011). KorAP provides two tokenization layers as part of the base foundry for every document – one with a rather aggressive splitting approach (‘greedy’) and one that only interprets white spaces as token boundaries (‘conservative’). Apart from the tokenization layers, the base foundry contains two layers that store sentence and paragraph boundaries.

```
<raw_text docid="WPD_AAA.00001">
  <metadata file="metadata.xml"/>
  <text>A bzw. a ist [...]</text>
</raw_text>
```

Figure 1: Extract from a file storing primary text.

```
<metadata docid="WPD_AAA.00001">
  <doc file="text.xml" />
  <foundry name="base"
    path="base/" />
</metadata>
```

Figure 2: Extract from a document metadata file.

The first deviation from the standard Lucene indexing process is that KorAP does not want Lucene to perform the tokenization during indexing because it uses tokenizations produced externally and independently of the indexing process. One way to achieve this would be to reimplement the tokenization algorithm(s) embedded into a Lucene analyzer so that the Lucene tokenizer exactly re-produces the results of the external tokenizer. This would comply with the Lucene-native solution, but not allow the engine to work with a tokenization from which only the results are known, without the algorithm that produced it, e.g. from a closed-source tool or uploaded by a user.

The KorAP implementation uses a Lucene analyzer that takes as input the location of an XML file instead of the actual primary text. That file lists the spans or tokens (cf. Figure 3) and provides a pointer to the primary text; this is implemented indirectly through the foundry metadata. The KorAP analyzer thus parses three files – the tokenization layer, the foundry metadata, and the primary text file – and generates the tokens for indexing. This method yields a Lucene index based on the tokens defined in the tokenization layer file without introducing any own tokenization logic during the indexing.

```
<layer docid="WPD_AAA.00001">
  <spanList>
    <span from="64" to="67" />
    <span from="68" to="73" />
  </spanList>
</layer>
```

Figure 3: Extract from an annotation file segmenting the primary text.

4.2 Annotations

On an abstract level, the indexing engine interprets all annotations in the same way: as character spans to which values are assigned. From an implementation point of view, an annotation either only defines a span by character offsets or it additionally provides a value to that span, e.g. a part-of-speech tag. In the former case, the actual character sequence is taken as the term to be indexed, while in the latter case, the tag value is the relevant information.

In the KorAP XML representation, a span has an optional feature structure in an `<fs>`-element. Figure 4 shows an annotation where different values (lemma, certainty, and morpho-syntactical tag) are assigned to a span. If a ``-element contains no `<fs>`-element, it is a purely segmenting annotation (cf. Figure 3).

4.3 Concurrent Tokenizations

KorAP shifts the definition of a token away from the engine and towards the tools and users that provide tokenizations. It is designed to accept any tokenization logic, leaving the judgement about its meaning and usefulness entirely to the user.

```

<span from="0" to="1">
  <fs type="lex">
    <f name="lex">
      <fs>
        <f name="lemma">A</f>
        <f name="certainty">0.780715</f>
        <f name="ctag">NN</f>
      </fs>
    </f>
  </fs>
</span>

```

Figure 4: A span annotated with a feature structure.

This flexibility resolves an issue that has appeared whenever text analysis is based on tokens: different results in tokenization are the norm rather than the exception (Chiarcos et al., 2009). While providing two basic tokenizations, KorAP allows for custom tokenizations as well, so that users can opt to base their queries either on one of the built-ins or on any other tokenization for their research. The user always specifies a tokenization layer to query, although the user interface can generally set a default when she has not. Also, multiple tokenizations can be queried at the same time in one search request.

4.4 Architecture and Implementation Details

4.4.1 Storing Higher-Level Information

As described before, an inverted index is always bound to some definition of tokens. In Lucene, a token object has an optional *Payload* field that holds an array of bytes. This is the place where higher-level information can be stored, for instance the sentence in which a specific token instance has occurred; cf. Brin and Page (1998). This is implemented by assigning IDs to such above-token spans: when parsing the XML file that holds sentence boundary information, the analyzer derives an ID for each sentence that encodes a reference to the annotation layer, a pointer to the primary text, and positional information.

In order to find, for instance, two or more tokens that occur in the same sentence, the search engine matches their sentence IDs, stored in the payloads. Because the IDs represent the positional order between sentences, the engine can as

well search for a word occurring in two subsequent sentences. This allows for more complex queries, for instance to find two words that occur in the same verb phrase, but in different noun phrases: the IDs for the verb phrase have to match and the IDs for the noun phrase have to differ.

Inverted indexes are efficient only when querying on the base of token strings. However, annotations have to be searchable efficiently, too, which is why they cannot just be stored as payloads on tokens. This is where the previously mentioned abstraction of segmenting and labelling annotations comes into play: by applying the same techniques as presented for tokens, further indexes can be built using keys that are not (only) based on the actual text character sequences, but rather on the vocabulary applied by the annotation tool or human annotator. For example, an index can be introduced that uses part-of-speech tags as keys and lists the occurrences for each tag. This method can also be applied for constituents, lemmas etc.

Different indexes can be queried separately and in parallel. Using the MapReduce paradigm (Dean and Ghemawat, 2004), queries are combined with queries on other indexes and merged using the appropriate set operation (see Section 4.4.2). The efficiency and scalability of this approach depends mainly on the vocabulary, i.e. the number of index keys, used in a given annotation, and will be evaluated during further development.

4.4.2 Map and Reduce in Practice

KorAP aims to be scalable, which means in practice that increasing corpus sizes must not lead to a growth in computational cost to an extent that could not be compensated by increasing hardware resources. In order to achieve this goal, KorAP parallelizes as many operations as possible, applying the MapReduce paradigm: each query is divided into independent sub-queries that can be processed independently of each other and eventually, the results of these sub-queries are merged to the final result. This allows for parallelization and distribution of work load to many processors, distributed across different machines, making the platform scale up for very large corpora by adding more machines to the system.

The Lucene index structure is well suited for

this strategy because any index can be split into smaller segments that can be distributed across different disks and machines. A central header node divides incoming queries into atomic sub-queries and distributes them to the machines in the cluster that hold the actual indexes, the worker nodes. Each worker node queries its index(es) and sends back a set of results to the header node where the result sub-sets are merged to the full result set.

For instance, if the index of a full corpus is to be distributed across five machines, five indexes are created, each based on a different (possibly overlapping) sub-set of the total document collection. When a simple term query comes in to the head node, e.g. “find all occurrences of word ‘A’”, the head node forwards that request to all five worker nodes. Each of them searches its respective sub-index and returns a set of occurrences to the head node. The head node joins the five result sets in order to produce the final result.

For a more complex sample query like “find word ‘A’ in sentences that also contain word ‘B’”, the query is split into two queries “find word ‘A’” and “find word ‘B’” and these two queries are sent to the worker nodes; this is the map step in MapReduce terminology. After the worker nodes have returned their results for both queries, each of the sub-results for ‘A’ and ‘B’ are joined to two sets. These sets are intersected – corresponding to the Boolean AND-operator – so that only results from the two sub-sets end up in the final result set that share the same sentence ID; the reduce step in MapReduce speak. Other Boolean operators – OR and NOT – can be realised through the union and difference set operations. Some queries could limit the search to a sub-set of the full corpus (virtual corpora, cf. Kupietz et al. (2010)) so that only those partitions have to be addressed that hold the relevant sub-corpora.

In the KorAP implementation, each query is at first divided into the annotation layers that are queried. In a second step, each of the single terms is isolated. The latter forms the smallest unit of a query and corresponds to a callable thread at runtime. The reducing is performed in the inverse order: the term results for a layer, returned by the different threads, are joined according to the conjunctions specified in the query (AND, OR, NOT).

Subsequently, the layer results are joined to form the total result set.

In order to speed up actual query performance upon intensive querying, index partitions can be stored redundantly, so that similar queries can be handled by different machines in parallel. This allows scaling up the platform to an almost unrestricted extent: when requests that query a certain partial index turn out not to be answered fast enough, another machine can be added to the cluster that helps out at frequent tasks.

Another potential bottleneck is the head node, although its main tasks – atomizing and distributing queries – are computationally less costly. However, multiple head nodes can parallelize these tasks as well when necessary.

5 Results

Benchmarking indexing and querying performance of a platform like KorAP has to consider a large number of factors, some of which are mutually dependent. Corpora of relevant size do not fit on a single hard disk, so that the file system choice matters as well as the distribution strategy, across multiple hard disks and across a network. Also, RAM capacity could lead to unexpected result because the system might not show significant speed reductions with increasing data amounts as long as it can hold the indexes in memory, but as soon as the platform has to store intermediate results on a hard disk, processing performance could decrease all of a sudden.

Performance comparisons between solid-state disks (SSDs) and magnetic hard disks (HDDs) could show surprising impacts as well because the physical location and ordering of data on the disk plays an important role. If data is dispersed across the disk, an HDD reading head has to jump across the platter while SSDs do not suffer speed losses in that scenario. When reading a sequence of bytes, the reading speed might not differ significantly though.

In order to find comparable statistics that yield insights about how the presented Lucene-based implementation scales in relation to the data volume, indexes from corpora of different sizes were created, all of them sub-samples of DeReKo (IDS, 2011). The number of documents in each test corpus are reported in Table 1.

#	#Documents	Index Size
1	29,704	1.6 GByte
2	196,854	16 GByte
3	512,542	33 GByte
4	974,722	42 GByte
5	1,487,264	75 GByte
6	2,080,111	76 GByte
7	3,597,079	151 GByte

Table 1: Sample corpora – sub-sets of the DeReKo corpus – were used to test query performance (cf. Tables 2, 3, 4, 5). The reported sizes sum over three tokenization layers and one annotation layer contained in the indexes.

We have applied two different tokenization algorithms, a sentence splitter, and TreeTagger (Schmid, 1994), including its own tokenizer, on the full corpus. Our self-made tokenizers follow different approaches in disputable cases such as hyphenations within words; the ‘conservative’ method treats such items as one token, while ‘greedy’ splits instances like “Ski-WM” into three tokens. By default, ‘conservative’ has been used in search.

The following queries were applied to all the sample corpora:

1. Simple token search: ‘Alphabet’.
2. Concurrent tokenization: ‘Ski-WM’ in *conservative tokenization* and ‘Ski’ in *greedy*.
3. Inter-layer search: ‘Buchstabe’ and ‘Alphabet’ occurring in one sentence.
4. Wildcard search: All tokens that start with ‘Alpha’ (‘Alpha*’).
5. Part-of-speech (POS) search: ‘Alphabet’ tagged as noun (NN) (both tokenized and tagged by TreeTagger).
6. Inter-layer annotation search: Token ‘Alphabet’ (*conservative*), tagged as Noun (NN) by TreeTagger.
7. High frequency: Tokens tagged as personal pronoun (PPER) that start with ‘er’.

Tables 2, 3, 4, and 5 report the average response times for the different queries executed

Sample	Query 1		Query 2	
	reponse	#hits	reponse	#hits
1	0.038s	10	0.021s	0
2	0.277s	996	0.198s	53
3	0.301s	1,078	0.291s	136
4	0.135s	82	0.578s	472
5	0.226s	1,160	0.616s	608
6	0.225s	380	0.418s	262
7	0.400s	1,550	0.803s	870

Table 2: Queries 1 and 2 for all sample corpora (Table 1), reporting response time and number of hits.

Sample	Query 3		Query 4	
	reponse	#hits	reponse	#hits
1	0.034s	0	0.397s	837
2	0.057s	89	0.201s	2,575
3	0.131s	89	0.351s	3,308
4	0.031s	1	0.175s	1,167
5	0.102s	90	0.487s	4,475
6	0.055s	8	0.287s	2,214
7	0.140s	98	0.828s	7,526

Table 3: Queries 3 and 4 for all sample corpora (Table 1), reporting response time and number of hits.

three times on each of the sample corpora listed in Table 1, every time disregarding the first call in order to allow caching to take effect. For each match, the surrounding sentence was returned as context; for two matches within one context sentence, that sentence was counted as one hit only. Therefore, the number of hits does not exactly represent the number of matches; this is especially important because searches were aborted after 1 million matches where necessary, yielding different numbers of hits.

The indexes were stored on a Linux machine with 48 CPU cores, 256 gigabytes of memory, on an Ext4 file system in a storage area network (SAN) on a RAID-5-volume. In order to retrieve the context properly, the Lucene SpanQuery class has been used in all cases. The present indexing implementation has been based on Lucene version 3.6.0.

Comparable index building times are not available here because they depend on factors like the XML parser, file system performance, and network load that lie beyond the scope of this pa-

Sample	Query 5		Query 6	
	reponse	#hits	reponse	#hits
1	0.015s	10	0.043s	9
2	0.066s	833	0.110s	564
3	0.044s	915	0.158s	624
4	0.018s	70	0.015s	52
5	0.055s	985	0.087s	676
6	0.026s	374	0.039s	279
7	0.068s	1,369	0.098s	964

Table 4: Queries 5 and 6 for all sample corpora (Table 1), reporting response time and number of hits.

Sample	Query 7	
	reponse	#hits
1	0.333s	8,489
2	4.234s	186,969
3	9.032s	421,229
4	10.721s	522,118
5	20.362s	910,038
6	19.543s	892,494
7	19.544s	894.905

Table 5: Query 7 for all sample corpora (Table 1), reporting response time and number of hits; search was aborted after 1 million matches.

per. However, parsing the XML input files and building indexes for the corpora reported in Table 1 took between 4 and 60 hours.

Anyway, indexing is not the most time-critical part in the KorAP scenario: new DeReKo versions are released only twice a year. At those points, they can be indexed in background while users can still use the previous release. The search, on the other hand, is expected to deliver instant results whenever possible. There might be compromises necessary in case of very complex queries, but the querying side is where KorAP places emphasis on performance. In short: the engine has not been optimized to build indexes fast, but to be queried fast.

The reported response times reveal one result very clearly: the querying time depends on the number of hits more than on the size of the corpora or indexes. In a simple token query (Query 1), sample corpus 3 contains many more hits than sample 4 which results in a threefold response time, despite having only approximately half of

the size. Less obvious differences between the corpora such as document sizes, vocabulary and other factors probably play a role too, but the number of hits seems to be most significant.

Not surprisingly, more complex queries take longer to process, but reply times still increase less than linearly in relation to corpus size. Intersections only increase response significantly when there are very many hits either, as in Query 7 (Table 5). In order to avoid intersections with large result sub-sets, one approach is a smarter distribution of indexes, so that joins can be performed at an early stage with smaller sets, before the final reduce step is performed on the full set. Another factor is query optimization so that complex queries are rewritten in order to avoid intersections and set differences whenever possible.

6 Summary & Outlook

We have used the Lucene engine to develop an indexing module for the KorAP corpus analysis platform. We have implemented a Lucene analyzer class that handles pre-analysed texts and corpora with multi-level stand-off annotations. The platform shifts away the task of segmenting, tokenizing, and analysing corpora towards the users and external tools and is ready to include new annotations of different kinds. We have presented a way to apply inverted indexes within a MapReduce-like environment, parallelizing tasks and making the platform scalable and ready to process very large corpora. Implicitly, this work demonstrates that techniques and software from the related field of IR can successfully be applied for linguistic search tasks.

Acknowledgements

This work has been made possible by the *Institut für Deutsche Sprache* (IDS) in the context of the KorAP project, funded by the *Leibniz-Gemeinschaft* and is a direct result of the work with the other KorAP core team members Piotr Bański and Elena Frick. Further IDS colleagues from within and outside the KorAP team have contributed important theoretical advice and practical help: Cyril Belica, Peter Fankhauser, Marc Kupietz, Roman Schneider, Oliver Schonefeld, and the Cosmas-II-Team: Franck Bodmer, Peter Harders, Helge Krause.

References

- P. Bański, C. Belica, H. Krause, M. Kupietz, C. Schnofer, O. Schonefeld, and A. Witt. 2011. KorAP data model: first approximation, December.
- P. Bański, Peter M. Fischer, E. Frick, E. Ketzan, M. Kupietz, C. Schnofer, O. Schonefeld, and A. Witt. 2012. The new IDS corpus analysis platform: Challenges and prospects. In *Proceedings of LREC-2012*, Istanbul, May.
- F. Bodmer. 2005. COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 21(3):2–5.
- S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.
- C. Chiarcos, J. Ritz, and M. Stede. 2009. By all these lovely tokens...: merging conflicting tokenizations. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 35–43. Association for Computational Linguistics.
- J. Dean and S. Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, December.
- R. Fiehler and P. Wagener. 2005. Die Datenbank Gesprochenes Deutsch (DGD)-Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion*, 6:136–147.
- S. Ghodke and S. Bird. 2008. Querying linguistic annotations. In *Proceedings of the 13th Australasian Document Computing Symposium*, pages 69–72.
- S. Ghodke and S. Bird. 2010. Fast query for large treebanks. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 267–275. Association for Computational Linguistics.
- IDS. 2011. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2011-II.
- D. Janus and A. Przeiórkowski. 2007. Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 85–88. Association for Computational Linguistics.
- A. Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147151.
- D.E. Knuth. 1997. *The Art of Computer Programming*. Addison-Wesley, Reading, Massachusetts, 3rd edition.
- M. Kupietz, C. Belica, H. Keibel, and A. Witt. 2010. The German reference corpus DeReKo: a primordial sample for linguistic research. In *LREC 2010 Main Conference Proceedings*. Malta.
- Lucene, 2012. *Apache Lucene – Index File Formats*. The Apache Software Foundation, April.
- M. McCandless, E. Hatcher, and O. Gospodnetić. 2010. *Lucene in Action*. Manning Publications Co., 2nd edition, July.
- R. Perkuhn, H. Keibel, and M. Kupietz. 2012. *Korpuslinguistik*. Fink. (UTB 3433), Paderborn.
- H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.
- R. Schneider. 2012. Evaluating DBMS-based access strategies to very large multi-layer corpora. In *Proceedings of the LREC 2012 Workshop: Challenges in the management of large corpora*. European Language Resources Association (ELRA).
- J. Sinclair, 2004. *Trust the text*, chapter 1. Routledge, Milton Park, UK.
- A. Sokirkö, 2003. *A technical overview of DWDS/Dialing Concordance*.
- Henry S. Thompson and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe*.
- M. Volk. 2002. Using the web as corpus for linguistic research. *Tähendusepüüja. Catcher of the Meaning. A Festschrift for Professor Haldur Oim*.
- A. Zeldes, J. Ritz, A. Lüdeling, and C. Chiarcos. 2009. Annis: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, pages 20–23.
- C. Zhang, J. Naughton, D. DeWitt, Q. Luo, and G. Lohman. 2001. On supporting containment queries in relational database management systems. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, volume 30, pages 425–436. ACM.

Three Approaches to Finding German Valence Compounds

Yannick Versley, Anne Brock, Verena Henrich, Erhard Hinrichs

SFB 833 / Seminar für Sprachwissenschaft

Universität Tübingen

firstname.lastname@uni-tuebingen.de

Abstract

Valence compounds (German: *Rektionskomposita*) such as *Autofahrer* ‘car driver’ are a special subclass in the otherwise very heterogeneous class of nominal compounds. As the corresponding verb (*fahren* ‘to drive’ in the example) governs the (accusative) object (*Auto* ‘car’), valence compounds allow for a straightforward (event-)semantic interpretation. Hence the automatic detection of valence compounds constitutes an essential step towards a more comprehensive approach to the analysis of compound-internal semantic relations. Using a hand-annotated dataset of 200 examples, we develop an accurate approach that finds valence compounds in large-scale corpora.

1 Introduction

German, Dutch, and other languages exhibit the phenomenon of word formation by compounding: In a process where nouns, verbs and other roots combine with a head noun, a novel word can be formed which is typically interpretable by considering its parts and the means of combination.

Previous research on compounds in German computational linguistics has concentrated on the question of accurately splitting them: Schiller (2005) and Marek (2006) present finite state approaches for accurate compound splitting, Koehn and Knight (2003) use a parallel corpus to find appropriate splits for compounds without oversplitting them.

For the English language, previous years have seen renewed interest in the semantic interpreta-

tion of noun-noun compounds which are the most conspicuous kind in English. Research such as that by Girju et al. (2005) and Ó Séaghdha (2007), *inter alia*, has concentrated on automatic classification of the compound-internal relations.

Compounds are a rich source of examples even for semantic relations crossing part-of-speech categories, e.g. when the head part of the compound is a nominalization. In this paper, we want to focus on the detection of one particular kind of cross-part-of-speech relation between nouns and deverbal nominalizations in so-called *valence compounds*. Our goal is to identify noun-object valence compounds among the words occurring in a corpus using a combination of morphological, statistical and semantic evidence.

Valence compounds are an interesting subset of all compounds due to the fact that they have a straightforward (event-)semantic interpretation, namely that the modifier noun fills an argument slot in the event expressed by the head’s nominalized verb, as in *Volkszählung* ‘people count’, which corresponds to an event where the people are counted.

Existing linguistic research cautions us that not everything that looks like the targeted construction (valence compounds with an accusative object modifier) is a valid example: The corpus-based studies of Wellmann (1975) and Scherer (2005) find that, even among -*er* derivations, the resulting noun does not realize an agent in all cases; Kohvakka and Lenk (2007) confirm that agentive nominals can also inherit other kinds of arguments such as prepositional objects (about 10% in their study). Finally, Gaeta and Zeldes

(2012), in their corpus study of valence compounds, remind us that not all such compounds correspond to an interpretable verb-object pair.

Therefore, the kind of valence compounds we are interested in (those that do correspond to an interpretable verb-object pair) constitute only a part of all compounds with a plausible morphology; but it is also the case that, since compounding presupposes a certain degree of semantic integration, such valence compounds are more specific than verb-object collocates in general.

2 Related Work

Lapata (2002) presents an approach to discriminate between compounds with an object- or subject-related modifier and a deverbal nominalization head. From the British Nominal Corpus (BNC), Lapata extracted a 1,277 item sample of such compounds that contained an actual nominalization head according to CELEX or NOMLEX. She then created a gold standard dataset containing a subset of 796 nominalization compounds where the premodifying noun corresponded either to a subject or to an object relation.

To predict whether a given compound is a subject or object nominalization, Lapata estimates the ratio of subject versus object occurrences among the co-occurrences of the noun and verb from which the valence compound is derived, and complements the raw frequency with class-based smoothing via hand-crafted resources (WordNet and Roget’s thesaurus), as well as distance-weighted averaging by distributional similarity across verbs. Lapata combines different smoothing methods using decision list classification (Ripper: Cohen, 1996), yielding a final accuracy of 86.1%.

Lapata’s work for English has a slightly narrower goal than ours, as she aims to discriminate verb-object nominalization compounds from verb-subject ones, rather than from all other types of compounds as is our goal. Nonetheless, her work is most similar in spirit to the specialized approach to valence compounds that we will advocate later in this paper.

3 Dataset

The target set of compounds for our experiments consists of 100 compound instances labeled as va-

lence compounds and 100 others. To find both the positive and negative examples, we prepared a list of nouns exhibiting a morphological structure compatible with being a valence compound.

Using the morphological analyzer SMOR (Schmid et al., 2004), we prepared a list of such compounds in which the nominalized verbs and nominal non-heads are simplex words (i.e., not compounds or derivations). Specifically, the SMOR analysis had to consist of a bare noun followed by a bare verb (in contrast to *Dienstagabend* ‘Tuesday night’, which does not have a deverbal head, or to *Netznutzungsentgelt* ‘network access fee’, which has a complex modifier *Netznutzung* ‘network access’). The dataset includes a variety of nominalization suffixes: *-ung*, *-en*, *-er*, *-erei*, as well as some less frequent ones.

A compound is labeled as a valence compound when its most common interpretation is that of a verb and its direct object, although some of them may be ambiguous. As a negative example, the word *Serienmörder* ‘serial killer’ consists of the nominalized verb *morden* ‘to murder’ and the noun *Serie* ‘series’. While the word could be used to denote a TV executive killing a series (i.e., canceling a show), the common meaning is different and we would not count the word as a valence compound.

4 Methods

As a text collection that provides contexts for the words or word pairs that interest us, we use the *TüPP-D/Z* corpus of *tageszeitung* news articles from 1986 to 1999 (Müller and Ule, 2002), and the *web-news* corpus, a 1.7 billion word collection of online news articles by Versley and Panchenko (2012). The parsing model used in the pipeline is based on MALTParser, a transition-based parser (Hall et al., 2006), and uses part-of-speech and morphological information from RFTagger (Schmid and Laws, 2008) as input. Using the MALTParser support for linear classification by Cassel (2009), we reach a parsing speed of 55 sentences per second.

4.1 Simple Association Statistics

One very straightforward idea for the identification of valence compounds is to check for a valence of the verb (corresponding to the deverbal

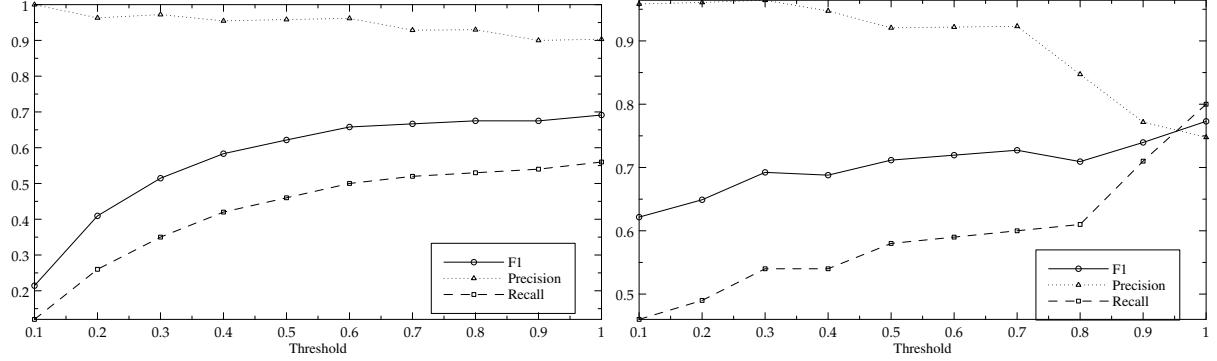


Figure 1: Precision/Recall/F₁ values for different thresholds on pointwise mutual information (left) and log-likelihood (right)

<i>SVM</i>	Acc	F ₁
GN+triples+w1w2	0.695	0.697
GN+triples/scale	0.705	0.619
GN	0.615	0.617

Table 1: Results for feature-rich SVM classification

part of a putative valence compound) with a selectional preference that would admit the noun (i.e., the premodifying part of a candidate).

From the verb-object pairs in our corpus (i.e., accusative objects tagged as OBJA), we calculate association statistics that consider counts for this relation across different verbs and arguments seen in the corpus. The most common statistics are *pointwise mutual information* and conservative estimates thereof, and the G^2 significance statistic (*log-likelihood*) proposed by Dunning (1993).

Figure 1 shows precision and recall for the task of identifying valence compounds when using different thresholds on the respective statistic.

Another way to look at the problem is to look at the *relative frequency* of a verb and a noun that co-occur in a sentence being connected by an OBJA edge (or the subject of a passive construction) rather than by some other dependency (e.g., the noun being the subject of the verb or occurring inside a prepositional phrase). As seen in Figure 2, imposing different thresholds on the ratio between object and non-object occurrences yields a recall between none and about 80% of valence compounds, and precision above 75%.

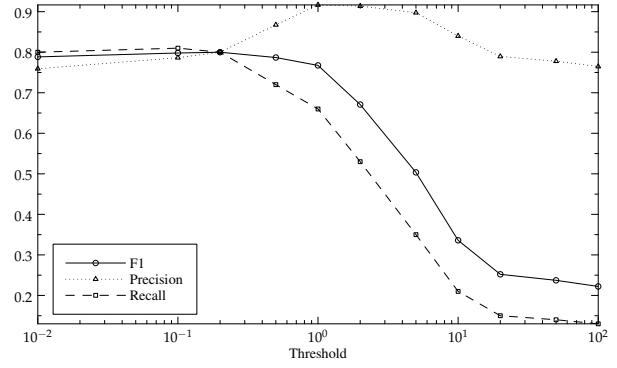


Figure 2: Precision/Recall/F₁ values for different thresholds on relative frequency

4.2 Feature-rich Supervised Classification

In the domain of more general approaches to the prediction of a relation between two words – for example, the two nouns in a noun-noun compound – more feature-rich approaches have been developed that can take into account *all paths* that occur between the two target words as well as *taxonomic* information.

In our version of such an approach, we used taxonomic relations in GermaNet (Kunze and Lemnitzer, 2002; Henrich and Hinrichs, 2010), denoted as *GN* in Table 1, as well as features derived from corpus co-occurrences, namely *triples* along the dependency path as well as words occurring in-between or around the target words in a co-occurrence (*w1w2*). The corpus-based features are illustrated in Figure 3.

4.3 Decision-tree Based Combination

To combine multiple statistics such as those described in subsection 4.1 and possibly smoothed

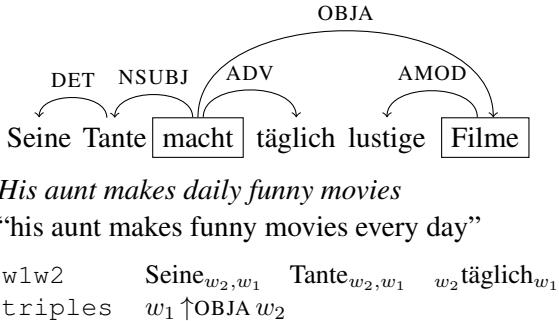


Figure 3: Features for SVM classification

<i>JRip</i>	Acc	F_1
mi+ll	0.770	0.776
rel	0.860	0.863
mi+ll+rel	0.840	0.837
mi+ll/avg	0.740	0.743
rel/avg	0.845	0.841
mi+ll+rel/avg	0.840	0.833
<i>J48</i>	Acc	F_1
mi+ll	0.785	0.792
rel	0.860	0.863
mi+ll+rel	0.850	0.851
mi+ll/avg	0.765	0.759
rel/avg	0.850	0.854
mi+ll+rel/avg	0.845	0.846
<i>AdaBoost+J48</i>	Acc	F_1
mi+ll	0.780	0.786
rel	0.830	0.835
mi+ll+rel	0.810	0.812
mi+ll/avg	0.780	0.770
rel/avg	0.710	0.655
mi+ll+rel/avg	0.830	0.830

Table 2: Results for decision tree classification

variants thereof, we use symbolic learning techniques such as decision lists (Cohen, 1996) or decision trees (Quinlan, 1993) which are aimed at finding logical combinations of thresholds, possibly in combination with AdaBoost as a meta-learner (Freund and Schapire, 1997).

In our case, we try both the association statistics for the OBJA relation and the relative frequency heuristic, as either raw values or averaged over a group of 10 related nouns. The set of related nouns was determined by taking 30 surrounding terms from GermaNet, then ranking by distributional similarity (using premodifying adjective and accusative object collocates, and similarity based on Jensen-Shannon divergence).

5 Results and Discussion

The statistics in Figures 1 and 2 have been determined on the full dataset, as the risk of overfitting is very low when fixing a single threshold parameter. The more comprehensive approaches using high-dimensional features (Subsection 4.2) or combining multiple statistics using symbolic learning (Subsection 4.3) were run as ten-fold crossvalidation where each 10% slice was predicted using a classifier built from the remaining 90% of the data.

The statistics show that the relative frequency heuristic yields the best F_1 measure (0.80) for a 1:5 cutoff ratio between accusative objects and other paths. In contrast, association statistics yield a best F_1 of 0.77 (log-likelihood).

In comparison, the feature-rich approach (see results in Table 1) does not seem particularly attractive: taxonomic information from GermaNet alone yields an F_1 measure of 0.62, and even the most sophisticated feature set that additionally takes into account surface-based and dependency-based features from the collocation contexts of noun and verb only yield an F_1 measure of 0.70.

The decision list and decision tree based methods allow to explore (and potentially to combine) larger sets of features with different corpora and smoothed/unsmoothed variants of the statistics. In our experiments, we found that the best classifier selected the relative frequency heuristic (but selecting statistics from the larger *web-news* corpus instead of the smaller TüPP-D/Z ones), reaching an F_1 measure of 0.86.

6 Summary

We demonstrated several methods to analyze the relations inside nominalizations and identify valence compounds. While a generic feature-rich approach works moderately well, we find that tightly focused statistics such as those investigated by Brock et al. (2012) can be easily combined using symbolic machine learning methods to yield a highly accurate discrimination between valence compounds and non-valence compounds.

Acknowledgements We are grateful to the three anonymous reviewers for insightful comments. Anne Brock and Yannick Versley were supported by the DFG as part of SFB 833.

References

- Brock, A., Henrich, V., Hinrichs, E., and Versley, Y. (2012). Automatic mining of valence compounds for German: A corpus-based approach. In *Digital Humanities Conference Abstracts*, Hamburg. Hamburg University Press.
- Cassel, S. (2009). MaltParser and LIBLINEAR – transition-based dependency parsing with linear classification for feature model optimization. Master's thesis, Uppsala University.
- Cohen, W. W. (1996). Learning trees and rules with set-valued features. In *Proc. 13th National Conf. on Artificial Intelligence (AAAI 1996)*.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Freund, Y. and Schapire, R. (1997). Decision-theoretic generalization of on-line learning and an application to boosting. *J Comp Sys Sci*, 55(1):119–113.
- Gaeta, L. and Zeldes, A. (2012). Deutsche Komposita zwischen Syntax und Morphologie: Ein korpusbasierter Ansatz. In Gaeta, L. and Schlücker, B., editors, *Das Deutsche als kompositionsfreudige Sprache: Strukturelle Eigenschaften und systembezogene Aspekte*, pages 197–217. De Gruyter, Berlin.
- Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions*, 19(4):479–496.
- Hall, J., Nivre, J., and Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 316–323.
- Henrich, V. and Hinrichs, E. (2010). GernEdiT – the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.
- Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *EACL 2003*.
- Kohvakka, H. and Lenk, H. E. H. (2007). Streiter für Gerechtigkeit und Teilnehmer am Meinungsstreit? Zur Valenz von Nomina Agentis im Deutschen und Finnischen. In *Wahlverwandtschaften. Valenzen - Verben - Varietäten*, pages 195–218. Georg Olms, Hildesheim, Zürich, New York.
- Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*.
- Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.
- Marek, T. (2006). Analysis of German compounds using weighted finite state transducers. Bachelor of arts thesis, Universität Tübingen.
- Müller, F. H. and Ule, T. (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proc. 19th Int. Conf. on Computational Linguistics (COLING 2002)*.
- Ó Séaghdha, D. (2007). Annotating and learning compound noun semantics. In *Proceedings of the ACL07 Student Research Workshop*.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Scherer, C. (2005). *Wortbildungswandel und Produktivität: Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Niemeyer.
- Schiller, A. (2005). German compound analysis with wfsc. In *Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2005)*.
- Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC*.
- Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *COLING 2008*.
- Versley, Y. and Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. In *Proceedings of the 7th Web as Corpus Workshop (WAC-7)*, pages 44–52.
- Wellmann, H. (1975). *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache*. Schwann, Düsseldorf.

A Computational Semantic Analysis of Gradable Adjectives

Mantas Kasperavicius

Ruhr-Universität Bochum

Department of Linguistics

Mantas.Kasperavicius@rub.de

Abstract

This paper describes ongoing work towards a computational model for the analysis of gradable adjectives, including dimensional/evaluative adjectives, modifiers, and comparative and incommensurable adjectives. The approach is based on a representation of conceptual comparison classes and a flexible construction of scales. Input sentences are compositionally analysed by means of the λ -calculus. First evaluation results support the theoretical approach reported in this paper.

1 Introduction

The goal of this work is to automatically analyse adjectival constructions for natural language-based database queries and web searches. These queries often involve adjectival forms. Users might be interested in certain attributes of entities or in a comparison of such attributes, as illustrated in (1).

- (1) a. Is Spain large/How large is Spain?
- b. Is BVB Dortmund more successful than Bayern Munich?/How successful is BVB Dortmund?

Adjectives like *large* or *successful* express judgements w.r.t. contextually determined scales. Therefore, an analysis of adjectives as simple one-place predicates is not sufficient. According to Klein (1980, p. 9) and Kennedy (2007, p. 4), adjectives denote functions which map their argument to the *positive* or *negative extension* of a

scale, or an undetermined middle section, the *extension gap*. Their meaning arises from the interpretation in context of a standard of comparison, which is derived from comparison classes and fixed by a certain degree on a scale.

The present approach shows how such an analysis can be carried out computationally, by bringing together distinct theories and taking into consideration modification (2a), dimensional/evaluative adjectives (2b), and incommensurable adjectives (2c).

- (2) a. Is Spain *very large*?
- b. Is Dale *taller/more intelligent* than Andy?
- c. Is the cupboard *taller* than the desk is *clean*?

2 Theoretical Foundation

2.1 A Hierarchy of Adjectives

Following Kennedy (2007, p. 21), different types of adjectives are interrelated and brought together in a hierarchical order as shown in fig. 1.

Relative adjectives are interpreted against the background of a standard of comparison, which is derived from comparison classes. The interpretation of absolute adjectives does not depend on context.

As illustrated in (3), arguments modified by minimum standard adjectives need to have only some amount of the described property in order to be interpreted as true, whereas arguments modified by maximum standard adjectives need to exhibit the maximum amount of a property.

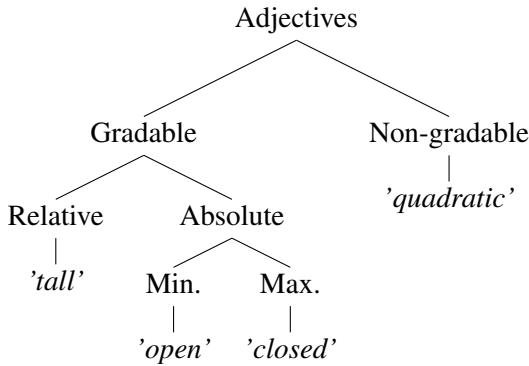


Figure 1: A Hierarchy of Adjectives

- (3) a. The floor is *slightly/completely* wet.
 b. The floor is **slightly/completely* dry.

2.2 Scale Construction & Comparison Classes

Kennedy & McNally (2005, p. 351) point out that scales are definable as triples $\langle D, \prec, \delta \rangle$, where D is a set of degrees, \prec describes a total ordering on the set D , and δ is the dimension. A closer examination of the set of degrees D reveals four types of possible structures for D . Either it lacks a minimal or maximal degree or both, or it possesses one or both degrees. Thus, four types of scales are assumed: *totally open*, *lower closed*, *upper closed*, and *totally closed*.

Yoon (1996, p. 222) makes a further distinction concerning *total adjectives* like *clean/safe* or *partial adjectives* like *dirty/dangerous*. The former describe the lack of a property (dirt, danger), while the latter denote its existence. Rotstein & Winter (2004, p. 272) propose the following structure for such adjectives. The minimum end of the partial adjective P_{min} is equal to the standard of the total adjective d_t , since e.g., a minimally dirty object is at least on the verge of being clean. While the standard of a partial adjective d_p can appear anywhere on the *partial* scale. Rotstein & Winter conclude that *total adjectives* can be a degree on a scale, yet do not necessarily have to be. *Partial adjectives* on the other hand always denote an interval.

The standard of comparison of gradable adjectives depends on context and is calculated considering an appropriate comparison class, which can be specified implicitly as in (4a) or explicitly as in

(4b) (Bale, 2011, p. 169). In the first case Andy is (most likely) considered to be tall for the class *men*. In the second sentence the *for-clause* determines the set of basketball players as comparison class.

- (4) a. Andy is tall.
 b. Dirk is tall for a basketball player.

Although the vagueness of gradable adjectives appears to be *fuzzy*, fuzzy set analysis turns out to be inappropriate for the analysis of gradable adjectives (Dubois, 2011).

2.3 Dimensional (DA) vs. Evaluative Adjectives (EA)

While **DA** can be exactly measured by a system of measurement (e.g., a metric system), there is no objective measure for **EAs** like *smart*. Thus, their interpretation neither depends on a concrete average value, nor on a contextual standard. Bogal-Allbritten (2011) argues that at least negative **EAs** can be treated as a subclass of min. standard adjectives since they show similar entailment patterns (5)¹.

- (5) a. Sandy is ruder than Ben. \models Sandy is rude.
 b. Sandy is ruder than Ben, # but Sandy isn't rude.

Therefore, I will treat them as min./max. endpoint adjectives, following the total/partial adjective analysis (Rotstein & Winter, 2004). Since scales associated with **EA** do not consist of numerical degrees, their structure has to be slightly modified: instead of degrees, they consist of concrete objects/individuals (Toledo, 2011, p. 38)².

2.4 Modifiers

Databases can be queried not only using positive or comparative constructions but also with modified sentences as shown in (6).

¹Bogal-Allbritten (2011, p. 6)

²For an alternative, trope-based analysis of evaluative adjectives, see Moltmann (2009)

- (6) a. Is Spain *very* large?
 b. Is BVB Dortmund *much* more successful than Bayern Munich?

The semantic forms of modifiers (*very/much*) can be derived from the semantics of the *posmorpheme*³ (see Bartsch & Vennemann (1972), von Stechow (1984)), given in (7)⁴.

$$(7) \llbracket pos \rrbracket = \lambda G \lambda x. \exists d [\text{standard}(d)(G)(c) \wedge G(d)(x)]$$

For a detailed discussion, see Kennedy & McNally (2005).

2.5 Incommensurability

Incommensurable adjectives like (8) differ from positives and comparatives in that they feature (at least) two different adjectives associated with different scales.

- (8) The table is longer than the desk is clean.

Bale (2008) provides an approach that facilitates the comparison of such constructions and suggests a universal scale ' Ω ' as a device of comparison. Universal degrees contain information about the relative position of objects/individuals on their associated scales (*primary scales*) and are derived by mapping degrees from primary scales to the universal scale.

Note that while this is a theoretically interesting phenomenon, it is rarely common in everyday use.

2.6 Summary of the Semantic Issues to Deal with

The challenges for our approach are to determine and associate comparison classes with the respective adjectives. Then, appropriate scales for **DAs** and **EAs** have to be derived from these comparison classes accordingly. Finally, a compositional analysis and evaluation of the adjectival constructions has to be carried out.

3 Implementation

The Python implementation comprises two significant parts. First, the compositional analysis,

³*pos* denotes the adjective's *positive* form in this context.

⁴(Kennedy & McNally, 2005, p. 350)

which analyses valid input sentences and delivers a first-order formula. Second, the evaluation, which checks the truth conditions of valid input sentences and yields a truth value according to a certain domain.

3.1 Prerequisites

Before going into deeper analysis, the input string is tokenized and tagged using NLTK tools⁵, yielding a list containing tuples, which consist of (*token*, *POS-tag*)-pairs. Constituents are accessed via POS-tags and together with their automatically associated λ -expression stored in dictionaries as *key:value*-pairs.

The manually compiled databases for domains and adjectives are similar, consisting of *key:value*-pairs, where *value* is another dictionary containing *proper_noun:measurement*-pairs as illustrated in (9).

$$(9) \text{domain} = \{ \text{'height'} : \{ \text{'Dale'} : 180 \} \}$$

Nested dictionaries in lexical entries for adjectives contain the features of an adjective in a *attribute:specification*-pairs. As shown in (10)⁶, each adjective is assigned three attributes: *pol* (*polarity +/-*), *type* (**DA/EA**), and *domain*.

$$(10) \text{adjectives} = \{ \text{'short'} : \{ \text{'pol'} : \text{'-'}, \text{'type'} : \text{'DA'}, \text{'domain'} : \text{'height'} \} \}$$

3.2 Scale Construction & Comparison Classes

Adjectives are associated with the corresponding comparison class according to the information in the lexical entry. Afterwards, a scale is derived from the comparison class via quasi orders following Krantz et al. (1971). Since quasi orders allow reciprocal relationships between two distinct elements and scales do not, this reciprocity has to be removed.

Bale (2011, p. 175) divides the process into three steps: associating each element *x* in the domain of the quasi order *R* with an equivalence class *E_x*, imposing an ordering relation on the

⁵nltk.word_tokenize, nltk.pos_tag

⁶Lexical entries are simplified here and neither claim to represent the full range of adjective features nor the ambiguity inherent in some adjectives.

equivalence classes in \geq_R ⁷, defining a measure function mapping every element to its equivalence class.

Bale also shows that comparison classes restrict quasi orders to the extent that elements can only be compared to a certain comparison class, if they are its members. Further, these restricted quasi orders can be used to create measure functions and scales, which serve as input for functions calculating the standard degree.

Scales for **DAs** consist of degrees (e.g., body height), while **EAs** require scales of individuals. In order to handle such scales computationally, individuals in the domain of **EAs** are associated with numerical values that represent the individual's affiliation with the corresponding property.

3.3 Covert Morphemes

Covert morphemes like *pos* are not made overt in the compositional analysis. The semantic representations of *pos* and *deg*⁸ were rather converted one-to-one to a Python function. Thus, making use of the information from the database described in 3.1, they evaluate input sentences according to their context and yield truth values.

3.4 Compositional Analysis

For the compositional analysis, constituents of the input sentence are associated with λ -expressions first, according to their POS-tag. Analytic comparative forms again (e.g. *more intelligent*) are treated differently than synthetic forms, in that, e.g. *more* requires functions as arguments, while the synthetic form takes simple arguments.

After preprocessing, the module responsible for the compositional analysis carries out β -reduction. The analysis is then presented as a step-by-step-tree.

4 First Evaluation Results

A total of 36 phrases covering the phenomena discussed in this paper were chosen from the BNC⁹ for a first evaluation of the programme. For each type, phrases had to consist of a certain fixed pattern. Of all phrases tested, 22 (~61%) were ana-

lysed correctly. Yet, the program yielded a wrong semantic analysis for only three sentences (~8%). For the remaining 11 sentences, the main source of errors was to be found in the inaccurate tagging process and thus not due to semantic misinterpretation.

4.1 Example analysis

The compositional analysis yields a tree that illustrates how the final first-order formula is composed, shown in fig. 2 for the sentence *Taligent is inherently risky*¹⁰. Note that first the meaning *Taligent is risky* is composed, which is afterwards modified by *inherently*.

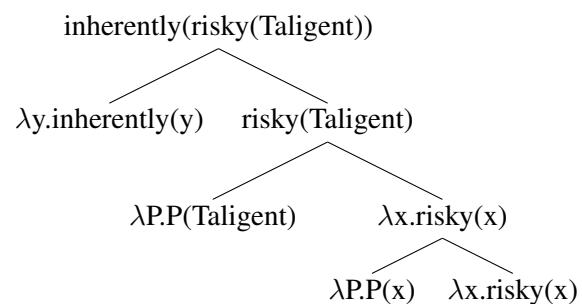


Figure 2: Analysis for *Taligent is inherently risky*

Taligent is then mapped to the corresponding scale, which consists of individuals (here: ventures) that are ranked according to their degree of “riskiness”.

5 Summary and Outlook

The aim of this paper was to show how different adjectival constructions can be compositionally analysed and evaluated w.r.t. scales and comparison classes. The evaluation gives evidence for the sustainability of the presented approach. As this is still ongoing work, the next step involves a more dynamic analysis. Instead of relying on a static database, it will be interesting to derive comparison classes from corpora. Furthermore, it would be desirable to extract information from context such as *dimension* or *type* of an adjective and thus automatically create lexical entries.

⁷ \geq_R denotes the derived scale

⁸Analogous to *pos*, the *deg*-morpheme denotes *degree modification* (see Kennedy & McNally (2005, p. 367))

⁹British National Corpus

¹⁰The original sentence *A venture like Taligent is inherently risky* has been simplified since the programme cannot handle modified head nouns.

References

- Alan C. Bale. 2008. A Universal Scale of Comparison. *Linguistics & Philosophy*(31), 1–55.
- Alan C. Bale. 2011. Scales and Comparison Classes. *Natural Language Semantics*, 19 (2), 169–190.
- Renate Bartsch and Theo Vennemann. 1972. *Semantic Structures: A study in the relation between semantics and syntax*. Athenaum: Frankfurt am Main.
- Elizabeth Bogal-Allbritton 2011. *Processing Evidence for the Scales of Negative Evaluative Adjectives*. (First Generals Paper. Ms., University of Massachusetts Amherst). URL: <http://blogs.umass.edu/eba/files/2011/08/GP-Rev.pdf> (accessed August 9, 2012).
- The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>
- Didier Dubois. 2011. Have Fuzzy Sets Anything to Do with Vagueness? In: P. Cintula, C. Fermüller, L. Godo, and P. Hájek (Eds.), *Studies of Logic*(36), 311–333. College Publications.
- Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30 (1), 1–45.
- Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification and the semantic typology of gradable predicates. *Language*, 81 (2), 345–381.
- Ewan Klein. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4 (1), 1–45.
- David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement*. Academic Press: New York, London.
- Friederike Moltmann. 2009. Degree Structure as Trope Structure: A Trope-Based Analysis of Positive and Comparative Adjectives. *Linguistics and Philosophy*(1), 51–94.
- Carmen Rotstein and Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higherorder modifiers. *Natural Language Semantics*(12), 259–288.
- Arnim von Stechow. 1984. Comparing Semantic Theories of Comparison. *Journal of Semantics*(3), 183–199..
- Assaf Toledo. 2011. *Measuring absolute adjectives*. The Hebrew University of Jerusalem. URL: http://www.hum.uu.nl/medeworkers/a.toledo/papers/Measuring_Absolute_Adjectives.pdf (accessed August 9, 2012).
- Youngeun Yoon. 1996. Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*(4), 217–236.

Extending Dependency Treebanks with Good Sentences

Alexander Volokh

alexander.volokh@dfki.de

DFKI, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

Abstract

For many resource-poor languages additional annotated data would be beneficial. However, annotation process is tedious and expensive. We propose a metric for selecting the most promising sentences for annotation. Annotating only good sentences saves time and would allow better results to be achieved even with a smaller amount of annotated data. We demonstrate how our method works on the example of parsing Finnish dependency treebank with MaltParser.

1 Introduction

Creation of annotated resources is a tedious and expensive work. Therefore for most languages resources like treebanks are rather small, despite the fact that learning curve experiments show that additional data would be beneficial and further improve performance of NLP applications.

In this paper we investigate whether it is possible to measure the usefulness of annotating and adding a sentence to an existing treebank and whether this value significantly differs across different sentences. We ascertain that it is indeed possible and show that a lot of time can be saved and much better results can be achieved, if one a) does not waste time on annotating *bad* sentences and b) concentrates on annotating *good* sentences. Finally, we construct a statistical model able to discriminate between those, evaluate it and discuss the results.

For our experiments we have used the Finnish treebank (Haverinen et al., 2010) and the dependency parser MaltParser (Nivre et al., 2007). Since we do not have any knowledge of the Finnish language and are not able to annotate sentences ourselves, we leave a portion of annotated sentences out of the treebank and add them in the process as if they were newly annotated by us. The reason for choos-

Günter Neumann

neumann@dfki.de

DFKI, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

ing Finnish is that it is a resource-poor language and its treebank is currently being annotated, so the work might be tested in a real-world application scenario, whereas the treebanks for languages that we speak, e.g. English or German, would lack this interesting property, because they are big and their development is completed.

2 Related Work

A similar work for dependency parsing exists in the area of active learning (Mirroshandel and Nasr, 2011). The work has exactly the same motivation of selecting only that material for annotation, which is beneficial for the performance of one's system and not wasting the time on the rest. They even propose not to annotate whole sentences manually, but rather process a sentence with a parser and only manually overwrite those dependencies which are actually useful.

The method of selecting good material proposed by Mirroshandel and Nasr is different from what we are proposing in this paper. Their idea is that those structures which are most error-prune require annotation, since obviously the model is missing the knowledge to process them. They then manually annotate the wrong parts, relearn the model and hope that the parser has learned to deal with the structures which went wrong before. Thus the most useful sentence is the one which contained the most mistakes prior to its manual annotation. On the other hand sentences which are already parsed correctly are not useful and do not require annotation.

Our proposal is that a sentence is useful and requires annotation if its inclusion to the training data improves the performance of the parser on the test data. This way the most useful sentence is the one whose annotation has the biggest positive effect on the performance of the parser on the test data.

There are also other works with the same motivation in the area of active learning for other domains, including other parsing areas, e.g. HPSG parsing (Baldridge and Osborne, 2008).

3 Sentence Usefulness

We define usefulness (su) of a sentence (s_i) as the influence of a sentence on the accuracy (acc) of a system for the test data (TE_m) when added to the training data (TR_n) of that system:

$$su(s_i) = acc(TR_n + s_i; TE_m) - acc(TR_n; TE_m)$$

For accuracy we use the metric called labeled attachment score (LAS), which represents the percentage of tokens for which both the parent and the dependency relation type were predicted correctly by the parser.

In order to compute a good estimate for sentence usefulness we use several training data sets and several test data sets and average the change in accuracies across all experiments.

For experiments described in this paper we have taken the Finnish treebank (6375 sentences) and split it into 2 training data sets (30% of the whole data each), 2 test data sets (10% each) and left the remaining 20% (1276 sentences) for experimenting. Thus for each of these 1276 sentences we have computed the accuracy for all (4) combinations and averaged over their number:

$$\begin{aligned} su(s_i) = & (acc(TR_1 + s_i; TE_1) - acc(TR_1; TE_1) + \\ & acc(TR_1 + s_i; TE_2) - acc(TR_1; TE_2) + \\ & acc(TR_2 + s_i; TE_1) - acc(TR_2; TE_1) + \\ & acc(TR_2 + s_i; TE_2) - acc(TR_2; TE_2)) / 4 \end{aligned}$$

In order to test whether our sentence usefulness measure serves its purpose, namely as a selection criterion for sentences which should be added to the training data foremost, we have performed several experiments. We have computed accuracies for randomly selected n sentences and compared it to the accuracies when the best n sentences where selected according to our metric. E.g. when 20 sentences were randomly added to the training data the accuracy improved by 0.08% LAS on average, whereas with 20 best sentences the improvement was 0.3% LAS. For 50 random sentences the improvement was 0.13% LAS, whereas for the best 50 it was 0.48% LAS.

4 Usefulness Classes

Having shown that different sentences have different impact on the accuracy and that it is beneficial to add good ones first, we have investigated whether it is possible to automatically predict this value using a statistical classifier.

Basically, we are not interested in the exact sentence usefulness decimal value, but we rather want to know whether a sentence is worth annotating or not. Therefore we can group sentences into some discrete usefulness classes. The core problem is binary, since we want to discriminate between good sentences we want to annotate and the rest. However, the number of good sentences is small and the rest is big, which makes the problem very imbalanced and the prediction difficult. Therefore we have experimented with several different divisions and could achieve the best results when the data is split into four classes of equal number: 1 being the worst, 2 being slightly better, 3 – good, 4 – best. Even though that we are still interested only in the class 4, this partitioning allows us to better optimise the performance of the classifier. In cases when the decision is not very certain, usually the confusion is only between two classes (Keerthi et al., 2008). E.g. when two good classes (3 and 4) compete, a mistake is not that bad as e.g. in the case of a good and a bad class (1 and 4), where the risk of making a severe mistake is much higher. In the simple binary case one would not be able to differentiate between different types of mistakes that well.

5 Model

First, we have tried to represent each sentence as a feature vector and then learn the corresponding classes using linear SVMs (Lin et al., 2008; Keerthi et al., 2008). However, either because we were not able to come up with enough good sentence-level features or because the size of the training data was not big enough (we have used 20% of the data left for experimenting, which then again was randomly split into 90% for training and 10% for testing), the results were not satisfactory.

We have then decided to do the classification on the word level: for every word in a sentence of class c , we have constructed a feature vector and assigned it the class c in the training phase. E.g. for a class 3 sentence of length 10 we have constructed

10 feature vectors, one for each word, of the class 3. The feature templates which we used were:

sentence-level features:

length of the sentence, number of punctuation tokens, number of verbs, number of pronouns, number of conjunctions, number of adverbial modifiers

word-level features:

word form of the current word, word form of the next word, POS of the current word, POS of the next word, word length, dependency type, word novelty

These particular features were selected after some semi-automatic feature engineering, during which we have tried all kinds of different POS tags and dependency types as features.

In order to detect punctuation, verbs and pronouns we have simply used the Finnish POS tags: PUNCT, V and PRON respectively. For adverbial modifiers and dependency types in general we have parsed every sentence with MaltParser and used the label predicted by the parser (not the gold standard that we also had). The novelty of a word is a binary feature: for the given word form we look whether it already ever occurred in the training and test data or whether it is new.

In the test phase we would then predict a class for every word in a sentence and perform a voting procedure in order to infer the sentence usefulness class for the whole sentence. Given a set of votes

$$V = \{c_1, c_2, c_3, \dots, c_n\} \text{ , where}$$

$c_n \in \{1, 2, 3, 4\}$, the voting procedure looks like that:

```

if  $4 \in V$ 
    weight(1) = count(1)*2 in V
    weight(2) = count(2) in V
    weight(3) = count(3) in V
    weight(4) = count(4) in V
    return argmax_i(weight(i))
else
    return majority_voting(V)

```

So basically if V does not contain words classified as class 4, we simply perform majority voting for the remaining classes. Otherwise all votes for other classes are counted, whereas the votes for the worst class are always counted twice in order to avoid a severe mistake whenever there is a chance that a sentence might belong to class 1.

6 Evaluation

The evaluation of our model was not straightforward. The first metric which we have tried out was accuracy: the percentage of correctly classified classes. However, since we are not equally interested in all classes, but are rather interested in correctly finding the class 4, the metric was not very helpful.

That is why the second attempt was to compute precision and recall only for the class 4. However, as we have already implied in section 4, different mistakes have different impact on the result. A confusion between class 1 and class 4 is different from a confusion between classes 3 and 4. Precision and recall metrics are thus not helpful either.

The next step was to compute a confusion matrix for all classes and try to minimise the number of actual instances of 1 and 2 predicted as 4 on the one hand and maximising the number of actual 4 predicted as 4 on the other hand.

In a real-world scenario the number of potential sentences which can be annotated is infinite and therefore the recall of the method might be arbitrary low and it will still find enough good candidates to annotate. It is merely important that the precision is high, such that the sentences which are annotated are really beneficial for the task. However, in our experiment the amount of data was small: in 20% of the sentences left for testing, i.e. 250 sentences, 57 belonged to class 4. Therefore we could not decrease recall because otherwise not enough sentences would have been found in order to clearly demonstrate that our approach outperforms the baseline of selecting sentences randomly.

We have run 10 tests randomly selecting 80% sentences for training and 20% for testing and averaged the performance on the confusion matrix:

In ~16% of all cases actual 1s were classified as 4s. In ~26% of all cases actual 2s were classified as 4s, in ~8% of all cases actual 3s were classified as 4s and in ~50% of all cases actual 4s were correctly classified as 4s.

Thus this method allows to select good sentences significantly more often (namely twice as often) than when doing it randomly. Again, if we had more data available for testing, we could tune the precision at costs of recall, improving the performance even further. Our experiment described in the section 3 with gold-standard usefulness

classes, suggests that it is possible to gain accuracy up to four times faster when annotating sentences proposed by our method compared to random selection.

7 Discussion

We have presented an approach which allows to predict sentence usefulness and extend a treebank only with the most promising sentences for the given task. However, we suppose that the sentence usefulness is not a constant parameter, but rather changes with the growing size of the training data. Because of the limited size of data for experimenting we could not investigate whether our metric is still better than random when hundreds or thousands of sentences are added to the original corpus and/or whether the model has to be adapted and re-trained on the new data, because in the course of annotation different units might become useful than it was the case prior to the treebank extension. Most probably the latter case is true and the model has to be regularly adopted to the changes in the treebank, however, we do not know whether it will be an easier or a harder task to predict new good candidates for annotation. On the one hand with the growing size it should become more difficult to find new beneficial sentences for the treebank, but on the other hand with more data available a more accurate model for prediction can be constructed.

Another interesting point is to analyse what actually makes some sentences better than other. It was important to make sure that longer sentences are not always automatically better, simply because they have more tokens and thus provide more information. So in one of experiments we have sorted all sentences according to their usefulness and looked at their length. The result was that even though the best sentences are on average slightly longer, there is still a huge amount of long sentences which were not useful and there are a lot of short sentences that were useful. In principle the approach might be further optimised by not only finding most promising but also short sentences, which do not require much time to annotate.

Unfortunately, we are not speakers of Finnish and we were not able to analyse neither the content of the most useful sentences nor their properties. It would be interesting to find out whether these sentences have certain linguistic structures or contain specific lexical entries. We will certainly try to co-

operate with someone to perform this analysis in the future.

Neither could we investigate whether the useful sentences we are able to detect overlap with those predicted by the approach proposed by Mirroshan-del and Nasr. The approaches are very different: they work in a bottom-up fashion by optimising the model for the sentences with poorest performance until the model performs well for most sentences and we work in a top-down fashion by optimising the model for sentences which improve the performance for as many other sentences as possible in the test data. Intuitively, the first and the latter sentence types are not the same and thus it would also be interesting to combine both approaches.

8 Conclusions

In this paper we have shown the necessity of extending annotated resources on the example of Finnish. Since the annotation process is expensive and tedious we have proposed an approach for selecting only the most promising sentences for annotation. We explain our sentence usefulness metric and how it can be predicted for new sentences. Our experiments show that it is possible to select good sentences significantly better than annotating random sentences. If gold standard values for sentence significance are used, the accuracy of the parser increases four times faster than when random sentences are added to the training data. With the values predicted by our model it is still twice as fast. We have also discussed some potential problems with the metric if the size of the initial treebank changes considerably. However, on the other hand it might become easier to accurately predict good sentences if more data becomes available.

Acknowledgements

The work presented here was partially supported by a research grant from the German Federal Ministry of Education and Research (BMBF) to the DFKI project Deependance (FKZ. 01IW11003) and partially funded by the EU project Excitement (FP7-IST). Additionally, we would like to thank Turku BioNLP group, in particular Filip Ginter and Katri Haverinen for their fantastic support. They have provided us with various versions of the Finnish treebank, replied numerous emails and even responded to some feature requests.

References

- J. Baldridge and M. Osborne. 2008. *Active learning and logarithmic opinion pools for HPSG parse selection.* Natural Language Engineering Journal, 2, pp. 191-222.
- Haverinen, K.; Viljanen, T.; Laippala, V.; Kohonen, S.; Ginter, F. & Salakoski, T. 2010. *Treebanking Finnish.* Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9), pp. 79-90. 2010.
- S. Sathiya Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. 2008. *A sequential dual coordinate method for large-scale multi-class linear SVMs.* KDD 2008.
- C.-J. Lin, R.-E. Fan, K.-W. Chang, C.-J. Hsieh and X.-R. Wang. *LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research 9(2008), pp. 1871-1874.
- Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanov, Gulsen Eryigit, Sandra Kubler, Svetoslav Marinov and Erwin Marsi. 2007. *MaltParser: A Language-Independent System for Data-Driven Dependency Parsing.* Natural Language Engineering Journal, 13, pp. 99-135.
- Seyed Abolghasem Mirroshandel and Alexis Nasr. 2011. Active Learning for Dependency Parsing Using Partially Annotated Sentences. In Proceedings of the 12th International Conference on Parsing Technologies, pages 140–149, October 5-7, 2011, Dublin City University.

Using subcategorization frames to improve French probabilistic parsing

Anthony Sigogne

Université Paris-Est, LIGM
sigogne@univ-mlv.fr

Matthieu Constant

Université Paris-Est, LIGM
mconstan@univ-mlv.fr

Abstract

This article introduces results about probabilistic parsing enhanced with a word clustering approach based on a French syntactic lexicon, the Lefff (Sagot, 2010). We show that by applying this clustering method on verbs and adjectives of the French Treebank (Abeillé et al., 2003), we obtain accurate performances on French with a parser based on a Probabilistic Context-Free Grammar (Petrov et al., 2006).

1 Introduction

Dealing with data sparseness is a real challenge for Probabilistic Context-Free Grammar parsers [PCFG], especially when the PCFG grammar is extracted from a small treebank¹. This problem is also lexical because the richer the morphology of a language is, the sparser the lexicons built from a treebank will be for that language. Nevertheless, the effect of lexical data sparseness can be reduced by word clustering algorithms. Inspired by the clustering method of (Koo et al., 2008), (Candito and Seddah, 2010) have shown that by replacing each word of the corpus by automatically obtained clusters of words, they can significantly improve a PCFG parser on French. Recently, (Sigogne et al., 2011) proposed a clustering method based on a French syntactic lexicon, the Lexicon-Grammar [LG] (Gross, 1994). This method consists in replacing each word of the corpus by the combination of its part-of-speech tag and its cluster, pre-computed from the lexicon. A

cluster corresponds to a class of the lexicon that gathers items sharing several syntactic properties. They applied this method on verbs only and reported significant gains.

In this article, we propose a clustering method of verbs and adjectives based on another French lexicon, the Lefff (Sagot, 2010). This lexicon does not offer a classification of items as in the LG but for each entry, information about subcategorization frame is available. Clusters of words are now computed by aggregating items that have a similar frame, a frame being reduced to a vector of syntactic functions linked to possible syntactic arguments.

In sections 2 and 3, we describe the probabilistic parser and the treebank used in our experiments. In section 4, we describe more precisely previous work on clustering methods. Section 5 introduces the syntactic lexicon, the Lefff, and then we present the clustering approach based on this lexicon. In section 6, we describe our experiments and discuss the obtained results.

2 Berkeley Parser

The probabilistic parser, used in our experiments, is the Berkeley Parser² [BKY] (Petrov et al., 2006). This parser is based on a PCFG model which is non-lexicalized. The main problem of non-lexicalized context-free grammars is that nonterminal symbols encode too general information which weakly discriminates syntactic ambiguities. The benefit of BKY is to try to solve the problem by generating a grammar containing

¹Data sparseness implies the difficulty of estimating probabilities of rare rules extracted from the corpus.

²<http://code.google.com/p/berkeleyparser/>

complex symbols, following the principle of latent annotations introduced by (Matsuzaki et al., 2005). Parameters of the latent grammar are estimated with an algorithm based on Expectation-Maximisation [EM]. In the case of French, (Seddah et al., 2009) have shown that BKY produces *state-of-the-art* performances.

3 French Treebank

For our experiments, we used the French Treebank³ (Abeillé et al., 2003) [FTB]. It is composed of articles from the newspaper *Le Monde* where each sentence is annotated with a constituent tree. Currently, most papers about parsing of French use a specific variant of the FTB, namely the FTB-UC described for the first time in (Candito and Crabbé, 2009). It is a partially corrected version of the FTB that contains 12.351 sentences and 350.931 tokens with a part-of-speech tagset of 28 tags and 12 nonterminal symbols⁴.

4 Previous work on word clustering

Numerous works used a clustering approach in order to reduce the size of the corpus lexicon and therefore reduce the impact of lexical data sparseness on treebank grammars. Several methods have been described in (Candito and Seddah, 2010). The best one, called *Clust*, consists in replacing each word by a cluster id. Cluster ids are automatically obtained thanks to an unsupervised statistical algorithm (Brown et al., 1992) applied to a large raw corpus. They are computed on the basis of word co-occurrence statistics. Currently, this method permits to obtain the best results on the FTB-UC. Recently, (Sigogne et al., 2011) described a method, called *LexClust*, based on a French syntactic lexicon, the Lexicon-Grammar (Gross, 1994), that consists in replacing each verbal form of the corpus by the combination of its POS tag and its cluster. These clusters follow the particular classification of entries offered by this lexicon, that aggregates items sharing several syntactic properties (e.g. subcategorization

³Available under licence at <http://www.11f.cnrs.fr/Gens/Abeille/French-Treebank-fr.php>

⁴There are also 7 possible syntactic functions attached to nonterminal nodes. Those annotations were removed for our experiments.

information). For example, a class of this lexicon, called *3IR*, indicates that all verbs belonging to this class are intransitive. By only modifying the verbs, this approach obtains significant results on the FTB-UC.

5 Word clustering based on a syntactic lexicon

5.1 A syntactic lexicon, the Lefff

The Lefff is a French syntactic and morphological wide-coverage lexicon (Sagot, 2010)⁵ that contains 110.477 lemmatized forms (simple and compound) and 536.375 inflected forms. This lexicon describes for each lemmatized entry a canonical subcategorization frame, composed of all possible arguments of the entry, and a list of possible redistributions from this frame. Inflected entries are built from lemmatized form and for each possible redistribution. For each argument of a subcategorization frame, it is stated the mandatory nature, a syntactic function, syntagmatic productions (pronoun *cln*, noun phrase *np*, infinitive phrase *sinf*, ...), and some semantic features (human, abstract, ...). A syntactic function takes a value among a set of nine functions, *Suj* (subject), *Obj* (direct object), *Obj_a* (indirect object introduced by the preposition *à*), *Obj_{de}* (indirect object introduced by the preposition *de*), *Loc* (locative), *Dloc* (delocative), *Att* (attribute), *Obl* and *Obl₂* (obliques). Figure 1 shows a simplified sample of the Lefff for an entry of the French verb *chérir* (to cherish). The frame of this entry is composed of two arguments, indicated by the two syntactic functions *Suj* and *Obj*. The coverage of the lexicon on the FTB-UC is high, with 99.0% and 96.4% respectively for verbs and adjectives, that are the only two grammatical categories that have available subcategorization frames in the Lefff.

$$\text{chérir} \rightarrow \text{Suj} : (\text{cln}|\text{sinf}|\text{sn}), \text{Obj} : (\text{cln}|\text{sn})$$

Figure 1: Sample of the Lefff for an entry of the verb *chérir* (to cherish).

⁵<http://atoll.inria.fr/~sagot/lefff.html>

5.2 Word clustering based on the Lefff

The clustering method of verbs and adjectives that we propose in this paper follows the principle of the experiment *LexClust*. A word in the corpus is replaced by the combination of its part-of-speech tag and its cluster. These clusters are computed from the Lefff by exploiting subcategorization frames of entries. First, for each lemmatized form of the lexicon, we reduce its frame to the vector of syntactic functions linked to arguments. If a form appears in several entries (depending on meanings), we merge all vectors into a single one. Then, clusters are determined by grouping forms that have the same vector. Vectors are composed of syntactic functions taken from a subset of the seven most frequent ones, *Suj*, *Obj*, *Obj_j*, *Objde*, *Loc*, *Att* et *Obl*. This subset allows for creating less clusters and improving results. Table 1 shows an example of the clustering process on several verbs of the Lefff. Each verb is associated with its vector of syntactic functions and its cluster. In this example, vectors of verbs *abolir* and *cibler* are identical and are composed of a subject and a direct object. Therefore, they belong to the same verb cluster, while other verbs are associated with a distinct cluster. Table 2 shows a similar example for adjective clusters.

Verb	Vector	Cluster
abolir (to abolish)	Suj, Obj	1
cibler (to target)	Suj, Obj	1
prouver (to prove)	Suj, Obj, Obj _j , Obl	2
gratifier (to gratify)	Suj, Obj, Objde	3

Table 1: Verb clusters obtained from the *Lefff*.

Adjective	Vector	Cluster
celtique (celtic)	Suj, Objde, Obj _j	1
censuré (censored)	Suj, Obl2	2
chanceux (lucky)	Suj, Objde, Obj _j	1
lavé (washed)	Suj, Obj, Obl2	2

Table 2: Adjective clusters obtained from the *Lefff*.

However, this approach requires a POS tagger and a lemmatizer in order to analyze a raw text (clusters being determined from lemmatized forms). Therefore, we chose one of the best tagger for French called *LGTagger* (Constant and Sigogne, 2011) which is based on a Conditional Random Field probabilistic model. Lemmatization is made

with the *Bonsai* tool⁶ which is based on the Lefff and some heuristics in case of ambiguities.

6 Experiments and results

6.1 Evaluation metrics

As the FTB-UC is a small corpus, we used a *cross-validation* procedure for evaluation. This method consists in splitting the corpus into p equal parts, then we compute training on $p-1$ parts and evaluations on the remaining part. We can iterate this process p times. This allows us to calculate an average score for a sample as large as the initial corpus. In our case, we set the parameter p to 10. Results on evaluation parts for all sentences are reported using several standard measures, the *F₁score* and *unlabeled attachment* scores. The labeled *F₁score* [F1]⁷, defined by the standard protocol called PARSEVAL (Black et al., 1991), takes into account the bracketing and labeling of nodes. In order to establish the significance of results between two experiments, we used an unidirectional t-test for two independent samples⁸. The *unlabeled attachment score* [UAS] evaluates the quality of unlabeled dependencies between words of the sentence⁹. Punctuation tokens are ignored in all metrics.

6.2 Berkeley parser settings

We used a modified version of BKY enhanced for tagging unknown and rare French words (Crabbé and Candito, 2008)¹⁰. We can notice that BKY uses two sets of sentences at training, a learning set and a validation set for optimizing the grammar parameters. As in (Candito et al., 2010), we used 2% of each training part as a validation set and the remaining 98% as a learning set. The number of split and merge cycles was set to 5. The random seed was set to 8.

⁶<http://alpage.inria.fr/statgram/frdep/>

⁷Evalb tool available at <http://nlp.cs.nyu.edu/evalb/>

⁸Dan Bikel's tool available at <http://www.cis.upenn.edu/~dbikel/software.html>

⁹This score is computed by automatically converting constituent trees into dependency trees. The conversion procedure is made with the *Bonsai* tool.

¹⁰Available in the *Bonsai* package.

6.3 Clustering methods

We evaluated the impact of our clustering method on verbs and adjectives of the FTB-UC (respectively noted *Verb* and *Adj*). Those of each training part are replaced by the corresponding cluster and, in order to do it on the evaluation part, we used LGTagger and a lemmatizer. Tagging TAG and lemmatization LEM accuracies of these tools are reported in the Table 3 according to cross-validation on the FTB-UC. In addition to the overall score for all words in the corpus, F1 score is also reported for verbs and adjectives¹¹. First, we can see that verbs are efficiently tagged and lemmatized. About adjectives, there is a greater number of errors (about 5%), and this is mainly because of the ambiguity involved with the past (31% of all errors).

	All	Verbs	Adjectives
TAG	97.75	97.83	94.80
LEM	96.77	97.15	95.84

Table 3: Tagging and lemmatization accuracies of LGTagger and Bonsaï lemmatizer according to cross-validation on the FTB-UC.

6.4 Results

The experimental results are shown in the Table 4¹². The columns *#cls* and *#lex* respectively indicate the number of created clusters and the size of the FTB-UC lexicon according to clustering methods. Note that all results are significant compared to the baseline¹³ ($t\text{-test} < 10^{-4}$). Absolute gains of experiment *Verb* are about +0.4 for both F1 and UAS. By just modifying verbs, we can drastically reduce the size of the corpus lexicon. About experiment *Adj*, despite lower tagging and lemmatization accuracies, clusters allow to obtain gains of about +0.3 for both F1 and UAS. However, combining *Adj* to *Verb* has no positive effect compared to *Verb* and *Adj*.

So as to compare our results with previous work on word clustering, we report, in Table 5, results of the method *Clust* described in section 4. More-

¹¹We can compute this score because words can be, for example, labeled incorrectly as a verb, or verbs may be labeled incorrectly.

¹²All experiments have a tagging accuracy of about 97%.

¹³Baseline experiment consists in training and evaluating BKY on FTB-UC with original words.

	#cls	#lex	F1	UAS
Baseline	-	27.143	84.03	89.58
Verb	96	20.567	84.44	89.96
Adj	16	23.982	84.30	89.79
Verb+Adj	112	17.108	84.42	89.92

Table 4: Results from cross-validation evaluation according to our clustering methods.

over, we tried some combination of methods *Verb*, *Adj* and *Clust*. In this case, *Clust* only replaces words of other grammatical categories.

	#cls	#lex	F1	UAS
Verb	96	20.567	84.44	89.96
Clust	1000	1.987	85.25	90.42
Verb+Clust	1096	2.186	85.13	90.25
Verb+Adj+Clust	1112	730	84.93	89.98

Table 5: Results from cross-validation evaluation according to our clustering methods.

We can see that *Clust* obtains the best scores, with an absolute gain of +0.9 for F1 and +0.4 for UAS compared to *Verb*. Nevertheless, we obtain similar results to *Clust* when our verb clusters are combined with method *Clust*, applied on all other words of the corpus ($t\text{-test} > 0.2$). Therefore, it would mean that verb clusters computed from a lexicon are as powerfull as clusters from a statistical model.

7 Conclusion

In this article, we have shown that by using information about verbs (and to a lesser extent, adjectives) from a syntactic lexicon, the Lefff, we are able to improve performances of a statistical parser based on a PCFG grammar. In the near future, we plan to reproduce experiments with other grammatical categories like nouns available in other French lexicons.

References

- A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, Kluwer, Dordrecht.
- E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic cov-

- erage of english grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*.
- P. F. Brown, V. J. Della, P. V. Desouza, J. C. Lai, and R. L. Mercer. 1992. Class-based n-gram models of natural language. In *Computational linguistics*, 18(4), pages 467–479.
- M. Candito and B. Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of IWPT'09*, pages 138–141.
- M. Candito and D. Seddah. 2010. Parsing word clusters. In *Proceedings of SPMRL'10*, pages 76–84.
- M. Candito, B. Crabbé, and P. Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC10*.
- M. Constant and A. Sigogne. 2011. MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources. In *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*, France.
- B. Crabbé and M. Candito. 2008. Expériences d’analyse syntaxique statistique du français. In *Proceedings of TALN'08*.
- M. Gross. 1994. Constructing Lexicon-grammars. In Atkins and Zampolli, editors, *Computational Approaches to the Lexicon*, pages 213–263.
- T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08*.
- T. Matsuzaki, Y. Miyao, and J. Tsujii. 2005. Probabilistic CFG with latent annotations. In *Proceedings of ACL-05*, pages 75–82, Ann Arbor, USA.
- S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of ACL'06*.
- B. Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC'10*.
- G. Sampson and A. Babarczy. 2003. A test of the leaf-ancestor metric for parsing accuracy. In *Natural Language Engineering*, 9 (4), pages 365–380.
- D. Seddah, M. Candito, and B. Crabbé. 2009. Adaptation de parsers statistiques lexicalisés pour le français : Une évaluation complète sur corpus arborés. In *Proceedings of TALN'09*, Senlis, France.
- A. Sigogne, M. Constant, and E. Laporte. 2011. French parsing enhanced with a word clustering method based on a syntactic lexicon. In *Proceedings of SPMRL'11*, pages 22–27, Dublin, Ireland.

Statistical Denormalization for Arabic Text

Mohammed Moussa, Mohamed Waleed Fakhr

College of Computing and Information Technology,
Arab Academy for Science and Technology,
Heliopolis, Cairo, Egypt
mohammed.moussa@live.com,
waleedf@aast.edu

Kareem Darwish

Qatar Computing Research Institute,
Qatar Foundation, Doha, Qatar
kdarwish@qf.org.qa

Abstract

In this paper, we focus on a sub-problem of Arabic text error correction, namely Arabic Text Denormalization. Text Denormalization is considered an important post-processing step when performing machine translation into Arabic. We examine different approaches for denormalization via the use of language modeling, stemming, and sequence labeling. We show the effectiveness of different approaches and how they can be combined to attain better results. We perform intrinsic evaluation as well as extrinsic evaluation in the context of machine translation.

1 Introduction

Arabic Text Denormalization (ATD) is considered a sub-problem of Automated Text Error Correction (TEC), which is an important topic in Natural Language Processing (NLP). TEC can be used in many applications such as OCR error correction, query-spelling correction, or as pre- or post- processing for other NLP tasks, such as Machine Translation (MT). For example, the training of an MT system that translates between Arabic and other languages is typically improved by normalizing some of the Arabic letters that replace each other depending on context or are commonly confused by document authors. When translating into Arabic, these letter normalizations need to be de-normalized to recover the proper forms of the letters. We test ATD in the context of a:

1. Standalone system for correcting common Arabic mistakes, which are made by users who are not linguistically proficient or who write casually, e.g. bloggers, tweeters, etc. It is also helpful for users who need an automatic system to help them identify such spelling mistakes.

2. Denormalizing MT output when translating into Arabic. We evaluated English to Arabic MT output against properly spelled output and we achieved a BLEU score of 8.78. Upon inspecting the output we saw that normalization during training contributed the most to the low score.

The specific letter normalizations that we will address in ATD are as follows:

1. Restoring ‘ه p’ or ‘ه h’ from ‘ه h’ at the end of a word¹:

Normalized	Denormalized
سنه snh	سنة snp (year); سنه snh (his age)
هذه h*h	هذه h*h (this fm.)

2. Restoring ‘ي Y’ or ‘ي y’ from ‘ي y’ at the end of a word:

Normalized	Denormalized
wHdY	وحدى wHdy (alone 1 st person)
قصوى qSwY	قصوى qSwY (maximum)

3. Restoring ‘ا <’, ‘ا >’, ‘ا |’ or ‘ا A’ from ‘ا A’:

Normalized	Denormalized
اسلام AslAm	إسلام<slam (Islam)
ارض ArD	أرض>rD (land)
ات At	أت t (coming)
قال qAl	قال qAl (he said)

These constitute normalizations that are commonly performed for machine translation and information retrieval. Another less common normalization entails conflating ‘و’ & ‘ء’, ‘ع’ , and ‘ئ’ } (Darwish and Ali, 2012).

The problem is that a normalized form of an Arabic word may be denormalized into multiple different forms depending on context. For example, both قرآن ‘qrAn’ (Qur'an) and قرآن ‘qrAn’ (marriage) are normalized to قرآن ‘qrAn’.

We used two main approaches to solve the problem, namely: using language modeling at word and stem levels; and using Conditional Random Fields (CRF) sequence labeling to handle

¹ We use Buckwalter transliteration throughout the paper

cases not disambiguated using language modeling. We also examined the use of a CRF labeler in isolation of language modeling for comparison.

The main contributions of the paper are:

1. Using stemming in conjunction with language modeling to improve language modeling coverage.
2. Treating the ATD problem as a sequence labeling problem.
3. Using a combined system to achieve a denormalization accuracy greater than 99%.

2 Related Work

Spelling correction is a well-studied problem (Kukich, 1992; Manning and Schutz, 1999). The problem of detecting and correcting misspelled words in text usually involves finding out-of-vocabulary words then finding most similar words in a dictionary using some measure of distance (Levenshtein, 1966; Wagner and Fisher, 1974). Heuristic approaches are also used as in (Shaalan et al., 2003) to find replacement candidates for the misspelled word by adding or removing letters, or splitting words. A finite-state automaton based approach proposed by (Hassan et al., 2008) models letter mapping probabilities and letter and word sequence probabilities.

Though there are commercial systems that perform such denormalization as part of their pipelines, the literature is quite scant on denormalization. (El-Kholi and Habash, 2010) used the MADA analyzer to perform de-tokenization and denormalization.

3 Data and Tools

3.1 Training Data

In our experiments, we used 8 million Arabic sentences, containing 182 million words, from Aljazeera.net news articles to train our language models. Aljazeera.net has very high editorial standards, making spelling mistakes very rare. We constructed language models as follows:

- Word-level models where Arabic text was properly tokenized, and all diacritics (short Arabic vowels), kashidas (word elongations), and numbers were removed.

- Stem-level models where words were stemmed using a statistical stemmer that is akin to AMIRA (Mona Diab, 2009). However we did not remove both ‘s p’ and ‘s h’ from the end of words as they are letters of interest.

We built our language models using the SRILM toolkit with Good-Turing smoothing (Stolcke, Andreas, 2002).

We trained a character-level CRF model using 5 thousand and 50 thousand sentences. We henceforth refer to this model as the ‘CRFModel’. We used the CRF++ implementation (Kudo, 2009) of CRF for all our experiments. The features that we used for the CRF character-level model were as follows:

- Features 1 to 9: Current letter, preceding 4 letters and following 4 letters, each as a feature. This serves as a character language model.
- Features 10 and 11: letter bigram features, namely the current letter with preceding letter, and current letter with following letter.
- Features 11 to 13: letter trigram features, namely current letter with 2 preceding letters, current letter with 1 preceding and 1 following letter, and current letter with 2 following letters
- Features 14 to 17: letter 4-gram features
- Features 18 to 22: letter 5-gram features. Features 10-22 further model Arabic letter sequences, by attempting to capture common 2, 3, 4, or 5 letter sequences that may have consistent denormalization patterns. For example, the sequence ‘ج Al’ is most likely a determiner that has a single form.
- Features 23 and 24: position of the letter from the beginning and end of a word respectively.

The output labels of the CRFModel are the proper denormalized form of the letter. For example, in case of input ‘ا A’ the output could be one of the 4 classes ‘ا <’, ‘ا >’, ‘ا |’ or ‘ا A’, in case of input ‘ي y’ the output could be one of the two classes ‘ي y’ or ‘ي Y’ and in case of ‘ه h’ the output could be one of the 2 classes ‘ه p’ or ‘ه h’. The output for the remaining letters would be ‘S’ (standing for same letter). Thus we had 9 output labels in all.

3.2 Test Data

To test the effectiveness of ATD, we used a test set of three thousand sentences, containing

106,859 words. The test sentences were obtained from islamonline.net, an online site, and were manually checked for errors. Testing involved attempting to perform ATD on a normalized version of the sentences. Around 75% of words required denormalization. Based on the unigram word-level model, 0.8%, 41.8%, 34.3%, 7.5% and 15.6% of the words in the test set have 0, 1, 2, 3 and more than 3 denormalized forms respectively.

To test the effectiveness of ATD in the context of machine translation, we used a parallel English-Arabic test set of 4 thousand sentences containing 36,839 words. We used the Bing online translator to perform MT from English to Arabic. We used BLEU with a single reference translation as the measure of effectiveness. We used two baselines, namely: the output of the Bing system, where the Bing translator performs some sort of ATD, and the same output with an additional letter normalization step. In discussions with the team that worked on the English to Arabic translation in Bing, they indicated that they are using a proprietary denormalization component.

4 ATD Experimental Setups

We used several experimental setups as follows:

1. Unigram LM: In this setup, we simply picked the most common denormalized form of a word regardless of context. If a word is Out Of Vocabulary (OOV), meaning not seen in training, it is left as is. We consider this as our baseline experiment. The approach has the following disadvantages:

- It ignores contextual ambiguity. For example, though the normalized form علی Ely has the possible denormalized forms: {علی Ely “proper name Ali”, على EIY “on”}, the second form will consistently be chosen.
- Coverage is limited by previously seen words.

2. Unigram Stem LM: Since attached clitics in Arabic typically have 1 form, then the stem (without prefixes and suffixes) is usually the portion of the word that requires denormalization. This setup is identical to the Unigram LM, but denormalization is done at stem level. The advantage of this approach is that it should have better coverage than the Unigram LM. However,

it does not take context into account. Also, ambiguity is increased, because attached clitics often disambiguate the correct denormalized form.

3. Unigram LM + Unigram Stem LM: In this setup, we used the Unigram LM setup for all words, and we backed-off to Unigram Stem LM for OOV words. This has the effect of increasing coverage, while using the disambiguation information of clitics. It still ignores context.

4. Bigram LM: In this setup, we generated all known denormalization of a word, and then we used the Viterbi algorithm (bigram model) to ascertain the best denormalized form in context. OOV words were left unchanged. This setup uses context to pick the best denormalization, but it is limited by the previously seen words.

5. Bigram Stem LM: This is identical to Bigram LM, except that the language model is constructed on stems and not words.

6. Bigram LM + Unigram Stem LM: This is identical to Bigram LM, but with back-off to the Unigram Stem LM. This accounts for context and backs-off to better handle OOV words.

7. Bigram LM + Bigram Stem LM: This is identical to Bigram LM, but with back-off to the Bigram Stem LM.

8. CRF Model: We trained the CRF sequence labeler using the aforementioned features. We used the generated CRF model in two ways:

- a. As a back-off to handle OOV words after we apply the entire language model based approaches.
- b. As a standalone approach that attempts to denormalize directly.

5 ATD Experimental Results

5.1 Intrinsic ATD Evaluation

Table 1 reports on the results of using the different language modeling based approaches. Table 2 reports CRFModel as a standalone approach with two different data sizes. Table 3 reports on the same approaches but with CRF-based back-off for OOV words.

Not surprisingly, the results show that using a bigram language model generally produces slightly better accuracy than using a unigram model. Using a stem language model helps only when being used as back-off and that appears clearly for words with more than 1 candidate. This can be explained by the fact that attached clitics can help disambiguate the correct denormalized form. (Results in Table 1 show that 73.7% of OOV's were unchanged after proper denormalization).

Also not surprisingly Table 2 shows that using more data to train CRFModel leads to better accuracy. We chose to use the better CRFModel combined with the language models to report the results in Table 3. Stem and CRF models help in handling OOV's. Table 4 and 5 shows how well these models perform on the words that are left over from word and stem-level models. The CRF model was effective in guessing the proper denormalization for more than 87% of the words.

Candidates/word in LM	All	0	1	> 1
% of test data	100	0.8	41.8	57.4
Setup	Accuracy (%)			
1. Unigram LM	98.2	73.7	99.9	97.3
2. Unigram Stem LM	97.6	86.7	99.6	96.3
3. Unigram LM + Unigram Stem LM	98.3	86.7	99.9	97.3
4. Bigram LM	98.9	73.7	99.9	98.4
5. Bigram Stem LM	98.6	86.4	99.7	97.9
6. Bigram LM + Unigram Stem LM	99.0	86.7	99.9	98.4
7. Bigram LM + Bigram Stem LM	99.0	86.4	99.9	98.4

Table 1: Results of using language modeling for ATD

Candidates/word in LM	All	0	1	> 1
% of test data	100	0.8	41.8	57.4
Setup	Data	Accuracy (%)		
8. CRF standalone	5k	95.1	87.5	97.0
	50k	97.0	87.7	98.2

Table 2: CRFModel w/ training sets of different sizes.

5.2 ATD Results in MT

Table 6 reports on the BLEU scores for translating 4 thousand sentences from English to Arabic and then performing denormalization using the

different approaches. Table 6 reports on two baselines. The first involves not using denormalization at all and the other relies on the denormalization of Bing online translator system. The results show that using our best ATD system edges the Bing system, but the difference is not statistically significant. Using CRF model alone yields results that are 0.38 BLEU points lower than the best system. This shows that even a 2% drop in ATD accuracy may noticeably adversely impact translation quality. When comparing with the Bing translation system, which is nearly state-of-the-art, our proposed ATD system is at par with it. Note that the Bing system has an advantage over our proposed system in that the MT system does not have OOVs in the denormalization phase because it only generates Arabic words that appear in training.

Candidates/word in LM	All	0	1	> 1
% of test data	100	0.8	41.8	57.4
Setup	Accuracy (%)			
1. Unigram LM	98.3 +0.1	88.7 +15.0	99.9	97.3
2. Unigram Stem LM	97.7 +0.1	93.0 +6.3	99.6	96.3
3. Unigram LM + Unigram Stem LM	98.4 +0.1	93.0 +6.3	99.9	97.3
4. Bigram LM	99.0 +0.1	88.7 +15.0	99.9	98.4
5. Bigram Stem LM	98.6 0.0	92.7 +6.3	99.7	97.9
6. Bigram LM + Unigram Stem LM	99.0 0.0	93.0 +6.3	99.9	98.4
7. Bigram LM + Bigram Stem LM	99.0 0.0	92.7 +6.3	99.9	98.4

Table 3: Results of using language modeling with CRF back-off with relative change over results in Table 1

Setup	Coverage (%)	Accuracy (%)
Unigram Stem LM		97.0
Bigram Stem LM	54.6	96.7

Table 4: Coverage of stem-based models on OOVs

Setup	Accuracy (%)
Word-Based	88.7
Stem-Based	87.5

Table 5: Accuracy of CRF model on OOVs of word-based models and combined stem-based models

Denormalizer System	BLEU non-CRF	BLEU w/CRF
Without ATD	8.78	
Bing Translator	20.79	
Unigram LM	20.75	20.77
Unigram Stem LM	20.64	20.65
Unigram + Stem Unigram LMs	20.76	20.77
Bigram LM	20.80	20.82
Bigram Stem LM	20.76	20.77
Bigram + Stem Unigram LMs	20.81	20.82
Bigram + Stem Bigram LMs	20.81	20.82
CRF Standalone		20.44

Table 6: Results for using ATD in MT

6 Conclusion

In this paper, we presented different approaches for performing automatic denormalization of Arabic text to overcome common spelling mistakes and to recover from the normalization that is typically done while training MT systems that translate into Arabic. The different approaches used word language modeling with back-off to a stem-based language models and a CRF model. We tested the different approaches on naturally occurring Arabic text and we evaluated their effectiveness intrinsically and extrinsically in the context of MT. The best technique according to our experiments is a bigram word-level language model with cascaded back-off to a unigram stem language model and then a CRF model to handle the OOVs.

References

- A. El-Kholi and N. Habash. 2010. *Techniques for Arabic morphological detokenization and orthographic denormalization*. In Proceedings of Language Resources and Evaluation Conference (LREC)
- A. Hassan, S. Noeman and H. Hassan. 2008. *Language independent text correction using finite state automata*. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP).
- A. Stolcke. 2002. *SRILM – an extensible language modeling toolkit*. In Proceedings of the International Conference on Spoken Language Processing.
- C.D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- K. Darwish and A. Ali. 2012. *Arabic Retrieval Revisited: Morphological Hole Filling*. The Association of Computational Linguistics (ACL)
- K. Kukich. 1992. *Techniques for automatically correcting word in text*. ACM Computing Surveys.
- K. Shaalan, A. Allam and A. Gomah. 2003. *Towards Automatic Spell Checking for Arabic*. In Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE).
- M. Diab. 2009. *Second generation tools (AMIRA 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking*. In Proceedings of 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt
- R.A. Wagner and M.J. Fischer. 1974. *The String-to-String Correction Problem*. Journal of the ACM
- T. Kudo. 2009. *CRF++: Yet another CRF toolkit*. <http://crfpp.sourceforge.net>.
- V.I. Levenshtein. 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady.

Automatic Identification of Language Varieties: The Case of Portuguese

Marcos Zampieri

University of Cologne

Albertus-Magnus-Platz 1, 50931

Cologne, Germany

mzampier@uni-koeln.de

Binyam Gebrekidan Gebre

Max Planck Institute for Psycholinguistics

Wundtlaan 1, 6525 XD

Nijmegen, Holland

bingeb@mpi.nl

Abstract

Automatic Language Identification of written texts is a well-established area of research in Computational Linguistics. State-of-the-art algorithms often rely on n-gram character models to identify the correct language of texts, with good results seen for European languages. In this paper we propose the use of a character n-gram model and a word n-gram language model for the automatic classification of two written varieties of Portuguese: European and Brazilian. Results reached 0.998 for accuracy using character 4-grams.

1 Introduction

One of the first steps in almost every NLP task is to distinguish which language(s) a given document contains. The internet is an example of a large text repository that contains languages that are often unidentified. Computational methods can be applied to determine a document's language before undertaking further processing. State-of-the-art methods of language identification for most European languages present satisfactory results above 95% accuracy (Martins and Silva, 2005).

This level of success is common when dealing with languages which are typologically not closely related (e.g. Finnish and Spanish or French and Danish). For these language pairs, distinction based on character n-gram models tends to perform well. Another aspect that may help language identification is the contrast between languages with unique character sets such

as Greek or Hebrew. These languages are easier to identify if compared to language pairs with similar character sets: Arabic and Persian or Russian and Ukrainian (Palmer, 2010).

Martins and Silva (2005) present results on the identification of 12 languages by classifying 500 documents. Results varied according to language ranging from 99% accuracy for English to 80% for Italian. The case of Italian is particularly representative of what we propose here: among 500 texts classified, 20 were tagged as Portuguese and 42 as Spanish. Given that Italian, Portuguese and Spanish are closely related Romance languages, it is evident why algorithms have difficulty classifying Italian documents.

This example shows that a seemingly simple distinction task gains complexity when used to differentiate languages from the same family. In this study, we aim to go one step further and apply computational methods to identify two varieties of the same language: European and Brazilian Portuguese.

2 Related Work

The problem of automatic language identification is not new and early approaches to it can be traced back to Ingle (1980). Ingle applied Zipf's law distribution to order the frequency of stop words in a text and used this information for language identification. Ingle's experiments are different from those used in state-of-the-art language identification, which relies heavily on n-gram models and statistics applied to large corpora.

Dunning (1994) was one of the first to use character n-grams and statistics for language identi-

fication. In this study, the likelihood of n-grams was calculated using Markov models and this was used as the key factor for identification. After Dunning, other studies using n-gram models were published such as (Cavnar and Trenkle, 1994), which developed a language identification tool called TextCat¹ (Grafenstette, 1995), and more recently (Vojtek and Belikova, 2007).

Given its vast amount of multilingual material, the Internet became an important application of language identification. Documents are often unidentified regarding source language and the same document may contain more than one language. Examples of language identification applied to Internet data are (Martins and Silva, 2005) and later (Rehurek and Kolkus, 2009).

2.1 Identifying Similar Languages

If general purpose methods for automatic language identification were substantially explored, the same is not true for methods designed to deal specifically with similar languages or varieties. The identification of closely related languages seems to be the weakness of most n-gram based models and there are few recent studies published about it.

Recently this aspect of language identification received more attention, including a study by Ljubesic et al. (2007). Ljubesic et al. propose a computational model for identification of Croatian texts in comparison to other Slavic languages, namely: Slovenian, Bosnian and Serbian. The study reports 99% recall and precision in three processing stages.

The distinction between languages, dialects or varieties can be political. Serbian, Bosnian and Croatian were all variants of the Serbo-Croatian language spoken in the former Yugoslavia. After their independence, each of these countries adopted their variety as a national language. In the case of Portuguese, although there are substantial differences between Brazilian and European Portuguese, it is widely accepted that these two are varieties of the same language.

Another study worth mentioning is Piersman et al. (2010) on lexical variation. Piersman et al. applied distributional lexical semantics to synonymy retrieval and it was also used for the identi-

fication of distinctive features (which the authors call lectal markers) in Dutch and Flemish. Experiments measuring lexical variation, focusing on convergences and divergences in lexicons, were recently carried out for Brazilian and European Portuguese by (Soares da Silva, 2010).

A couple of recent studies for spoken Portuguese were published by scholars like (Rouas et al., 2008) and later (Koller et al., 2010). These studies model the substantial phonetic differences among Brazilian, African and European Portuguese and discussed how to distinguish them automatically. For written Portuguese, however, to our knowledge there have been no studies published.

The experiments presented here can open new research perspectives in two areas: contrastive linguistics and NLP. For contrastive linguistics, our experiments provide a quantitative estimation on the differences between the two varieties, hence the three groups of features used: lexis-syntactical, lexical and orthographical. Secondly, in real-world NLP applications. Brazilian and European Portuguese do not share a common orthography and identifying the variety of a Portuguese text will help NLP tasks such as spell checking.

3 Linguistic Motivation

Although they are considered to be the same language, there are substantial differences between European and Brazilian Portuguese in terms of phonetics, syntax, lexicon and orthography. For the analysis of written texts, differences in syntax, lexicon and orthography were considered.

Orthography in these language varieties differs in two main aspects: graphical signs and mute consonants. Due to phonetic differences, Brazilian and European Portuguese use different orthographical signs for the same word, such as:

- *econômico* (BP); *económico* (EP): *economic* (EN)

Mute consonants are still used in the Portuguese orthography and are no longer used in Brazil:

- *ator* (BP); *actor* (EP); *actor* (EN)

¹<http://odur.let.rug.nl/vannoord/TextCat/>

Differences also appear at the syntactic level. Some contractions are only used in one of the varieties; for instance: *mo* (pronoun *me* + definite masculine article *o*) is exclusive to Portugal. Past verb tenses (perfect and imperfect) are used in different contexts. The use of pronouns also differs, the Brazilian variety tends to prefer the pronoun before the verb whereas the European variety uses it primarily afterwards:

- *eu te amo* (BP) and *eu amo-te* (EP): *I love you* (EN)

Lexical variation is also a distinctive characteristic of these varieties. Some words are frequent in one of the varieties and rare in the other: *nomeadamente* (EP), *namely* (EN) is widely used in Portugal and rare in Brazil. Additionally, there are cases in which each variety may heavily favor a different word in a set of synonyms, such as: *coima*(EP), *multa* (BP), *fine*, *penalty* (EN).

4 Methods

In order to create a reliable classification method to distinguish Brazilian and European Portuguese, we compiled two journalistic corpora containing texts from each of the two varieties. The Brazilian corpus contains texts published in 2004 by the newspaper *Folha de São Paulo* and the Portuguese corpus contains texts from *Diário de Notícias*, published in 2007. Texts were pre-processed using Python scripts: all meta-information and tags were removed.

The length of texts in the corpora varies and we therefore classified them and grouped them together according to their length in tokens. Language identification and classification tasks tend to be favoured when using large documents and we explore this variable in section 5.2.

4.1 Experiments

The features used take into account differences in lexicon, syntax and orthography. For the orthographical differences, we used character n-grams ranging from 2 to 6-grams. At the lexical level identification was performed using word uni-grams and finally, to explore lexico-syntactical differences we used word bi-grams. The language models were calculated using the Laplace proba-

bility distributions using a function available in NLTK (Bird et al., 2009) as shown in equation 1:

$$P_{lap}(w_1 \dots w_n) = \frac{C(w_1 \dots w_n) + 1}{N + B} \quad (1)$$

In the equation number 1: C is the count of the frequency of w_1 to w_2 in the training data, N is the total number of n-grams and B is the number of distinct n-grams in the training data. For probability estimation, we used the log-likelihood function (Dunning, 1993) represented in equations 2 and 3:

$$P(L|text) = argmax_L P(text|L)P(L) \quad (2)$$

$$P(L|text) = argmax_L \prod_{i=1}^N P(n_i|L)P(L) \quad (3)$$

Equation 3 is a detailed version of equation number 2 and it shows how the classification is made. N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability (log-likelihood) for each of the language models. The language model with higher probability determines the identified language of the text.

5 Results

Evaluation was done using each of the feature groups in a set of 1,000 documents sampled randomly. The sample contains 50% of the texts from each variety and it is divided into 500 documents for training and 500 for testing. We report results in terms of accuracy calculated as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (4)$$

The formula can be interpreted as the number of instances correctly classified ($tp + tn$) divided by all instances classified.

5.1 Word Uni-Grams

The word uni-gram features were used to perform classification taking into account lexical differences. Accuracy results are reported using texts of maximum 300 tokens each.

Max. Len.	Accuracy
300 words	0.996

Table 1: Word Uni-Gram Results

Proper nouns play an important role when using word uni-grams. It is very likely that texts from Portugal will contain named entities that are almost exclusively used in Portugal and vice-versa (e.g. names of important or famous people from Brazil/Portugal or names of companies).

5.2 Word Bi-Grams

For the word bi-gram model evaluation, we explored how the maximum length of texts affects the performance of the algorithm. The results were classified according to the maximum text size, ranging from 100 words to 700 words. The best results were reached with a maximum length of 500 words, after that, the model seems to indicate saturation, as can be seen in table 2:

Max. Len.	Accuracy
100 words	0.851
200 words	0.886
300 words	0.889
400 words	0.904
500 words	0.912
600 words	0.912
700 words	0.905

Table 2: Text Size and Word N-Grams

Results from table 2 are presented in figure number 1:

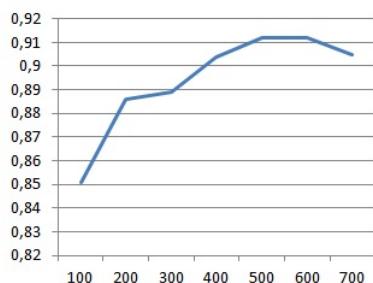


Figure 1: Text Size and Word Bi-Grams

One explanation for the results is that only a few journalistic texts in both corpora are larger than

500 words. Adding these few texts to the classification brings no improvement in the algorithm's performance.

5.3 Character N-Grams

The best results obtained in our experiments relied on a character n-gram model and were designed using texts of maximum 300 tokens. Results reached 0.998 accuracy for 4-grams, and they are presented in table number 3:

N-Grams	Accuracy
2-Grams	0.994
3-Grams	0.996
4-Grams	0.998
5-Grams	0.988
6-Grams	0.990

Table 3: Character N-Gram Results

The good results obtained by character n-grams in comparison to the word n-gram models presented in the previous sections, indicate that the orthographical differences between Brazilian and European Portuguese are still a strong factor for distinguishing these varieties.

6 Conclusions and Future Work

This paper explored computational techniques for the automatic identification of two varieties of Portuguese. From the three groups of features tested, the character-based model using 4-grams performed best. The results for word bi-grams were not very good reaching an accuracy result of 0.912. These outcomes suggest that for this task lexico-syntactic differences are not as important as differences in orthography and lexicon.

The small number of classes contributes to the encouraging results presented here. When performing binary classification, the baseline result is already 0.50 and when provided with meaningful features, algorithms are quite likely to achieve good results.

Experiments are being carried out to integrate the character-based model described in this paper in a real-world language identification tool. Preliminary results for this model in a 6-fold classification reached above 90% accuracy and f-measure.

Acknowledgments

The authors express gratitude to *Folha de São Paulo* for the Brazilian corpus and Sascha Diwersy for the European corpus. Many thanks to the anonymous reviewers who provided important feedback to increase the quality of this paper.

References

- Bird, S.; Klein, E.; Loper, E. 2009. *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit* O'Reilly Media
- Cavnar, W.; Trenkle, J. 1994. N-gram-based Text Categorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*
- Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics - Special Issue on Using Large Corpora* Volume 19, Issue 1. MIT Press
- Dunning, T. 1994. Statistical Identification of Language *Technical Report MCCS-94-273* Computing Research Lab, New Mexico State University
- Grafenstette, G. 1995. Comparing two Language Identification Schemes. *JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data* Rome
- Ingle, N. 1980. *Language Identification Table* Technical Translation International
- Koller, O.; Abad, A.; Trancoso, I.; Viana, C. 2010. Exploiting Variety-dependent Phones in Portuguese Variety Identification Applied to Broadcast News Transcription. *Proceedings of Interspeech 2010*
- Ljubesic, N.; Mikelic, N.; Boras, D. 2007. Language Identificaiton: How to Distinguish Similar Languages? *Proceedings of the 29th International Conference on Information Technology Interfaces*
- Martins, B.; Silva, M. 2005. Language Identification in Web Pages *Proceedings of the 20th ACM Symposium on Applied Computing (SAC), Document Engineering Track* Santa Fe, EUA. p. 763-768
- Palmer, D. 2010. Text Processing. In: Indurkhy, N.; Damerau, F. (Ed.) *Handbook of Natural Language Processing - Second Edition..* CRC Press. p. 9-30
- Piersman, Y.; Geeraerts, D.; Spelman, D. 2010. The Automatic Identification of Lexical Variation Between Language Varieties. *Natural Language Engineering 16 (4)..* Cambridge University Press. p. 469-491
- Rehurek, R.; Kolkus, M. 2009. Language Identification on the Web: Extending the Dictionary Method. *Proceedings of CICLing. Lecture Notes in Computer Sciences*. Springer. p. 357-368
- Rouas, J.; Trancoso, I.; Ribeiro, M.; Abreu, M. 2008. Language and Variety Verification on Broadcast News for Portuguese. *Speech Communication vol. 50 n.11*. Elsevier
- Soares da Silva, A. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese: endo/exogeneous and foreign and normative influence. *Advances in Cognitive Sociolinguistics*. Berlin. De Gruyter
- Vojtek, P.; Belikova, M. 2007. Comparing Language Identification Methods Based on Markov Processes. *Slovko, International Seminar on Computer Treatment of Slavic and East European Languages*

Extending the STTS for the Annotation of Spoken Language

Ines Rehbein

SFB 632 ‘Information Structure’
German Departement
Potsdam University
`{irehbein, soeren.schalowski}@uni-potsdam.de`

Sören Schalowski

SFB 632 ‘Information Structure’
German Departement
Potsdam University
`{irehbein, soeren.schalowski}@uni-potsdam.de`

Abstract

This paper presents an extension to the Stuttgart-Tübingen TagSet, the standard part-of-speech tag set for German, for the annotation of spoken language. The additional tags deal with hesitations, backchannel signals, interruptions, onomatopoeia and uninterpretable material. They allow one to capture phenomena specific to spoken language while, at the same time, preserving inter-operability with already existing corpora of written language.

1 Introduction

Language resources annotated with part-of-speech (POS) information are a valuable resource for linguistic studies as well as for research in the humanities in general. Most existing corpora for German, however, include only written language data, often from the domain of newspaper text.

Recent years have seen an increasing interest in building language resources with data from a variety of domains like spoken language, historical language or computer-mediated communication. This has started a discussion on best practices for annotating and processing *non-canonical* language,¹ where *non-canonical* refers to all kinds of language data which deviate from standard written text. Important issues which have been ad-

¹See, e.g., the workshop on *Annotation of Corpora for Research in the Humanities* (ACRH), the LREC 2012 workshops *Best Practices for Speech Corpora in Linguistic Research*, *Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects*, *NLP can u tag #user-generated content?!*, or the NAACL 2012 workshop on *Syntactic Analysis of Non-canonical Language*.

dressed are the need for normalisation a) to enable corpus searches for all (pronunciation or spelling) variants of one word token, and b) to support the use of off-the-shelf NLP tools developed for written text. Another topic of discussion is the adequacy of existing annotation schemes for new types of data (e.g. a new POS tag set for annotating Twitter data (Gimpel et al., 2011)).

The main objective for using existing annotation schemes for annotating a new variety of data is *inter-operability* with existing resources. There are two aspects of inter-operability. First, we want to be able to use different corpora in cross-linguistic studies and compare results obtained from different corpora, which is only possible if all resources employ the same annotation scheme. Second, we would also like to use existing off-the-shelf NLP tools for the semi-automatic annotation of new data, which again would not be possible when using newly developed annotation schemes for which no training data is available.

We acknowledge the importance of the first objective while, at the same time, arguing for the need to provide a more adequate linguistic description of spoken language phenomena on the POS level. Areas of application are linguistic investigations of e.g. communication strategies or disfluencies in language production, amongst others. We thus propose an extension to an existing tag set with additional tags for phenomena not yet covered by the annotation scheme. This approach guarantees the comparability with other corpora using the original tag set while providing the means for a more adequate description of spoken language.

2 Related Work

Previous work on POS tagging spoken language mostly relies on existing annotation schemes developed for written language, using only minor additions (if any). The Switchboard corpus (Godfrey et al., 1992) provides a fine-grained annotation of disfluencies in spoken dialogues. On the POS level, however, the annotations do not distinguish between interjections, backchannel signals, answer particles, filled pauses or other types of discourse particles. The same is true for the Corpus of Spoken Netherlands (CGN) (Schuurman et al., 2003) and the spoken part of the BNC (Burnard, 2007). Nivre and Grönqvist (2001) extend a tagset developed for written Swedish with two tags designed for spoken language (*feedback* for answer particles and adverbs with similar function, and *own communication management* for filled pauses).

The only linguistically annotated, publicly available corpus of spoken German we are aware of is the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000). The TüBa-D/S was created in the VerbMobil project (Wahlster, 2000) and is annotated with POS tags and syntactic information (phrase structure trees, grammatical dependencies and topological fields (Höhle, 1998)).

2.1 POS annotation in the TüBa-D/S

The TüBa-D/S uses the Stuttgart-Tübingen TagSet (STTS) (Schiller et al., 1995), the standard POS tag set for German which was also used (with minor variations) in the creation of the three German newspaper treebanks, NEGRA (Skut et al., 1998), TIGER (Brants et al., 2002) and TüBa-D/Z (Telljohann et al., 2004).

There are a number of phenomena specific to spoken language which are not captured by the STTS, including hesitations, backchannel signals, question tags, onomatopoeia, and non-words. As the TüBa-D/S does not use additional POS tags to label these phenomena,² it is interesting to see how they have been treated in the corpus.

Concerning hesitations, the TüBa-D/S encodes neither silent nor filled pauses such as *ahm*, *äh*

²The only additional tag used in TüBa-D/S is the BS tag used for isolated letters, which is not defined in the STTS.

(uhm, er). Occurrences of these seem to have been removed from the corpus. Particles expressing surprise (*ah*, *oh*), affirmation such as *gell* (right), or discourse particles such as *tja* (well) have been included in the transcription and assigned the label for interjections (ITJ). Backchannel signals as in (1) are also annotated as interjections in TüBa-D/S .

- (1) A: also ab zwölf Uhr habe ich bereits
well from twelve o'clock have I already
einen Termin
a date
- B: **mhm** welche Uhrzeit
mhm what time
- A: Well, from 12:00 on I already have a date.
B: Mhm, what time?

Question tags like *nicht/ne* (no), *richtig/gell* (right), *okay* (okay), *oder* (or) have been labelled as interjections, too (Example 2).

- (2) A: es war doch Donnerstag , **ne** ?
it was however Thursday , no ?
It was Thursday, right?

As a result, there is no straight-forward way to search for occurrences of these phenomena in the corpus. This is due to the fact that the VerbMobil corpus was created with an eye on applications for machine translation of spontaneous dialogues, and thus phenomena specific to spoken language were not the focus of the annotation.

3 Extending the STTS for spoken language

Our extension to the STTS provides 11 additional tags for annotating spoken language phenomena (Table 1).

3.1 Hesitations

Our extended tag set allows one to encode silent pauses as well as filled pauses.

POS	description	POS	description
PTKFILL	particle, filler	PAUSE	pause, silent
PTK	particle, unspec.	NINFL	inflective
PTKREZ	backchannel	XYB	unfinished word
PTKONO	onomatopoeia	XYU	uninterpretable
PTKQU	question tag	\$#	unfinished
PTKPH	placeholder		utterance

Table 1: Additional POS tags for spoken language data

The **PAUSE** tag is used for silent (unfilled) pauses which can occur at any position in the utterance.

- (3) das ist irgend so ein (-) Rapper
this is some so a rapper
This is some eh rapper.

The **PTKFILL** tag is used for filled pauses which can occur at any position in the utterance.

- (4) das ist irgend so ein **äh** Rapper
this is some so a eh rapper
This is some rapper.

3.2 Other particles

The **PTKONO** tag is used for labelling onomatopoeia and forms of echoism.

- (5) das Lied ging so **lalalala**
the song went like lalalala

The **PTKREZ** tag is used for backchannel signals. We define backchannel signals as plain, non-emotional reactions of the recipient to signal the speaker that the utterance has been received and understood.

- (6) A: stell dir das mal vor !
A: imagine you this PART. VERB PART. !
Imagine that !

- (7) B: **m-hm**
B: uh-huh

Preliminary annotation experiments showed a very low inter-annotator agreement for the distinction between answer particles and backchannel signals for *ja* (yes). To support consistency of annotation, we always label *ja* as an answer particle and not as a backchannel signal.

The **PTKQU** tag is used for question tags like *nicht/ne* (no), *richtig/gell* (right), *oder* (or), added to the end of a positive or negative statement.

- (8) wir treffen uns am Kino , **ne** ?
we meet REFL at the cinema , no ?
We'll meet at the cinema. Right ?

The **PTK** tag is used for unspecific particles such as *ja* (yes), *na* (there, well) when occurring in utterance initial position.

- (9) **ja** wer bist du denn ?
yes who are you then ?
And who are you now?

Please note that most occurrences of *ja* (yes) in the middle field are modal particles (Example 10) which are assigned the ADV label (adverb) in the German treebanks. Occurrences of *ja* in the pre-field, on the other hand, should be considered as discourse markers and thus should be treated differently (also see Meer (2007) for a discussion on the different word classes of *ja*).

- (10) die hat **ja** auch nicht funktioniert .
this has PTK.MOD also not worked .
This didn't work, either.

The **PTKPH** tag is used as a placeholder when the correct word class can not be inferred from the context. Example (11), for instance, has many possible readings. In (a), the correct POS tag would be noun (NN), while in (b) we would assign a past participle (VVPP) tag. The placeholder might also stand for a whole VP, as in (c).

- (11) er hat **dings** hier .
he has thingy here .
a. er hat MP3-Player_{NN} hier .
he has MP3 player here .
b. er hat gewonnen_{VVPP} hier .
he has won here .
c. er hat (Schuhe gekauft)_{VP} hier .
he has shoes bought here .

3.3 Non-words

Our tag set distinguishes 3 types of non-words.

1. uninterpretable
2. non-word in abandoned utterances
3. other

The **XYU** tag is used for lexical material which is uninterpretable, mostly because of poor audio quality of the speech recordings or because of code-switching. This tag should also be used for word tokens where it is not clear whether they are unfinished or simply non-words.

- (12) wir waren gestern bei (**fremdsprachlich**).
we were yesterday at (FOREIGN).
Yesterday we've been at (FOREIGN).

The **XYB** tag is used for abandoned words.

- (13) ich **ha** # sie kommt **Sams** äh Sonntag .
I ha- she comes Satur- eh Sunday .
I ha- she'll come on Satur- eh Sunday.

The **XY** tag is used for all non-words which do not fit one of the categories above. This category is consistent with the XY category used in the STTS where it is used for non-words including special symbols.

3.4 Inflective

The **NINFL** tag is used for non-inflected verb forms (Teuber, 1998) which are a common stilistic device in comics and computer-mediated communication, but are also used in spoken language.

- (14) ich muss noch putzen . **seufz** !
 I must still clean . sigh !
 I still have to clean. Sigh!

3.5 Punctuation

The **\$#** tag is used to mark interrupted/abandoned utterances. These can (but not necessarily do) include unfinished words, as in Example (15).

- (15) sie war ge #
 she was (UNINTERPRETABLE) #

4 Inter-Annotator Agreement

We measured inter-annotator agreement for three human annotators using the extended tagset on a test set (3415 tokens) of spontaneous multi-party dialogues from the KiDKo corpus (Wiese et al., 2012) and achieved a Fleiss' κ of 0.975 (% agr. 96.5). Many of the errors made by the annotators concern the different functions of *ja* in spoken data (discourse marker vs. answer particles).

5 Discussion

A major pitfall for the annotation of spoken language is the danger of carrying over annotation guidelines from standard written text which, at first glance, seem to be adequate for the description of spoken language, too. Only on second glance does it become obvious that what looked similar at first does not necessarily need to be the same.

A case in point is *ja* (yes), which in written text mostly occurs as a modal particle in the middle field, labelled as ADV, while in spoken dialogues occurrences of *ja* in utterance-initial position, labelled as answer particles (PTKANT), are by far the more frequent (Table 2). Motivated by the difference in distribution, we took a closer look at these instances and observed that many of them

POS	TIGER	TüBa-D/Z	TüBa-D/S
PTKANT	43	147	27986
ADV	154	372	4679
ITJ	2	0	0
NN	0	16	0
total	199	536	32664

Table 2: Distribution of *ja* (yes) in different corpora, normalised by corpus size

are in fact discourse markers (Example 9). We thus added the label PTK to our tag set, which is defined by its position in the utterance and its function.

As a second example, consider *weil* (because, since) which, according to standard grammars, is a subordinating conjunction. In TIGER as well as in the TüBa-D/S, all occurrences of *weil* are annotated as KOUS (subordinating conjunction). However, in TüBa-D/S we also find examples where *weil* is used to coordinate two main clauses (indicated by V2 word order) and thus should be labelled as a coordination (KON) (Example 16).

- (16) [...] fahren wir nicht zu früh los , **weil** sonst
 [...] drive we not too early PTK , because else
 bin ich unausgeschlafen
 am I sleepdeprived
 let's not start too early, else I'll be tired out

Finally, it is important to keep in mind that all types of linguistic annotation not only provide a description, but also an interpretation of the data. This is especially true for the annotation of learner data, where the formulation of target hypotheses has been discussed as a way to deal with the ambiguity inherent to a learner's utterances (Hirschmann et al., 2007; Reznicek et al., 2010). When annotating informal spoken language, we encounter similar problems (see Example 11). Adding an orthographic normalisation to the transcription might be seen as a poor man's target hypothesis where decisions made during the annotation become more transparent.

6 Conclusion

In the paper we extended the Stuttgart-Tübingen TagSet, the standard POS tag set for German, for the annotation of spoken language. Our extension allows for a more meaningful treatment of spoken language phenomena while also maintaining the comparability with corpora of written text annotated with the original version of the STTS.

Acknowledgments

This work was supported by a grant from the German Research Association (DFG) awarded to the Collaborative Research Centre (SFB) 632 “Information Structure”. We gratefully acknowledge the work of our annotators, Nadja Reinhold and Emiel Visser.

References

- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, Sofia, Bulgaria.
- Lou Burnard. 2007. Reference guide for the British National Corpus XML edition. Technical report, <http://www.natcorp.ox.ac.uk/XMLedition/URG/>.
- Kevin Gimpel, Nathan Schneider, Brendan O’Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanagan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT ’11, pages 42–47.
- John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520, San Francisco, California, USA.
- Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.
- Tilman Höhle. 1998. Der Begriff ”Mittelfeld”, Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.
- Dorothee Meer. 2007. ”ja er redet nur Müll hier.” – Funktionen von ’ja’ als Diskursmarker in täglichen Talkshows. *gidi Arbeitspapierreihe*, 11.
- Joakim Nivre and Leif Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics*, 6(1):47–78.
- Marc Reznicek, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas, 2010. *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin.
- Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.
- Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton van der Woulsen. 2003. CGN, an annotated corpus of spoken Dutch. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.
- Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, pages 705–711.
- Rosmary Stegmann, Heike Telljohann, and Erhard W. Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.
- Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2229–2235.
- Oliver Teuber. 1998. fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. *Germanistische Linguistik*, 26(6):141–142.
- Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.
- Heike Wiese, Ulrike Freywald, Sören Schalowski, and Katharina Mayr. 2012. Das Kiezdeutsch-Korpus. spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 40(2):97–123.

Comparing Variety Corpora with *Vis-À-Vis* – a Prototype System Presentation

Stefanie Anstein

Institute for Specialised Communication and Multilingualism
European Academy of Bozen/Bolzano (EURAC)
Viale Druso 1
I – 39100 Bolzano
stefanie.anstein@eurac.edu

Abstract

In this paper, the prototype system *Vis-À-Vis* to support linguists in their comparison of regional language varieties is presented. Written corpora are used as an empirical basis to extract differences semi-automatically. For the analysis, existing and adapted as well as new tools with both pattern-based and statistical approaches are applied. The processing of the corpus input consists in the annotation of the data, the extraction of phenomena from different levels of linguistic description, and their quantitative comparison for the identification of significantly different phenomena in the two input corpora. *Vis-À-Vis* produces sorted ‘candidate’ lists for peculiarities of varieties by filtering according to statistical association measures as well as using corpus-external knowledge to reduce the output to presumably significant phenomena. Traditional regional variety linguists benefit from these results using them as a compact empirical basis – extracted from large amounts of authentic data – for their detailed qualitative analyses. Via a user-friendly application of a comprehensive computational system, they are supported in efficiently extracting differences between varieties e.g. for documentation, lexicography, or didactics of pluri-centric languages.

1 Background and related work

Pluri-centric languages are languages with more than one national center and with specific national varieties (Clyne, 1992). The latter usually differ

to a certain extent on different levels of linguistic description, mostly on the lexical level – an example for variants in German being *Marille* (used in Austria and South Tyrol) vs. *Aprikose* (used in Germany and Switzerland) for ‘apricot’.

The question to be answered in the framework research project is to what extent the comparison of varieties for supporting variety linguists’ manual analyses can be automated with natural language processing (NLP) methods. The analysis results obtained with such computational systems will contribute to variety documentation, lexicography, and language didactics.

Vis-À-Vis has been developed for the case of the pluri-centric language German (Ammon, 1995) for the time being; its development originated in the initiatives *Korpus Südtirol*¹ and *C4*². The former is preparing a written text corpus of South Tyrolean German³ (Anstein et al., 2011), which can also be queried together with other German variety corpora with the help of the distributed query engine implemented in the *C4* project (Dittmann et al., 2012). In addition to interactively run single queries in the *C4* corpora, variety linguists can use *Vis-À-Vis* to exploratively and empirically analyse and compare corpora on the desired levels of linguistic description. This is especially relevant since the amount of electronically available data constantly increases and can no longer be handled purely manually. The benefit of supportive tools from the NLP community for em-

¹<http://www.korpus-suedtirol.it>

²<http://www.korpus-c4.org>

³South Tyrolean German is the German variety used as an official language in the Autonomous Province of Bolzano / South Tyrol in Northern Italy (Egger and Lanthaler, 2001).

pirical analyses in traditional linguistics is clearly evident in this scenario, which is where the usefulness of this approach can be seen.

Other work that is related to this topic has been done in general comparative corpus linguistics as e. g. described in McEnery et al. (2006) or in Schmied (2009). Comparative regional variety linguistics has first been handled mostly manually with single introspective studies or later as well with the help of the Internet, e. g. in the development of the *Variant Dictionary of German* (Ammon et al., 2004). By now, more and more projects use variety corpora and automated comparison methods, e. g. the *ICE*⁴ initiative or Baeclar do Nascimento et al. (2006) studying the varieties of Portuguese.

2 The system *Vis-À-Vis*

In this section, the toolkit's implementational and functional details as well as its accessibility are described.

2.1 Design and implementation

Vis-À-Vis is written in the programming language *Perl*⁵ with a modular approach.

Input and output The main script takes as input (i) written text corpora of two varieties and, if available, (ii) lists with known peculiarities (e. g. named entities or regionalisms) of the variety to be investigated with respect to the so-called reference variety. The output is composed by (i) general quantitative information on the corpora and their comparability as well as (ii) lists of phenomena occurring in the two corpora, sorted by frequency and statistical values including filtering information for identifying new regionalism candidates.

Architecture As a first step, the two input corpora are checked with regard to their comparability. Then the corpora are annotated including corpus-external linguistic knowledge. In the extraction module, phenomena from different levels of linguistic description are identified. These are further compared by frequency and by statistical association measures and are presented to the user

⁴International Corpus of English; <http://ice-corpora.net/ice/index.htm>; Nelson (2006)

⁵<http://www.perl.org>

together with filter information for their interpretation. The overall *Vis-À-Vis* design can be seen in figure 1; details of the modules are given in section 2.2.

Approaches Both top-down / corpus-based and bottom-up / corpus-driven methods are applied; the former for the revision of possibly existing, manually compiled variant lists and the latter for their enhancement. Morpho-syntactic patterns according to part-of-speech (PoS) tags are used as well as explorative statistical approaches on the basis of significance measures.

2.2 Functionalities

In the following, the system's functional features are elaborated on.

Comparability check As a measurement for the comparability of the two corpora to judge the reliability of the comparison results (see also Gries, 2007), their 'complexity', as also investigated e. g. in learner corpus studies, is taken. On the one hand, the type-token ratio is calculated, which is an indicator for vocabulary richness and lexical variability. On the other hand, the proportion of lexical to grammatical words is given, measured as lexical density (Stubbs, 1986).

Annotation After tokenisation, the corpora are PoS-tagged and lemmatised with the *TreeTagger*⁶. The corpus-external lexical lists are used to lemmatise words that are not known to the tagger. In a bootstrapping process, new findings can be integrated via the annotations into a new *Vis-À-Vis* run by providing new lexicon entries as additional input.

Analysis levels On the lexical level, all word forms or lemmas of the two corpora are counted with the *Corpus Query Processor (CQP)*⁷. On the bi-gram level, the extraction of co-occurrences is done by searching for PoS patterns (e. g. adjective + noun or adverb + adjective) via *CQP* corpus queries. On an exemplary higher level of linguistic description, frequencies of main and subordinate clauses for both corpora are provided and the

⁶<http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger>; Schmid (1994)

⁷<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench>; Christ (1994)

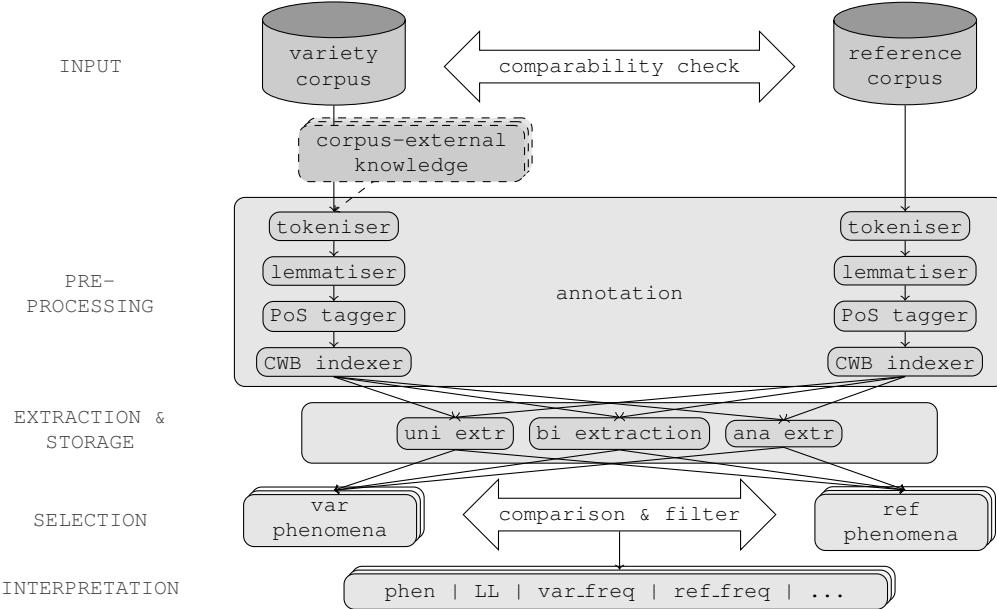


Figure 1: Overall architecture of *Vis-À-Vis*

word order in subordinate clauses is investigated. The extraction is done by *CQP* queries for specific PoS patterns, e.g. a subordinating conjunction with verb-second word order, which is an anacoluthon (sentence ‘break’) in written language.

Comparison and statistics The difference of phenomenon occurrences in the two corpora is determined (i) with absolute as well as relative frequencies with respect to corpus sizes and (ii) with statistical association measures. Two measures indicate how significantly different the frequency of one phenomenon is in the variety corpus with respect to the reference corpus. The log-likelihood (LL) measure (Dunning, 1993) was chosen as an association measure recommended e.g. by Rayson and Garside (2000) for co-occurrences and also by Evert (2004) both for words and collocations. As a second measure, $LL \cdot \log(frequency)$ is given, since Kilgarriff and Tugwell (2002) state that LL values over-emphasise the significance of low-frequency items and thus suggest to adjust these values for measuring e.g. lexicographic relevance.

Filtering The output lists are marked according to several external knowledge lists, partly system-internal and, if available, also provided by the user. They consist of known regionalism

lists (e.g. ‘Südtirolisms’⁸ taken from Abfaltrerer, 2007), place and person name lists, and lists of ‘reality’ descriptions (Heid, 2011) such as currency names. By filtering out known information, new regionalism candidates can be identified by sorting the output lists accordingly. In addition, the statistical measures and a filter according to expected frequency values serve as a guideline to the probability of the candidates to be relevant.

2.3 Access and extensibility

The system will be released with a free software license for the download as a stand-alone application, and it can also be used online from the *Korpus Südtirol* website. The download comprises two possibilities of usage – Unix command line use and a graphical user interface (GUI) for both Unix and Windows environments. A comprehensive system documentation with all details for its usage is provided for all scenarios.

Command line use The script *visavis.perl* supports several parameters as described in the following. With the option *-f*, the user decides

⁸Südtirolisms (‘South-Tyrol-isms’) are the specific variants of linguistic units used in South Tyrolean German. The terms ‘primary’ (exclusively used in South Tyrol) and ‘secondary’ (shared with other varieties) Südtirolisms have been coined by Abfaltrerer (2007).

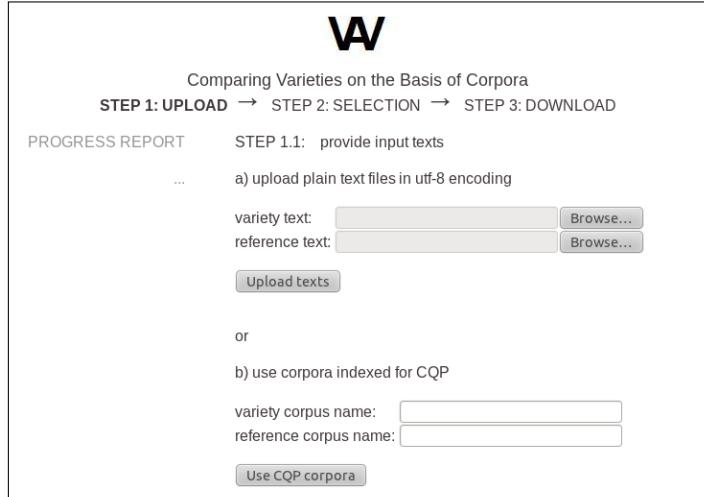


Figure 2: *Vis-À-Vis* GUI start page – corpus upload

if either word forms or lemmas are to be considered in the analysis process. The option `-l` chooses the level of extraction and comparison (lexical, co-occurrence, or anacolutha). The option `-e` takes corpus-external knowledge for the variety of all kinds as input, lexicon entries (one-word units) with PoS and lemma information and external knowledge lists with certain prefixes to each of their filenames. Finally, the option `-i` is used to specify if the two input corpora are in text format or if they are available corpora indexed for *CQP*. As results, the user gets general data, e. g. regarding the comparability of the two corpora, as well as the locations of the comparison output files to view or further process printed in the terminal window.

Graphical user interface For easier accessibility of *Vis-À-Vis*, users can upload their data over a GUI and are guided through the options for the comparison process up to the download of their analysis results. In figure 2, the start page of the *Vis-À-Vis* GUI is shown by a screenshot to give an idea of its layout. After choosing the kind of corpus input and providing the data locations, lists with corpus-external knowledge can be uploaded, if available. In the second step, the desired analysis and comparison level is chosen, and after the *Vis-À-Vis* run, the result data can be viewed and downloaded for further processing. Through the GUI, also a direct link to *Korpus Südtirol* for the verification of South Tyrolean phenomena and for context search is provided.

Extensions In the stand-alone version, several possibilities to adapt and extend the system are given, for example: integration of additional annotation, extraction, and comparison tools, usage of other comparability measures, extension of the analysis levels, enhancement of the filtering, application of additional statistical measures for comparison, or adaptation to other languages. Also the integration of the tool into a larger corpus processing architecture is a possible and promising development to be followed further.

3 Evaluation and conclusion

The system evaluation using Abfalterer's Südtirolisms as a gold standard showed promising results for *Vis-À-Vis*' approach. First concrete outcomes obtained by using *Vis-À-Vis* for lexicographic tasks are new as well as refined dictionary entries for South Tyrolean German, e. g. for the lemmas *ehestens* (as soon as possible), *Konsortium* (consortium), *ober* (above), or *weiters* (furthermore); for details see Abel and Anstein (2011). Detailed precision, recall, and f-score values on the basis of different parameters can be found in Anstein (to appear).

Given such findings, it seems worth following *Vis-À-Vis*' approach and develop it further to provide an even more useful NLP tool for traditional linguistics, also to serve in other fields than regional variety linguistics – wherever corpora are to be compared in order to find significantly different phenomena.

References

- Andrea Abel and Stefanie Anstein. 2011. Korpus Südtirol - Varietätenlinguistische Untersuchungen. In Andrea Abel and Renata Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, Bolzano. University Press.
- Heidemaria Abfalterer. 2007. *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht: lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*, volume 72 of *Germanistische Reihe*. Innsbruck University Press.
- Ulrich Ammon, Hans Bickel, Jakob Ebner, Ruth Esterhammer, Markus Gasser, Lorenz Hofer, Birte Kellermeier-Rehbein, Heinrich Löffler, Doris Mangott, Hans Moser, Robert Schläpfer, Michael Schloßmacher, Regula Schmidlin, and Günter Valaster. 2004. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. De Gruyter, Berlin / New York.
- Ulrich Ammon. 1995. *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. De Gruyter, Berlin / New York.
- Stefanie Anstein, Margit Oberhammer, and Stefanos Petrakis. 2011. Korpus Südtirol - Aufbau und Abfrage. In Andrea Abel and Renata Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, Bolzano. University Press.
- Stefanie Anstein. to appear. Computational Approaches to the Comparison of Regional Variety Corpora – Prototyping a Semi-automatic System for German. PhD thesis.
- Maria Fernanda Bacelar do Nascimento, José Bettencourt Gonçalves, Luísa Pereira, Antónia Estrela, Alfonso Pereira, Rui Santos, and Sancho M. Oliveira. 2006. The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon. In *Proceedings of the Fifth International Language Resources and Evaluation Conference (LREC 2006)*, pages 1791–1794, Genoa, Italy.
- Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX)*, pages 23–32, Budapest.
- Michael G. Clyne, editor. 1992. *Pluricentric languages: Differing norms in different nations*. De Gruyter, Berlin / New York.
- Henrik Dittmann, Matej Ďurčo, Alexander Geyken, Tobias Roth, and Kai Zimmer. 2012. Korpus C4 – a distributed corpus of German varieties. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, number 14 in Hamburg Studies in Multilingualism (HSM). John Benjamins, Amsterdam.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1).
- Kurt Egger and Franz Lanthaler, editors. 2001. *Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt*. Folio, Vienna / Bolzano.
- Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.
- Ulrich Heid. 2011. Korpusbasierte Beschreibung der Variation bei Kollokationen: Deutschland – Österreich – Schweiz – Südtirol. In Stefan Engelberg, Anke Holler, and Kristel Proost, editors, *Sprachliches Wissen zwischen Lexikon und Grammatik*, Jahrbuch 2010. De Gruyter, Institut für Deutsche Sprache, Mannheim.
- Adam Kilgarriff and David Tugwell. 2002. Sketching words. *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137.
- Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies - An Advanced resource book*. Routledge Applied Linguistics. Routledge.
- Gerald Nelson. 2006. The core and periphery of world englishes: a corpus-based exploration. *World Englishes*, 25(1):115–129.
- Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester.
- Josef Schmid. 2009. Contrastive corpus studies. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, number 29 in Handbooks of Linguistics and Communication Science, chapter 54, pages 1140–1159. De Gruyter, Berlin.
- Michael Stubbs. 1986. Lexical density: A computational technique and some findings. In Malcolm Coulard, editor, *Talking about Text. Studies Presented to David Brazil on His Retirement*, pages 27–42. English Language Research, Birmingham.

Selecting Features for Domain-Independent Named Entity Recognition

Maksim Tkachenko

St Petersburg State University

St Petersburg, Russia

maksim.tkachenko@math.spbu.ru andrey.simanovsky@hp.com

Andrey Simanovsky

HP Labs Russia

St Petersburg

Abstract

We propose a domain adaptation method for supervised named entity recognition (NER). Our NER uses conditional random fields and we rank and filter out features of a new unknown domain based on the means of weights learned on known domains. We perform experiments on English texts from OntoNotes version 4 benchmark and see a statistically significant better performance on a small number of features and a convergence of performance to the maximum F_1 -measure faster than conventional feature selection (information gain). We also compare with using the weights learned on a mixture of known domains.

1 Introduction

Nowadays the majority of text analytics techniques require that named entities (people, companies, products, etc) are recognized in text. While domain-specific named entity recognition (NER), e.g. in newswire, can be quite precise (Ratinov and Roth, 2009), the accuracy of NER systems is significantly degraded in the presence of several domains (Evans, 2003), especially unknown ones. The generalization of this problem is known as a domain-adaptation (DA) problem. DA is a hard problem (Jiang, 2008). Another issue with NER systems is efficiency. Feature selection can address the efficiency issue in supervised NER systems but regular methods of fast feature selection under-perform in the presence of multiple domains (Satpal and Sarawagi, 2007).

In this paper we consider the problem of feature selection for a supervised NER system that

works with texts from multiple domains. We take a large set of feature types that we analyzed on CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) data and compare common feature selection methods (information gain) with methods that rank features based on the weights learned by a machine learning algorithm on the known domains (Jiang and Zhai, 2006). We perform experiments on OntoNotes version 4 (Hovy et al., 2006). We demonstrate that proposed by us feature ranking by domain weights mean is a better feature selection method than information gain and it is competitive against ranking by the weight learned over a mixture of domains.

The main contribution of this paper is that we propose the mean of feature weights learned by the algorithm on known domains as a feature ranking criterion for unknown domains. On large new OntoNotes benchmark, we observe that thresholding on the suggested ranking is a more effective feature selection method.

2 Related work

Named entity recognition is a task that is actively pursued in industry (for example, www.opencalais.com) and in academia (Ratinov and Roth, 2009) since the 6-th Message Understanding Conference (Grishman and Sundheim, 1996). A good overview of the area is given by (Nadeau and Sekine, 2007). We use supervised NER based on conditional random fields (CRF) that were first proposed in (McCallum and Li, 2003). The feature types that we consider come from various previous works (Ratinov and Roth, 2009) etc. We also consider novel feature types.

Features
Context features (token windows and bi-grams)
Token-based features (shape, affixes)
Part-of-speech tags
Brown clusters (Brown et al., 1992)
Clark clusters (Clark, 2003)
Phrasal clusters (Lin et al., 2010) (novel in NER)
Wikipedia gazetteers (Tkatchenko et al., 2011)
DBPedia gazetteers (novel)
Context aggregation (Ratinov and Roth, 2009)
2-stage prediction (Krishnan and Manning, 2006)
Date and hyphenation features

Table 1: Evaluated features.

Several works specifically focus on adapting NER to new domains. Jiang and Zhai (Jiang and Zhai, 2006) explored non-uniform Gaussian prior on the weights of the features to penalize domain specific features. In (Jiang and Zhai, 2007) the authors used the same idea and proposed a two-stage approach to DA. The first stage is recognition of generalizable features. The second stage is learning an appropriate weight with the use of a modified logistic regression framework. A team from Bombay (Satpal and Sarawagi, 2007) proposed an approach to choosing a feature space in which source and target distributions are close. Ando and Zhang (Ando and Zhang, 2005) proposed a semi-supervised multi-task learning framework which is used to identify generalizable features (Ben-david et al., 2007). Klinger and Friedrich (Klinger and Friedrich, 2009) explored information gain and iterative feature pruning in application to feature selection in NER but they did not consider DA perspective.

3 Domain adaptation method

We consider a supervised named entity recognition with a sequential labeling algorithm. The machine learning model uses a comprehensive set of features that represent tokens which are classified using appropriate common labelling schemes like BILOU. Table 1 contains the set of features that we evaluated in this work.

Our primary target is the case of domain adaptation, when there are several known domains but a new document comes from an unknown do-

main. In this setup we assume that the NER system has a lot of information on the known domains including recognizers that have been trained on the domains or their mixtures. At the same time limited information is available on the unknown domain apart from the document in which named entities are being recognized. Two options were suggested in the literature. One can map the problem dimension set into a higher dimensional space reserving one set of dimensions for each domain and a separate set of dimensions for a combination of all domains (Daume III, 2007). A recognizer for the target document is learned in this higher dimensional space. Alternatively, the case can be addressed by applying a machine learning algorithm that starts not from the default weights for the target document but from the weights that are functions of the known domains. The expectation is that the weights learned for the target document would be biased towards the weights learned on the known domains. The latter approach can also be interpreted as a feature selection strategy — features are ranked according to their weights learned on a mixture of known domains and filtered out based on a threshold value.

We propose to enhance the latter approach with the use of the mean of weights learned across several known domains. The intuition behind this proposition is that a feature that has big weights in several domains is more likely to be important in a new domain than a feature that has a big weight in only one large domain. Thus, macro-averaging makes more sense than micro-averaging that was applied in other works. We also claim and show that the weight-based method works significantly better than a naive feature selection by a common fast feature ranking like information gain.

Our method implies the following steps. NER recognizers are trained on known domains and feature weights produced by them are remembered. When a new domain is encountered, its features are ranked according to the means of the remembered feature weights. Features that do not appear in a domain have a zero weight in it. Previously unseen features are ranked lowest (smoothing can be applied). The obtained ranking is used as a feature utility metrics and top N features are selected.

	Training		Test	
	Range	Size (KB)	Range	Size (KB)
nw-xinhua	0-260	4336	261-325	883
mz-sinorama	0-1062	6047	1063-1078	1519
wb-eng	0-13	1934	14-17	778
bc-msnbc	0-5	2228	6-7	813
bn-cnn	0-375	2989	376-437	716

Table 2: The size of training and test sets for the subcorpora. The file ranges refer to the numbers within the names of the original OntoNotes files.

4 Experiments

We performed experiments on English texts from OntoNotes version 4.0 benchmark. It is a large set of mainly newswire texts of various genres. We used the CoNLL 2003 task NER classes. We compared our feature selection method to information gain and to a feature selection algorithm based on ranking features in accordance with weights learned on a mixture of domains. Five OntoNotes corpora from different domains were used. In each experiment one corpus was withheld as an unknown domain; the rest were used as known domains. Each corpus was split into training and test sets using the document ranges presented in Table 2. The subcorpora are MSNBC (broadcast conversation), CNN (broadcast news), Sinorama (magazine), Xinhua (newswire), and wb (web data).

Figure 1 shows feature selection results for five experiments on *sinorama* subcorpus; the results on the test sets of other subcorpora are similar. In the presented experiment the feature ranking was the same in each of the five *sinorama* experiments and was built using the means of the feature weights learned on the training sets of the four other subcorpora. In each of the experiments a training set of a different OntoNotes subcorpus was used. The most interesting setup is experiment (c), where the training set of *sinorama* was used to collect features for feature selection, since it is the case closest to the appearance of a new domain. Other setups check stability. We can see that our method clearly outperforms information gain and most of the time it reaches flat F_1 -measure values (falls into $\pm 1\%$ range) faster than the method based on weights learned from a mixture of domains.

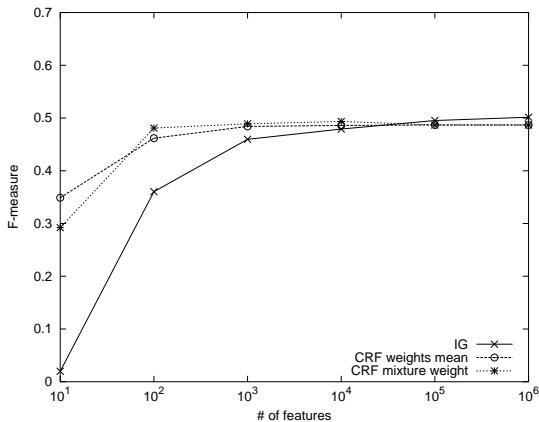
To test statistical significance of the obtained results, we used approximate randomization test described in (Yeh, 2000). It samples a mixture two algorithms being compared and tests if it is better than the baseline. With 100000 iterations we observed that performance advantage of our method against information gain is statistically significant up to more than 1000 features in all experiments. The same holds almost all the time for another weights-based method.

Apart from bias-reduction, feature selection also improves performance of the system. In our experiments the throughput grew from 31 to 45 and 66 tokens per millisecond with reducing 10^5 features to 10^3 and 10^2 respectively.

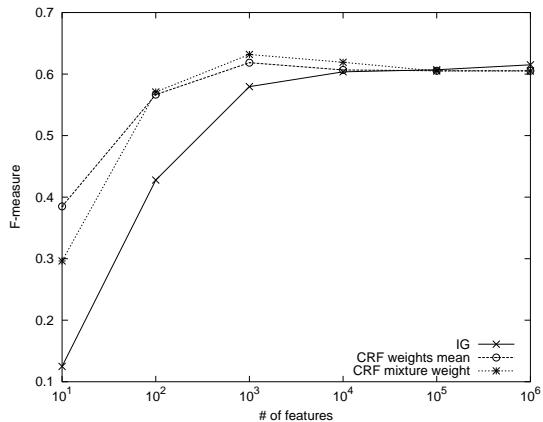
5 Conclusions

In this paper we presented evidence that, in terms of NER F_1 -measure, ranking by the mean of feature weights learned on the known domains is a better method of fast feature selection than regular ones (e.g. information gain). It is also competitive against ranking by a weight learned on the mixture of known domains. The experiments on OntoNotes benchmark show that our method obtains higher F_1 measure on a small number of features as compared to other fast feature selection methods. Consequently, our method is less prone to over-fitting.

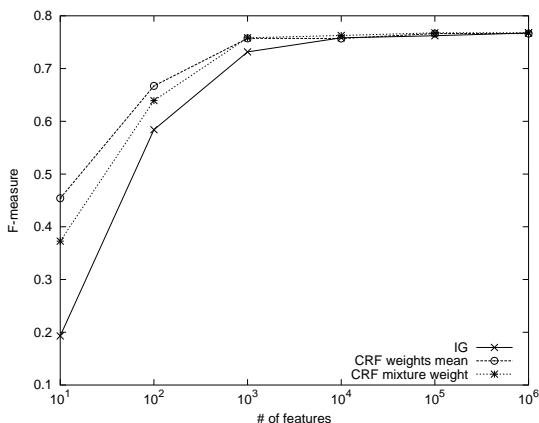
We have explored the two extremes: using a mixture of domains to learn feature weights and taking the mean of feature weights learned on each domain. While we show that the approaches are competitive, our future work is to explore possible combination of the two approaches, e.g., to automatically learn coefficients with which each domain should be taken into account.



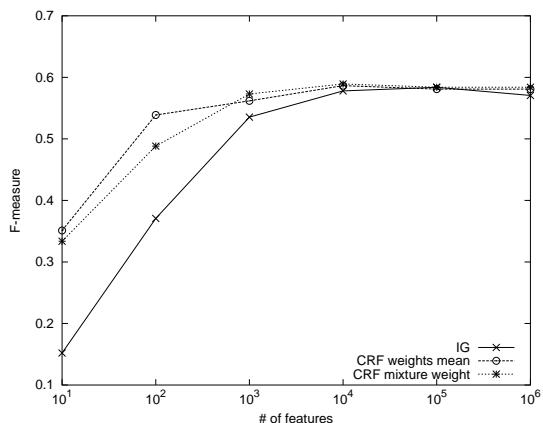
(a) bc-msnbc



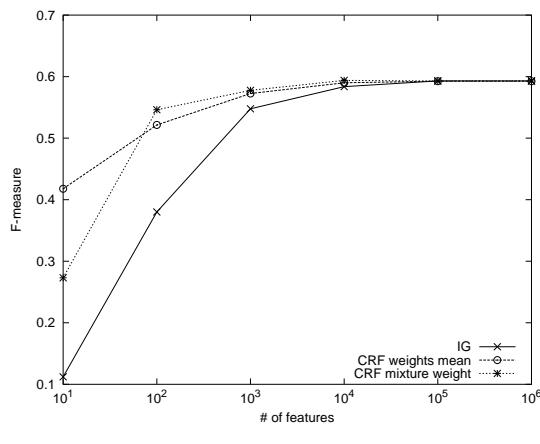
(b) bn-cnn



(c) mz-sinorama



(d) nw-xinhua



(e) wb-eng

Figure 1: Feature selection on different training data. The test data is mz-sinorama. The Y-axis on all charts is F_1 -measure. The X-axis is the number of features. Lines with crosses stand for feature selection based on information gain (IG). Lines with stars stand for feature selection based on the feature weight in a mixture of domains (CRF mixture weight). Lines with circles stand for feature selection based on the feature weights mean (CRF weights mean). One can see that the latter lines start at higher values of F_1 and most of the time reach flat part of the chart faster than the lines corresponding to other methods.

References

- Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL*. The Association for Computer Linguistics.
- Shai Ben-david, John Blitzer, Koby Crammer, and Presented Marina Sokolova. 2007. Analysis of representations for domain adaptation. In *In NIPS*. MIT Press.
- Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.
- Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.
- Richard Evans. 2003. A framework for named entity recognition in the open domain. In *In Proceedings of the Recent Advances in Natural Language Processing (RANLP)*, pages 137–144.
- Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90 In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.
- Jing Jiang and Chengxiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *In Human Language Technology Conference*, pages 74–81.
- Jing Jiang and ChengXiang Zhai. 2007. A two-stage approach to domain adaptation for statistical classifiers. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 401–410, New York, NY, USA. ACM.
- Jing Jiang. 2008. A Literature Survey on Domain Adaptation of Statistical Classifiers, March.
- Roman Klinger and Christoph M. Friedrich. 2009. Feature subset selection in conditional random fields for named entity recognition. In *Proceedings of the International Conference RANLP-2009*, pages 185–191, Borovets, Bulgaria, September. Association for Computational Linguistics.
- Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New Tools for Web-Scale N-grams.
- Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January.
- L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6.
- Sandeepkumar Satpal and Sunita Sarawagi. 2007. Domain adaptation of conditional probability models via feature subsetting. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 224–235. Springer.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA. Association for Computational Linguistics.
- Maksim Tkatchenko, Alexander Ulanov, and Andrey Simanovsky. 2011. Classifying wikipedia enti-

ties into fine-grained classes. In *ICDE Workshops*, pages 212–217.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences.

Ambiguity in German Connectives: A Corpus Study

Angela Schneider

Master Student

Department of Computational Linguistics
Universität Heidelberg / Germany
schneida@cl.uni-heidelberg.de

Manfred Stede

Applied Computational Linguistics
EB Cognitive Science
Universität Potsdam / Germany
stede@uni-potsdam.de

Abstract

For deriving information on text structure (in the sense of coherence relations holding between neighbouring text spans), connectives are the most useful source of evidence on the text surface. However, many potential connectives also have a non-connective reading, and thus a disambiguation step is necessary. This problem has received only relatively little attention so far. We present the results of a corpus study on German connectives, designed to estimate the magnitude of the ambiguity problem and to prepare the development of disambiguation procedures.

1 Introduction

Connectives are a central source of information for *discourse parsing*, i.e., the identification of structural relationships between sentences or clauses in text. As with many lexical items, however, there is an ambiguity problem that needs to be resolved before such structural information can be exploited. We have conducted a corpus study in order to determine the magnitude of the ambiguity problem for German connectives, and to pave the way for implementing effective disambiguation procedures. The results will be presented in Section 3. Before, we briefly introduce the notions of discourse parsing and connectives in the remainder of this section, and discuss related work in Section 2.

1.1 Discourse parsing

Coherence relations are often used to model the coherence of texts, and sometimes to also ascribe

structure to it. For instance, many researchers would analyze the discourse *The hotel is nice and clean. But I think it is much too expensive* as an instance of the *Concession* relation holding between the two sentences. Certain theories of discourse structure then take the observation of such relations a step further and postulate that, by means of recursive application of such relations, a structural description can be produced as a model of the text's internal coherence (e.g., (Polanyi, 1988) (Mann and Thompson, 1988), (Asher and Lascarides, 2003)). The step of automatically building these descriptions is commonly called *discourse parsing* (see, e.g., (Marcu, 2000)).

Regardless whether the aim is to derive a full text structure or to merely identify local structural configurations at selected points of a text, the identification of coherence relations is made much easier when explicit connectives are present: words that – more or less explicitly – signal the presence of a coherence relation.

1.2 Connectives

Connectives are non-inflectable, closed-class lexical items that denote two-place relations, i.e., they need two arguments in order to be used felicitously (Pasch et al., 2003). Traditional grammars often group connectives together according to their semantic function (e.g., contrastive, temporal, causal); the mapping to a coherence relation can be seen as a discourse-level extension of that grammatical analysis. Syntactically, connectives are not homogeneous; we find coordinating conjunctions (e.g., *and*, *but*), subordinating conjunctions (e.g., *while*, *because*), and discourse ad-

verbials (e.g., *therefore*, *still*). Some researchers also include certain propositions such as *due to* or *despite*.

The mapping from connective to coherence relation is non-trivial for three different reasons. (i) Connectives vary in their specificity, ranging from very clear ones (*although*) to vague ones (*and*). (ii) Some connectives are ambiguous between different semantic readings (or, coherence relations), such as *while*, which can be temporal or contrastive. (iii) Some words have a connective and a non-connective reading, such as *since*, which can be a subordinating conjunction or a preposition without discourse function (*She has been a widow since 1983*).

In the following, we address only problem (iii), and in particular study it for the German language. Our method, however, should be applicable to other languages just as well.

2 Related work

There are quite a few papers dealing with connective ambiguity of the kind described in point (ii) above, but regarding the ambiguity between a connective and non-connective reading for English words, we are aware of only (Pitler and Nenkova, 2009). In this study, Pitler and Nenkova investigate to what extent syntactic information is useful in solving this ambiguity problem.

For German, (Bayerl, 2004) had presented a pilot study on disambiguating the potential connective *wenn*, also using various syntactic features, but her focus was on ambiguity (ii). Regarding (iii), (Dipper and Stede, 2006) suggested the approach to incrementally retrain a POS tagger with non-/connective features, but their experiment was restricted to nine words. To our knowledge, there is no study of the “bigger picture” of connective ambiguity yet.

3 Corpus study

The first problem is to identify the set of ambiguous words: those that have a connective and a non-connective reading. For German, such a set has been determined in earlier work by (Dipper and Stede, 2006) who gave a list of 41 German connectives that can be ambiguous; these are listed below in Table 1.

For each of these words we now collected a small corpus consisting of 200-250 sentences taken from the *DWDS-Kernkorpus*¹. Then we manually annotated all sentences as to the candidate words having a connective or a non-connective reading.² In this way, we ended up with 1-200 sentences for each reading of each potential connective. Since the sentences were collected randomly, the distribution of non-/connective readings can be interpreted as approximating the actual distribution in language. Table 1 includes this information. Notice that some words have a very skewed distribution so that a “default” reading could be assumed (*auch*, *nur*, *wie* and others), whereas most show a balance of the two readings, so that disambiguation is indeed non-trivial. In the table, we also give the raw frequency of the words’ occurrences in the DWDS corpus (which is the parameter for sorting the table). For the highly infrequent connective readings of *auch* and *nur* we subsequently searched for another 50 occurrences, in order to have enough material for the disambiguation steps described below.

3.1 Standard POS tagging

An obvious first idea is to explore standard POS tagging for the disambiguation problem (iii) as stated in section 1.2. Therefore, we tagged our data with the TreeTagger (Schmid, 1994), which uses the STTS Tagset³. However, it turns out that POS tags disambiguate only eight potential connectives with an f-score > 0.75 (which we take as a threshold for “acceptable” performance). Tables 2 and 3 show the respective tags and their precision and recall.

A side result of our analysis is that for some POS tags, general rules can be established; in particular, words tagged with the tags PTKVZ, NN or NE are always non-connectives. Overall, however, the intermediate conclusion is negative: Standard POS tagging yields acceptable results only for eight of the 41 candidates. And, as Table 1 reveals, the eight “easy” words range in the

¹<http://www.dwds.de>

²The criteria that were applied in making the annotation decisions are documented in (Schneider, 2012).

³<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.htm>

middle of the frequency distribution, so that the problem at large can be solved only to a small extent by tagging.

3.2 POS-context disambiguation rules

For the remaining 33 potential connectives (where the POS-tag alone does not give enough information to disambiguate), our next step was to look at their immediate context and determine whether POS tag patterns can be identified for making the decision. We found that for ten potential connectives, it is indeed possible to formulate a set of context-rules that use the POS-tags of the tokens directly before and after the potential connective, and which perform with an f-score > 0.75. These connectives and the context rules are shown in table 4. The underlined POS-tag is the one the potential connective is annotated with, while the POS-tags before and/or after describe the directly adjacent words. To merge some context rules together the following symbols are used: \$ stands for any punctuation mark, VV.* for any full verb and V.FIN for any finite verb.

4 Summary and outlook

We provided a comprehensive analysis of the connective ambiguity problem in German and showed for our base set of 41 ambiguous words that

- a set of words has a relatively clear tendency to occur in either the connective or non-connective reading,
- for eight words, plain POS tagging yields good results for disambiguation, and
- for another ten words, fairly reliable POS-context patterns can serve to disambiguate the reading.

These results can serve as the basis for an implementation. For the remaining ambiguous words, a straightforward first step is to simply assume the majority reading as taken from Table 1. Alternatively, one could search more intensely for POS patterns by working with larger corpora and machine learning techniques. And another option to explore, of course, is to test whether syntactic chunking or even full parsing can be trusted

to resolve the ambiguities. For English, this has been done by (Hutchinson, 2004), using the Charniak parser. In an interesting experiment, (Versley, 2010) suggested to project connective annotations from English onto German data, thus circumventing the problem of lacking (German) training data for machine learning. Working directly on the output of a German parser, however, has to our knowledge not been attempted yet.

References

- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.
- Petra Bayerl. 2004. Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie. *Linguistik Online*, 18.
- Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives. In Miriam Butt, editor, *Proc. of KONVENS '06*, pages 167–173, Konstanz.
- Ben Hutchinson. 2004. Mining the web for discourse markers. In *Proc. of LREC-04*, Lisbon.
- William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.
- Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge/MA.
- Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Herrmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.
- Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.
- Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conference on New Methods in Language Processing*, pages 44–49, Manchester.
- Angela Schneider. 2012. Disambiguierung von Diskurskonnektoren im Deutschen. Bsc. thesis, Dept. Linguistics, Universität Potsdam, March.
- Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proc. of the Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu/Estland.

word	raw freq.	non conn. (of 200)	conn. (of 200)
und	1.550.931	121	79
als	382.237	183	17
auch	351.946	199	1
wie	268.878	188	12
so	263.201	163	37
nur	226.392	199	1
aber	219.248	55	145
dann	106.610	56	144
doch	92.350	120	80
da	88.776	110	90
denn	60.647	84	116
also	54.251	142	58
seit	39.670	179	21
während	39.320	98	102
darauf	32.630	187	13
dabei	27.616	181	19
allein	24.781	183	17
wegen	20.994	9	191
dafür	20.857	178	22
daher	18.874	16	184
sonst	18.139	144	56
statt	15.749	163	37
zugleich	15.093	119	81
allerdings	14.481	33	167
dagegen	13.388	52	148
ferner	12.622	18	182
trotz	10.921	20	180
darum	10.090	120	80
außer	10.084	185	15
soweit	8.884	31	169
entgegen	6.904	142	58
danach	6.351	85	115
wonach	3.042	186	14
worauf	2.840	103	97
weshalb	2.638	124	76
seitdem	2.396	139	61
womit	2.048	109	91
aufgrund	1.806	200	0
allenfalls	1.162	177	23
wogegen	614	54	146
nebenher	286	93	107
weswegen	188	111	89

Table 1: Ambiguous connectives with their raw frequencies and non-/connective distributions for 200 random occurrences

	POS = connective	Prec. in %	Recall in %
denn	KON	85.6	94.2
doch	KON	86.3	83.1
entgegen	APPO, APPR	97.9	65.7
seit	KOUS	77.8	84.0
seitdem	KOUS	81.0	77.0
trotz	APPR	99.5	100
während	KOUS	81.8	96.1
wegen	APPO, APPR	98.3	100

Table 2: POS tagging results for connective readings

	POS = non-connective	Prec. in %	Recall in %
denn	ADV	91.9	79.1
doch	ADV	90.2	92.1
entgegen	PTKVZ	85.7	99.3
seit	APPR	98.0	96.6
seitdem	PAV	90.2	92.1
trotz	NN	100	95.0
während	APPR	95.3	78.6
wegen	NN	100	60.0

Table 3: POS tagging results for non-connective reading

	Context-rules connective	Prec. in %	Recall in %	Context-rules non-connective	Prec. in %	Recall in %
also	\$, <u>ADV</u> V.FIN \$. <u>ADV</u> V.FIN V.FIN <u>ADV</u>	83.3	87.3	all else	95.0	93.2
auch	<u>ADV</u> VVFIN	98.0	78.1	all else	95.3	99.7
außer	\$ <u>APPR</u> \$, \$ <u>APPR</u> KOUS	100	86.7	all else	98.9	100
da	\$, <u>ADV</u> KON <u>ADV</u> KOUS	77.5	86.9	<u>ADV</u> <u>PTKVZ</u>	87.7	78.8
darum	all else	70.7	81.7	<u>PAV</u> \$ <u>PAV</u> VV	82.5	89.7
nebenher	\$. <u>ADV</u> VVFIN all else	88.2	83.3	<u>ADV</u> \$ <u>ADV</u> VV.* <u>ADV</u> KON	81.8	87.1
nur	\$. <u>ADV</u> VVFIN	100	74.2	all else	93.6	100
so	\$, <u>ADV</u> KOUS \$, <u>ADV</u> V.FIN KON <u>ADV</u> V.FIN	77.8	77.8	all else	95.1	95.1
sonst	\$ <u>ADV</u> V.FIN	87.2	70.7	all else	89.6	96.1
soweit	\$ <u>ADV</u> KOUS	98.9	92.3	all else	65.9	93.5

Table 4: Evaluation results for POS patterns

Corpus-based Acquisition of German Event- and Object-Denoting Nouns

Stefan Gorzitze and Sebastian Padó

Institut für Computerlinguistik

Universität Heidelberg

gorzitze, pado@c1.uni-heidelberg.de

Abstract

This paper presents a simple distributional method for acquiring event-denoting and object-denoting nouns from corpora. Its core is a bootstrapping cycle that alternates between acquiring new instances and new features, using a simple log odds ratio for filtering. We acquire 3000 German nouns for each class with precisions of 93% (events) and 98% (objects), respectively.

1 Events and Objects

While the majority of nouns in English and related languages refer to objects, either physical (*book*) or abstract (*idea*), some nouns refer to events and states (*christening, happiness*).¹

Being able to distinguish between these two classes is desirable for a number of reasons. From a model theoretic semantics point of view, event nouns and object should at the very least receive lexical entries of different types that mirror their different semantic behavior and associated information (event information vs. relational or qualia information). The distinction is also relevant for applications ranging from question answering (where entities and events usually correspond to different question types, cf. Hirschman and Gaizauskas 2000) to information extraction (where events are generally of primary importance, cf. Saurí et al. 2005) and to the modeling of reading times in psycholinguistics where the event/entity distinction often plays an important role (Traxler et al., 2002).

¹For simplicity, we will refer to the first class as *object nouns* and the second class as *event nouns*. We acknowledge that the event/object dichotomy is an oversimplification; see the discussion in Peris et al. (2012).

There is a great deal of literature on nominalizations and their ambiguities, but relatively little work on the acquisition of event and object nouns (but see Eberle et al., 2009; Peris et al., 2012). Ontologies like WordNet (Miller et al., 1990) distinguish events and objects, but coverage remains a problem even for English. For other languages, such resources are typically much smaller. For languages like German, the additional problem of productive *compounding* arises: it is impossible to cover all noun compounds in an ontology.

This paper addresses this problem with a simple method for acquiring event and object nouns. Its core is a corpus-based bootstrapping cycle which exploits distributional differences between the two classes, alternating between instances and features. It is largely language-independent; our evaluation on German shows promising results.

2 A Corpus-based Acquisition Method

2.1 Distributional Features

Our intuition is that event nouns refer to entities which have a *temporal dimension*, while object nouns refer to entities that do not (Peris et al., 2012). This conceptual difference is mirrored in the usage of event and object nouns and can thus be picked up with distributional semantics (Turney and Pantel, 2010). Within distributional models, it has been observed (Peirsman et al., 2008) that “loose” contexts (i.e., large bag-of-word contexts) tend to pick up semantic *relatedness* while “tight” contexts (small bag-of-words contexts or syntactic contexts) modeling semantic *similarity* better. Since the event/object distinction belongs to the second type, and since all target words are nouns, we consider three types of *syntactic* contexts.

The first type covers direct modifiers of the target nouns, namely adjectives, which typically refer

to properties of the nouns. Event nouns should therefore support adjectives that refer to temporal properties (*recent, frequent*) while object nouns occur with adjectives that refer to physical properties (*large, blind*). The second type covers occurrences of the target nouns as prominent arguments (subjects and objects) of verbs. Again, we expect that events occur preferably with verbs dealing with temporal or aspectual properties (occurring as the grammatical objects of transitive *begin, repeat, postpone*) while object verbs support a large variety of events (occurring as the grammatical objects of *drink, transport, love*). Finally, the third type covers occurrences of the target nouns embedded in larger NPs, such as *N of target*. As before, event nouns should occur in NPs with “temporal” heads such as *anniversary, period* while object nouns prefer nouns such as *collection, West*.

This approach comes with two main potential problems. The first one is the asymmetry between event and object nouns: event nouns should occur in restricted contexts, while object nouns support a wide variety of contexts. Furthermore, these contexts differ considerably among subgroups of object nouns (concrete vs. abstract objects). We will ignore this problem for the moment and assume a standard two-class classification process.

The second problem is the identification of predictive features. Clearly, using *all* verbs, adjectives, and nouns as features is infeasible. Furthermore, most of these features would be useless since they occur too infrequently, or because they can occur with both event and object nouns (e.g., *long* can refer to time length or physical dimensions). We require a method that can learn features directly from data and determine reliable ones.

2.2 A Bootstrapping Cycle

We approach the feature learning problem with the bootstrapping cycle shown in Figure 1. Bootstrapping has been applied to a variety of NLP tasks including lexical acquisition (Thelen and Riloff, 2002), question answering (Ravichandran and Hovy, 2002), and relation extraction (Pantel and Pennacchiotti, 2006). The idea is that knowledge about the nouns in a class can be used to acquire new features for the class, and vice versa.

Bootstrapping makes it possible to start out with a “seed” (in our case, a small set of either proto-

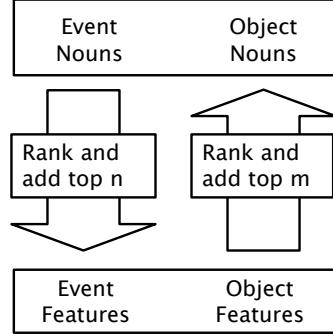


Figure 1: A bootstrapping cycle for learning event and object nouns as well as features for these classes

typical nouns or prototypical features) and acquire sets of nouns and features – in principle without further supervision. Crucial for the success of bootstrapping, however, is a sensible filtering mechanism, usually realized by way of a ranking function combined with selecting only the top n features and nouns, respectively. We follow Thelen and Riloff (2002) in taking advantage of a multi-class setup. Specifically, we score each feature f with respect to a class c using a simple log odds ratio: $\text{LOR}(f, c) = \log \frac{P(c|f)}{P(\neg c|f)}$ which measures how many times it is more likely that f indicates c than it does the other class.² In the inverse direction, we simply replace f by nouns n , employing the same formula to measure n ’s association with c : $\text{LOR}(n, c) = \log \frac{P(c|n)}{P(\neg c|n)}$.

3 Evaluation

3.1 Setup

We tested our model on the sdewac corpus. sdewac is a subset of the deWac corpus (Baroni et al., 2008), a web-crawled German corpus containing about 1.9M documents from 11,000 different .de subdomains. sdewac was drawn from deWac by removing duplicate and ungrammatical sentences, and parsing the remaining 900M tokens with a rule-based dependency parser (Schiehlen, 2003).

We seeded the bootstrapping cycle by manually specifying 100 noun lemmas for the event and object classes, respectively. To avoid that the cycle picks up features for specific domains rather than the event/object distinction, we included a wide range of nouns in both classes, based on the 26

²We use add-one smoothing.

	Events		Objects	
	IPrec	CPrec	IPrec	CPrec
Step 1	93.4	93.4	98.6	98.6
Step 2	93.5	93.5	98.0	98.3
Step 3	95.3	94.1	98.0	98.2
Step 4	93.6	94.0	95.2	97.5
Step 5	93.7	93.9	99.0	97.8
Step 6	93.1	93.8	99.1	98.0
Step 7	93.8	93.8	97.9	98.0
Step 8	92.1	93.6	98.6	98.1
Step 9	90.9	93.3	99.3	98.2
Step 10	90.0	92.9	99.0	98.3

Table 1: Precision of extracted event and object nouns: Values for individual batches (IPrec, 300 nouns each) and cumulative precision (CPrec)

WordNet “supersenses” / “lexicographer labels”.

The cycle ran for 10 iterations over the whole corpus, with n (the number of features selected in each bootstrapping step) set to 150 and m (the number of lemmas) to 300. This resulted in 1500 features and 3000 lemmas for each class.

3.2 Method

Given that we do not have complete lists of event and object nouns, it is hard to compute recall. Similar to work in relation extraction (Pantel and Pennacchiotti, 2006), this paper focuses on a precision-based evaluation.

In order to gauge the difficulty of assigning nouns to the event and object categories, we performed a pre-experiment on a small dataset of 25 nouns which were annotated by 4 annotators each as either events, nouns, or ambiguous. Using Fleiss’ κ (Fleiss, 1971), a measure of reliability appropriate for multiple annotators, we obtained an inter-annotator agreement of 0.76, which corresponds to substantial agreement. On the basis of this result, we decided to annotate the complete output of the method with single annotation, using the same annotation scheme as before. In the evaluation, we give full credit for each label of ambiguous nouns, also for minority senses.

3.3 Results

Table 1 shows our results. The two columns for each class list the individual precision of the 300 nouns acquired in each step (IPrec) and the cumulative precision of all nouns up to this step (CPrec). The results are fairly high across the board. With

figures around 98%, object nouns are easier to acquire than event nouns (in the low 90s), which is unsurprising since they form the majority class.³ The precision of objects remains at a high level for all ten steps. For events, IPrec remains between 93% and 94% up until step 7. It then begins to drop, however. It appears that current settings work well to acquire a “core set” of some 2000 events but becomes more unreliable afterwards.

3.4 Analysis

Table 2 shows a random sample of acquired nouns for both classes, including occasional errors such as **Religionsausübung (religious observance)* identified as an object. The high accuracy of the objects is due at least partly to the large number of concrete nouns in the object class which are easy to categorize. In contrast, abstract nouns are relatively rare. The list also contains a substantial number of compounds, highlighting the benefits of distributional analysis for this class.

Table 3 shows a sample of acquired features, which bear out our assumptions (cf. Section 2.1) rather well. Among the verb features for events, we find a number of aspectual verbs (*dauern / take time*) as well as of “scheduling” verbs (*vorverlegen / move forward*). Object nouns occur in agent-like positions (as grammatical subjects) and as grammatical objects of causative verbs (*waschen / wash*). Some of the nominal features are nominalizations of verbs (*Ableistung / serving, e.g. a jail sentence*). These are complemented by temporal nouns for events (*Vorabend / previous evening*) and person and physical position nouns for objects (*Bürgermeister / mayor*). Finally, almost all adjective features for events refer to durations or to participants of the event (*amtsärztlich / by an officially appointed doctor*) while adjective features for objects mostly express physical properties (*rund / round*) or are adjectival passives (“*Zustandspassive*”, *erworben / purchased*).

Finally, we sampled 800 correctly recognized nouns (400 events and 400 objects) and reclassified the nouns using three models that used just one feature type each. Table 4 evaluates them against the original 800-noun list. The noun model shows the worst results. The main culprit is a low

³In a random 100-word sample from the corpus, we found 56 object nouns, 32 event nouns, and 12 ambiguous nouns.

Events	Objects
Jahrestagung, Rundreise, Überprüfung, Exektion, Militärputsch, Auftauchen, Bergung, Pubertät, Freiheitsstrafe, Wiederkehr, Niederschlagung, Militäraktion, Wiedereröffnung, Auszählung, Praxissemester, Prophylaxe, Umfrage, Vorbereitungsphase, Feierstunde, Boom, Planungsphase, Währungsreform, Ratifizierung, Klausurtagung, 90er-Jahr, Machtkampf, Ersatzdienst, Tagung, Volksabstimmung, Ritt, Ultraschalluntersuchung	Antenne, Medaille, Nase, Mitmensch, Steuerzahler, Nadel, *Religionsausübung, Häuschen, Schlüssel, Wirtschaftsguts, Auge, Segel, Deckel, Nachbarstaat, Ministerin, Kapelle, Gefäß, Krankenkasse, Hase, Handschuh, Mitgliedsland, Tarifpartei, Konfliktpartei, Passagier, Beschwerdeführer, Linse, Schürze, Schwan, *Handlungsfähigkeit, Gebietskörperschaft, Flair, Fötus, 5tel, Ärztekammer, Elefant, Mehrheit, Gesundheitsamt, Eisenbahnlinie,

Table 2: Sample of acquired nouns

Verb features (events)	Verb features (objects)
anzetteln-OBJ, ableisten-OBJ, verstreichen-SUBJ, mitschneiden-OBJ, vertagen-OBJ, anberaumen-OBJ, jähren-SUBJ, verbüßen-OBJ, vorverlegen-OBJ, dauern-SUBJ	trinken-OBJ, erkranken-SUBJ, errichten-OBJ, investieren-SUBJ, engagieren-SUBJ, schütteln-OBJ, waschen-OBJ, erwerben-SUBJ, schöpfen-OBJ, spenden-OBJ, transportieren-OBJ
Noun features (events)	Noun features (objects)
Ableistung, Beendigung, Vorabend, Anmelder, Wirren, Schlußphase, Gräuel, Ausbruch, Veteran, Jahrestag, Vortag, Siegermacht, Versäumung, Verbüßung, Zurückverweisung, Ablegung, Ablauf	Seele, Osten, Beziehung, Ministerpräsident, Bürgermeister, Gründung, Erwerb, Auge, Wiederaufbau, Köpfen, Anstalt, Einwohner, Sohn, Wettbewerbsfähigkeit, Verkauf, Verabschiedung
Adjective features (events)	Adjective features (objects)
30jährig, erkennungsdienstlich, mündlich, medikamentös, anderweitig, monatelang, konzertant, amtsärztlich, ambulant, wochenlang, menschenunwürdig, physiotherapeutisch	mittelständisch, rund, behindert, rot, kreisfrei, ätherisch, tätig, erworben, gehandelt, hell, schwerbehindert, arm, beruflich, blau, interessiert, elektrisch, teilen, blind, gebildet, golden, ansässig

Table 3: Sample of acquired features

recall due to the low frequency of the “embedded NP” construction (cf. Section 2.1), but the precision is also imperfect: many nouns can embed events as well as objects (*Beziehung*, *Anmelder* (*relation*, *registrant*)). The verb model works substantially better. Notably, it has an almost perfect precision – there are verbs almost all subject (or objects, respectively) of which belong to one of the two classes. However, its recall is still fairly low. The best model is the adjective-based one, with the highest recall and an almost equal precision.

Open questions include the influence of target word frequency and domain. Given the space constraints, these must be left for future work.

4 Conclusion

We have presented a simple corpus-based method to learn event-denoting nouns and object-denoting nouns from corpora using a bootstrapping cycle to alternately acquire nouns and context features. Our method shows good results on German, but

	Recall	Precision	F ₁
Adjective features	82.89	92.86	87.59
Verb features	61.66	95.46	74.92
Noun features	17.80	72.95	28.62
All features	100	100	100

Table 4: Results of feature ablation analysis (evaluation measures relative to full model)

is essentially language-independent. It requires only a large parsed corpus and a seed set of nouns from both classes. Our feature ablation analysis indicates that full parsing may even be dispensable, as long as Adj-N-pairs can be identified reliably.

In the current paper, we have mostly ignored the issue of ambiguous nouns. In future work, we plan to apply our model to the disambiguation of nouns instances (rather than lemmas), which will involve considerably more sparsity.

Acknowledgments. This work was partially supported by the EC-funded project EXCITEMENT (FP7 ICT-287923).

References

- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2008. The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Kurt Eberle, Gertrud Faaß, and Ulrich Heid. 2009. Corpus-based identification and disambiguation of reading indicators for german nominalizations. In *Proceedings of Proceedings of the 5th Corpus Linguistics Conference*, Liverpool, UK.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Lynette Hirschman and Robert Gaizauskas. 2002. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.
- George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.
- Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of COLING/ACL*, pages 113–120, Sydney, Australia.
- Yves Peirsman, Kris Heylen, and Dirk Geeraerts. 2008. Size matters: tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41.
- Aina Peris, Mariona Taulé, and Horacio Rodríguez. 2012. Empirical methods for the study of denotation in nominalizations in spanish. *Computational Linguistics*. To appear.
- Deepak Ravichandran and Eduard H Hovy. 2002. Learning surface text patterns for a Question Answering System. In *Proceedings of ACL*, Philadelphia, PA.
- Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: a robust event recognizer for QA systems. In *Proceedings of HLT/EMNLP*, pages 700–707, Vancouver, BC.
- Michael Schiehlen. 2003. A cascaded finite-state parser for German. In *Proceedings of EACL*, pages 163–166, Budapest, Hungary.
- Michael Thelen and Ellen Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of EMNLP*, pages 214–221, Philadelphia, PA.
- Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in sentence processing: evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47(4):530–547.
- Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Artificial Intelligence*, 37(1):141–188.

KONVENS 2012

**PATHOS 2012: First Workshop on Practice and Theory of
Opinion Mining and Sentiment Analysis**

September 21, 2012
Vienna, Austria

Preface

The present proceedings contain the papers and invited talk presented at PATHOS-2012, the First Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis. PATHOS-2012 took place on September 21, 2012 in Vienna Austria, in conjunction with KONVENS-2012. Its goal was to provide a platform for researchers and developers in sentiment analysis and opinion mining to present their ongoing work, discuss open issues in the research area and address future challenges. PATHOS-2012 was initiated by IGGSA, the Interest Group on German Sentiment Analysis (iggsa.sentimental.li).

We proposed the following topics in the call for papers of the workshop:

- Semi-supervised/weakly supervised learning for sentiment analysis
- Contrast of machine learning vs. linguistic approaches vs. hybrid methods
- Sentiment analysis on twitter and social media in general
- Fine-to-coarse sentiment analysis
- Emotion detection and classification
- Representation of and calculus on emotions
- Multi-lingual sentiment analysis
- Lexical resources for opinion mining and sentiment analysis
- Evaluation gold standards and evaluation methodologies
- Real-world applications and large-scale sentiment analysis
- Trends and perspectives in the field

We received 12 submissions (7 long papers and 5 short papers) out of which we accepted 8 papers (5 long papers and 3 short papers). 4 of the submissions came from Germany, 3 from France, 2 from Switzerland and one each from the Czech Republic, Denmark and Italy. The high number of these international submissions underlines the significance of sentiment analysis in natural language processing.

In addition to the presentation of the technical papers, the workshop also included an invited talk given by Carlo Strapparava on Emotions and Creative Language.

We would like to thank the authors of the papers for their interesting contributions, the members of the program committee for their insightful reviews, and the presenter of our invited talk, Carlo Strapparava, for accepting the invitation to give a talk at the workshop. We also thank our Gold Sponsor TrendMiner (www.trendminer-project.eu) and our Silver Sponsor WebLizard (www.weblyzard.com), who provided financial support crucial for the success of PATHOS-2012. Eventually, we are also grateful to the organizers of KONVENS-2012 to support us at the different stages of organizing this workshop.

Stefan Gindl, Robert Remus and Michael Wiegand
Workshop Organizers

September 2012

Workshop Organizers:

Stefan Gindl (MODUL University Vienna, Austria)
Robert Remus (University of Leipzig, Germany)
Michael Wiegand (Saarland University, Germany)

Program Committee:

Alexandra Balahur (European Commission Joint Research Centre, Italy)
Simon Clematide (University of Zurich, Switzerland)
Manfred Klenner (University of Zurich, Switzerland)
Johann Mitlöhner (Vienna University of Economics and Business, Austria)
Stefanos Petrakis (University of Zurich, Switzerland)
Josef Ruppenhofer (Hildesheim University, Germany)
Manfred Stede (Potsdam University, Germany)
Ralf Steinberger (European Commission Joint Research Centre, Italy)
Veselin Stoyanov (Johns Hopkins University, USA)
Cigdem Toprak (Technische Universität Darmstadt, Germany)
Ulli Waltinger (Siemens AG, Corporate Technology, Germany)
Albert Weichselbraun (University of Applied Sciences Chur, Switzerland)
Cäcilia Zirm (University of Mannheim, Germany)

Additional Reviewers:

Jasna Tusek Weichselbraun (Vienna University of Economics and Business, Austria)

Invited Speakers:

Carlo Strappavara (Fondazione Bruno Kessler, Italy)

Sponsors:



Workshop Program

Friday, September 21, 2012

- 09:00–09:10 Welcome
- 09:10–10:10 Invited talk: *Emotions and creative language*
Carlo Strappavara
- 10:10–10:30 *Unsupervised sentiment analysis with a simple and fast Bayesian model using Part-of-Speech Feature Selection*
Christian Scheible and Hinrich Schütze
- 10:30–11:00 Coffee break
- 11:00–11:30 *Sentiment analysis for media reputation research*
Samuel Läubli, Mario Schranz, Urs Christen and Manfred Klenner
- 11:30–12:00 *Opinion analysis: The effect of negation on polarity and intensity*
Lei Zhang and Stéphane Ferrari
- 12:00–12:20 *Domain-specific variation of sentiment expressions: A methodology of analysis in academic writing*
Stefania Degaetano, Ekaterina Lapshinova-Koltunski and Elke Teich
- 12:20–13:50 Lunch break
- 13:50–14:20 *Creating annotated resources for polarity classification in Czech*
Kateřina Veselovská, Jan Hajič, Jr. and Jana Šindlerová
- 14:20–14:50 *A phrase-based opinion list for the German language*
Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari and Nikolaos Korfiatis
- 14:50–15:10 *Opinion mining in an informative corpus: Building lexicons*
Patrice Enjalbert, Lei Zhang and Stéphane Ferrari
- 15:10–15:45 Coffee break
- 15:45–16:15 *What is the meaning of 5*'s? An investigation of the expression and rating of sentiment*
Daniel Hardt and Julie Wulff
- 16:15– “World Café”

Emotions and Creative Language

Carlo Strappavara

Fondazione Bruno Kessler

Abstract

Dealing with creative language and in particular with affective, persuasive and possibly humorous language has been considered outside the scope of computational linguistics. Nonetheless it is possible to exploit current NLP techniques starting some explorations about it. We briefly review some computational experience about these typical creative genres. In particular we will focus on research issues relevant to how affective meanings are expressed in natural language, and it will introduce techniques for affective content detection and generation. As far as persuasive language is concerned, after the introduction of a specifically tagged corpus (exploiting for example public political speeches), some techniques for persuasive lexicon extraction and for predicting persuasiveness of discourses are provided. We conclude the talk showing some explorations in the automatic emotion recognition exploiting a combination of music and lyrics.

Unsupervised Sentiment Analysis with a Simple and Fast Bayesian Model using Part-of-Speech Feature Selection

Christian Scheible and Hinrich Schütze

Institute for Natural Language Processing

University of Stuttgart

scheibcn@ims.uni-stuttgart.de

Abstract

Unsupervised Bayesian sentiment analysis often uses models that are not well motivated. Mostly, extensions of Latent Dirichlet Analysis (LDA) are applied – effectively modeling latent class distributions over words instead of documents. We introduce a Bayesian, unsupervised version of Naive Bayes for sentiment analysis and show that it offers superior accuracy and inference speed.

1 Introduction

Unsupervised models for sentiment analysis remain a challenge. While sentiment is relatively easy to detect in supervised experiments (e.g. (Pang et al., 2002)), models lacking supervision are often unable to make the distinction properly. In some domains (e.g. book reviews), topics strongly interfere with sentiment (Titov and McDonald, 2008), and unsupervised models barely beat the chance baseline (Dasgupta and Ng, 2009).

Bayesian models have received considerable attention in natural language processing research. They enable the incorporation of prior knowledge in arbitrary graphical models while parameter inference can be accomplished with simple sampling techniques (Gelman et al., 2004). In this paper, we apply a Bayesian model for unsupervised sentiment analysis of documents. Although various unsupervised Bayesian sentiment models exist, almost all of them extend Latent Dirichlet Analysis (LDA, (Blei et al., 2003)) which was designed to model document-specific topic mixtures. While LDA is a well-understood and widespread model, it is not well-suited for labeling

documents. Instead, we propose a model that generates a single label per document which generates all words in it, instead of generating multiple word labels per document. Naive Bayes, which is commonly used supervisedly, meets this requirement. We will show that the unsupervised version of Naive Bayes with Bayesian Dirichlet priors achieves a higher classification accuracy than LDA on standard review classification tasks. In addition, we will demonstrate a significant speed advantage over LDA.

This paper is structured as follows: Section 2 describes related approaches, Section 3 contains model definition, Sections 4 and 5 presents the experimental setup and results.

2 Related Work

Most related work on Bayesian models for sentiment uses LDA-style models that predict one label per word. A notable exception is (Boyd-Graber and Resnik, 2010) who use document-level labels. Their focus however lies on supervised multilingual classification and they do not compare their model against any reference model.

Zagibalov and Carroll (2008) introduce an unsupervised model using lexical information about modifiers like negation and frequent adverbials. They automatically induce a lexicon of relevant seed words and report high classification accuracy. The model requires hand-crafted language-specific knowledge.

Dasgupta and Ng (2009) present a spectral clustering approach that yields highly competitive results. The method requires some human interaction: Selecting the best eigenvector automatically is difficult, so the authors have it manually selected, boosting the results. This constitutes a

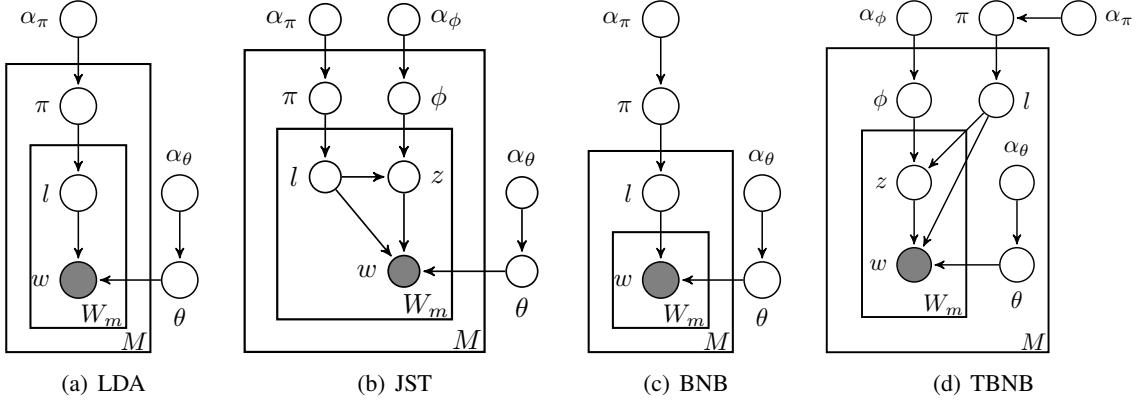


Figure 1: Four Bayesian topic models

form of supervised model selection.

Lin and He (2009) present the Bayesian joint sentiment-topic model (JST). It is an extension of LDA as it contains an intermediary topic layer. The authors experiment with both unsupervised and lexically supervised setups. Unfortunately, no advantage of using the additional layer could be demonstrated. We will investigate this model more closely in the following sections and compare it to our proposed model.

3 Bayesian models for document classification

We compare four Bayesian models: Latent Dirichlet Analysis (LDA, Blei et al. (2003)) presented as the latent sentiment model (LSM) by Lin et al. (2010), the joint sentiment-topic model (JST) by Lin and He (2009) that introduces an additional topic layer, Bayesian Naive Bayes (BNB) as used by Pedersen (1997), and TBNB, an extra-topic-layer version of BNB analogous to JST. We will give a short definition for each of the models in the following section. We refer the reader to the respective papers for details. Note that the terminology used in previous research is conflicting. We will thus refer to latent document or word sentiment classes as *labels* (l) and other latent classes as *topics* (z). Hyperparameters are generally referred to as α_x meaning that α is the hyperparameter for prior distribution of the multinomial with parameters x .

3.1 Model definitions

LDA. LDA (Figure 1(a)) is a model of label distributions over words in documents. Each document has a multinomial label distribution specified by

θ with a Dirichlet prior with hyperparameters $\vec{\alpha}$. As labels are inferred for words, obtaining a document label requires an additional step: The document label is the most probable word label in the document by majority vote. Although presented under a different name in related work, we will refer to the model as LDA to avoid confusion.

JST. The JST model (Figure 1(b)) is an extension of LDA. Applied to reviews, LDA usually finds topics instead of sentiment (Titov and McDonald, 2008). For that reason, JST contains both labels and topics that are intended to steer the sentiment classification. The number of topics T and the number of labels K are set separately. LDA is a special case of JST since setting either $K = 1$ or $T = 1$ removes the additional latent level.

BNB and TBNB. BNB (Figure 1(c)) is the Bayesian extension of the Naive Bayes model (Pedersen, 1997). It has a global multinomial label distribution θ with a Dirichlet prior and generates one label for each document. The generative story of BNB is:

- Choose a label distribution $\pi \sim \text{Dir}(\alpha_\pi)$
- Choose a label $l_d \sim \text{Multinomial}(\pi)$ for each document document d
- Choose each word w_i in d from $p(w_i|l_d)$

Analogously to JST, we define an extension of BNB that adds a layer word-level topics (Bayesian Naive Bayes with topics, TBNB). Again, TBNB becomes BNB when the number of topics T is set to 1. We will later show that $T = 1$ is actually the best choice for JST and BNB.

3.2 Parameter estimation

We use the Hierarchical Bayes Compiler (HBC, Daumé III, 2008) which implements Gibbs sam-

pling (Geman and Geman, 1984) to estimate the model parameters. For the LDA and JST model, collapsed sampling is possible where the continuous parameters of the models (θ , π , ϕ) are marginalized out. For the BNB model, this would lead to a violation of independence assumptions, making explicit sampling of the parameters by prior updating necessary (cf. (Resnik and Hardisty, 2010) for details). We did not find a difference in accuracy between the collapsed and uncollapsed LDA versions.

4 Experimental setup

Data preparation. We chose two standard datasets for a comparable evaluation: the movie review (MR, (Pang et al., 2002)) and multi-domain sentiment datasets (MDS, (Blitzer et al., 2007)), each containing 1000 positive and 1000 negative documents per domain, with the MDS containing data from multiple domains (**Books**, **DVDs**, **Electronics**, and **Kitchen**).

We intend to reduce the feature space based on parts of speech. The processed form of multi-domain dataset is unsuitable for this as all texts were lowercased. We reconstructed the reviews from the raw files offered by the authors by extracting the full text from HTML and applying the same sentence splitter. We will make the reconstructed data publicly available.

Feature selection. Naive Bayes can be sensitive to uninformative features, making feature selection desirable. Previous work on sentiment classification showed that certain part-of-speech classes are highly informative for sentiment analysis, e.g. Pang et al. (2002) report high results when using only adjective features. Since the movie and product reviews differ considerably in length (746 and 158 average tokens/document, respectively), we retain more features for the MDS than for the MR dataset. Feature representations contain only adjectives for MR and adjectives, adverbs, verbs, modals, and nouns (Penn Treebank tags JJ.*, MD, NN.*., RB.*., VB.*.) for MDS. We tag the documents with the Mate tagger (Björkelund et al., 2010). In addition, we remove stopwords and all features that occur less than 100 times in the corpus to clean the feature set and speed up computation.

Sampling. Latent class priors are set to $\frac{100}{N}$,

		MDS						
		MR		B	D	E	K	avg.
all feat.	LDA	63.5	51.3	53.4	59.5	57.2	55.4	
	BNB	61.2	51.9	53.0	61.4	62.8	57.3	
selection	LDA	60.3	54.1	53.3	54.7	54.6	54.2	
	BNB	70.4	57.8	56.1	64.1	65.2	60.8	

Table 1: Average accuracy (%) for each dataset

with N being the number of classes, and all word priors are set to 0.001. Priors are symmetric. Since both datasets contain two classes, positive and negative, K is set to 2. We vary the number of topics T to study its effect. We sample for 1000 iterations and report accuracies averaged over 10 runs since model quality may vary due to random sampling.

5 Experiments

This section describes our experiments on feature selection, the number of topics T , and the computation times of the models.

Feature selection We build models for (i) the full data (without stopwords and infrequent words) and (ii) with our feature selection. We try the least complex model first and set $T = 1$. As shown in Table 1, when using all features, BNB performs about equal to or slightly better than LDA except on the MR data. However, after applying feature selection BNB outperforms LDA in all cases. Feature selection shortens the documents which affects LDA negatively (Titov and McDonald, 2008), a problem whose solution would require additional computational effort (e.g. by modeling topic transitions (Blei and Moreno, 2001)). Conversely, BNB behaves as expected from a Naive Bayes model – feature selection improves the results. Errors can occur because of frequency effects: common words like *be*, *have*, ... receive high probabilities. Another problem is that many words are mistagged, making proper selection more difficult – particularly on the MDS where sloppy orthography is frequent. Normalization might correct this, although ungrammatical sentences might still produce erroneous results.

Number of topics. We are interested in the effects of the additional topic layer. Lin et al. (2010) do not perform an evaluation of the number of

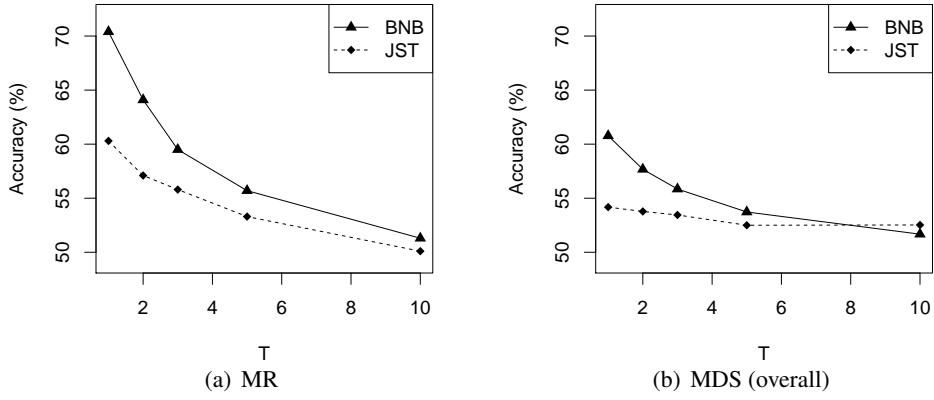


Figure 2: Classification accuracies for different values of T

MDS						
MR		B	D	E	K	avg.
all feat.	LDA	1064	209	226	190	141
	BNB	728	149	156	103	110
selection	LDA	163	109	115	79	67
	BNB	109	79	87	47	50
						66

Table 2: Average inference time (sec) for each dataset

topics in the unsupervised case and their results for lexically supervised classification indicate decreasing performance for higher topic numbers. To this end, we run experiments for values of T between 1 and 10 for both TBNN and JST. Note that the models simplify to BNB and LDA, respectively, if $T = 1$ since the class probabilities are then all conditioned on the same topic which essentially leads to no conditions at all.

Figure 2 shows the classification accuracy for each dataset individually and an overall average accuracy for the MDS data. We can observe that just like in the lexically supervised case, the additional topic layer leads to a decline in accuracy when more topics are introduced, with TBNN being more sensitive than JST. We achieve the best results for models with $T = 1$, where the best TBNN setup beats the best JST setup by 6.9% on the MR data and 5.4% on the MDS. Note that the best setup for TBNN uses feature selection while the one for LDA does not. We will briefly examine implications on the computational efforts.

Computation time. BNB is the better model for document classification because it models document labels instead of word labels. Conversely, LDA needs to estimate more latent classes than needed (one per word instead of only

one per document). This leads to higher computational effort which is unjustified as it is not reflected through better classification results. We measured the average time used for inference and labeling on an Intel Xeon 3.33 GHz CPU. We only report numbers for $T = 1$ as models with more topics are less accurate. Note however that these models can take significantly more time to compute since more label distributions need to be estimated. Table 2 shows the average inference time in seconds for each dataset. Using BNB saves 1/3 of computation time compared to LDA. Since LDA can only produce competitive results when run with all features, the differences become more drastic when comparing lines 1 and 4 of the table, yielding reductions up to around 90% on MR and 50% on MDS. Using a model with feature selection is thus even more desirable if efficiency is an issue.

6 Conclusion and Future Work

We presented Bayesian Naive Bayes, a Bayesian model for unsupervised document classification. We showed that BNB is superior to the LDA model on the standard unsupervised sentiment classification task. In future work, we would like to examine the behavior of our model in a semi-supervised setting where some document or feature labels are known.

Acknowledgements

This work was funded by the DFG as part of the SFB 732 project D7. We thank Thomas Müller for his helpful comments.

References

- Anders Björkelund, Bernd Bohnet, Love Hafnell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.
- David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 343–348, New York, NY, USA. ACM.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 3.
- J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440.
- Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sajib Dasgupta and Vincent Ng. 2009. Topic-wise, sentiment-wise, or otherwise? Identifying the hidden dimension for unsupervised text classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 580–589, Singapore, August. Association for Computational Linguistics.
- H. Daumé III. 2008. hbc: Hierarchical bayes compiler. *Pre-release version 0.7, URL <http://www.cs.utah.edu/~hal/HBC>.*
- A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. 2004. *Bayesian data analysis*. CRC press.
- S. Geman and D. Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.
- Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.
- Chenghua Lin, Yulan He, and Richard Everson. 2010. A comparative study of bayesian models for unsupervised sentiment detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 144–152, Stroudsburg, PA, USA. Association for Computational Linguistics.
- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL-EMNLP 2002*, pages 79–86.
- Ted Pedersen. 1997. Knowledge lean word sense disambiguation. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 814–814. AAAI Press.
- Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical report, University of Maryland.
- Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 111–120.
- Taras Zagibalov and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 1073–1080, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sentiment Analysis for Media Reputation Research

Samuel Läubli^φ Mario Schranz^γ

^φInstitute of Computational Linguistics,
University of Zurich
Binzmühlestrasse 14
8050 Zürich
{laeubli,klenner}
@cl.uzh.ch

Urs Christen^γ Manfred Klenner^φ

^γCenter for Research on the Public
Sphere and Society, University of Zurich
Andreasstrasse 15
8050 Zürich
{mario.schranz, urs.christen}
@foeg.uzh.ch

Abstract

As a subtask of qualitative media reputation research, human annotators manually encode the polarity of actors in media products. Seeking to automate this process, we have implemented two baseline classifiers that categorize actors in newspaper articles under six and four polarity classes. Experiments have shown that our approach is not suitable for distinguishing between six fine-grained classes, which has turned out to be difficult for humans also. In contrast, we have obtained promising results for the four class model, through which we argue that automated sentiment analysis has a considerable potential in qualitative reputation research.

1 Introduction

While opinion mining techniques have been successfully implemented in large-scale appliances such as social media monitoring on the web (e.g., Godbole et al. (2007), Chen et al. (2012)), detailed studies in the field of reputation research still raise the need for human assessments. By focussing on a sub-task of qualitative media reputation analysis—identifying fine-grained actor polarities in full newspaper articles—we seek to examine whether automated sentiment analysis approaches can be used in traditional encoding workflows.

From a pragmatic perspective, supporting encoding processes has the potential of saving precious time for annotation experts, be it by pre-

selecting texts for further examination or suggesting classifications for targeted variables, such as the centrality or polarity of the reputation objects in scope. For sentiment analysis research on the other hand, we see opportunities to evaluate selected methods in a real-world scenario. In our current work, we primarily seek to explore the feasibility of employing more fine-grained classes than the traditional trichotomic distinction between *positive*, *neutral*, and *negative* polarities in automated sentiment analysis.

First, we give an introduction to traditional media reputation analysis and detail the resources we use for our experiments. In section 3, we introduce a lightweight approach to sentiment composition, which we implemented in a prototype classifier for the polarity of actors in newspaper articles. The evaluation of our system is presented in section 4 and subsequently discussed in section 5. Finally, we conclude our report in section 6, also listing further work that is planned or currently pursued.

2 Background

From the very beginning, text classification has been a very active research direction in the area of sentiment analysis. However, the focus of attention was mostly on the classification of product or movie reviews (e.g., Hu and Liu (2004)). There are a few exceptions, e.g., work based on the MPQA corpus (Wilson et al., 2005), where newspapers are dealt with.

Most of the time, a three-partite classification

Class	# Events	in %
neutral	45'018	48.5
controversial	14'096	15.2
negative, explicit	13'251	14.3
negative, implicit	9'977	10.8
positive, explicit	6'244	6.7
positive, implicit	4'236	4.6
Total	92'822	100.0

Table 1: Class Distribution in the Media Sample Corpus

is carried out: a text is either *positive*, *negative* or *neutral*. In contrast, we cope with four and even six classes, among which is a rather demanding class for controversial texts. Moreover, the classes of *negative* and *positive* are split into the more fine-grained distinction of *implicit* and *explicit*, respectively. This makes a challenging demand of our application scenario—media reputation analysis.

We are also in the tradition of Moilanen and Pulman (2007) and, more basically, Polanyi and Zaenen (2004), since we regard the notion of compositionality as crucial for the analysis of advanced texts. Instead of the elaborated syntax- and rule-based approach of Moilanen and Pulman (2007), our approach (described in section 3) is closer to the grammar independent one proposed by Choi and Cardie (2008).

In the remainder of this section, we give an introduction to our application domain and list resources we rely on for our present work.

2.1 Media Reputation Analysis

The Center for Research on the Public Sphere and Society (*fög*) of the University of Zurich has analyzed the media reputation of Swiss companies since 1998. Media reputation is defined by Deephouse (2000, p. 1097) as “the overall evaluation of the firm presented in the media resulting from the stream of media stories about the firm”. Reputation arises and decays wherever information about the trustworthiness of an actor circulates in arenas of public communications, be it in the traditional mass media or the new internet-based media. Measurement instruments to determine relevant reputation dynamics must therefore necessarily be based on an analysis of public commu-

Actor	Category	# Events	in %
UBS	bank	43'440	46.8
Crédit Suisse	bank	30'662	33.0
ZKB	bank	5'897	6.4
Swisscom	telecom	5'070	5.5
Novartis	pharma	3'270	3.5
Roche	pharma	2'381	2.6
Cablecom	telecom	1'637	1.8
Sunrise	telecom	465	0.5
Total		92'822	100.0

Table 2: Actors in the Media Sample Corpus

nifications.

The *fög* has conducted a quantitative-qualitative media content analysis (Eisenegger et al., 2010). It is aimed at determining the media reputation of the 39 largest Swiss companies on a daily basis in thirteen leading Swiss media. Accordingly, the most significant Swiss business sectors such as banking, insurance, pharmaceuticals, telecommunications, as well as manufacturing, food, and retail are part of this monitoring. The content analysis examines how frequently and strongly (centrality) the media report on specific companies—to which we refer as *actors*—and how they were evaluated (polarity). The recorded encodings (*positive*, *neutral*, *negative*, and more fine-grained sub-classes) allow the *fög* to build a Media Reputation Index (Eisenegger and Imhof, 2008). The content analysis is focussed on a sample of thirteen major Swiss opinion-forming media, covering both key print media as well as newscasts by public service broadcasters.

2.2 Resources

Our current work is primarily based on two central resources: a big sample of manually encoded newspaper articles, stemming from the *fög* content analysis, and an unweighted sentiment lexicon. Both are written in or for German, respectively.

2.2.1 Media Sample Corpus

We extracted a sample corpus from the *fög* database of encoded texts (see section 2.1). It comprises newspaper articles that each include at least one of the actors listed in table 2. All ar-

Class	# Positive	# Negative	# Neutral
adjectives	1'677	2'097	91
nouns	1'202	2'144	595
verbs	528	1'001	2
Total ^a	3'407	5'242	688

^a Besides 9337 polar entries for adjectives, nouns, and verbs, the lexicon comprises 61 base forms with polarity functions (shifter, intensifier, diminisher).

Table 3: Polar Base Forms in the Polarity Lexicon (Clematide and Klenner, 2010)

ticles were published and encoded between 1998 and 2011.

Newspaper articles can contain more than one actor, hence we use the term *event* whenever we refer to an encoded occurrence of an actor in a text. The presented corpus consists of 85'817 articles that contain 92'822 events altogether; an article features 1.08 events on average (std. dev = 0.36). We point out that due to this fact we cannot just classify the polarity of a text in order to derive the polarity of its actor(s), which is a further challenge for automating the classification.

Despite the fine-grained annotations, there is no structural information stored for the articles in our sample corpus. Although titles, leads, and image captions could be particularly informative for classification, we regard this as a given constraint and leave according experiments to future work.

Used as a training set for our learning algorithm and a gold standard for our overall evaluation, the media sample corpus makes the cornerstone of our present study. Section 4 gives further details on how the corpus was used for training and evaluation.

2.2.2 Polarity Lexicon

As a further resource, we use a polarity lexicon compiled by Clematide and Klenner (2010). It contains 9'398 base forms (see table 3), each of which has either assigned a polarity—*positive* (+), *neutral* (=), *negative* (−)—or a polarity function—*shifter* (⊖), *intensifier* (<), *diminisher* (>). As we pursue an unweighted approach for sentiment computation, we did not use intensifiers and diminishers, and neither did we use any weights associated with polar base forms.

In the following section, we describe how the

sentiment lexicon is used for the purpose of sentiment composition, and we give an example of how the approach can be used to generate features for machine learning algorithms.

3 Method

While human experts can rely on rather loosely defined coding instructions and their world knowledge to classify an actor’s polarity, classification systems call for a set of well-defined features that capture characteristics of the samples to be classified. For example, the *fög* class definition of *positive*, *implicit*—“the reputation object is discussed in a positively connotated context”¹—needs to be translated into computable features such as “a context is defined as a sentence unit”, “a context is positive if it only contains positive words” or “an actor is implicit to a context if it is not directly mentioned, but strongly related to the context’s subject”. Such rules are clearly error-prone; they are only heuristic, and in addition, they need to cope with noise caused by preceding system components such as a syntactic parser, thus raising the need for, e.g., a machine learning algorithm.

We implemented a lightweight sentiment analysis approach in our prototype, which is explained in the following section.

3.1 Lightweight Lexical Sentiment Composition

Moilanen and Pulman (2007) have proposed a model for calculating global polarities of syntactic phrases via their subordinate constituents. The model is based on sentiment lexica that assign prior polarities to leaf constituents, which are subsequently propagated or reversed by applying a considerable number of rules for weighting, filtering, and conflict resolution.

Aiming at robustness and simplicity, we propose a sentiment composition approach that is entirely based on an unweighted sentiment lexicon and head-dependencies of candidate words (tokens).

¹“Das Reputationsobjekt wird in einem positiv konnotierten Kontext thematisiert.” (*fög*, 2011)

3.1.1 Rules

Let $t_{1..n}$ be tokens of a candidate word sequence T . An initial *polarity-marker* function assigns prior polarity tags m to all $t \in T$ that are contained in the lexicon:

$$\text{polarity_marker} : t \rightarrow t_m \in \{+,=,-,\neg\}$$

The model consists of two simple rules that operate on accordingly tagged and dependency-parsed word sequences:

- (A) Each shifter t_\neg can invert its parent $t_{+/-}$ exactly once.
- (B) Each polar token $t_{+/-}$ can change its parent $t_{+/-}$ exactly once:
 - (i) child t_+ parent $t_=$ \rightarrow parent t_+
 - (ii) child t_- parent $t_=$ \rightarrow parent t_-
 - (iii) child t_- parent t_+ \rightarrow parent t_-

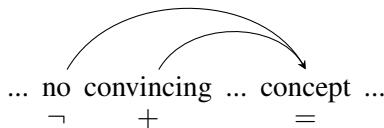
Rules (A) and (B) are applied repeatedly until convergence, that is, until no marked token t can be altered any further. Composition rules (i)–(iii) are taken from the pattern-matching approach described in (Klenner et al., 2009). Note that we do not list relationships where none of the tokens need to be altered, e.g., $t_+ + t_+ \rightarrow t_+$.

3.1.2 Example

To give an example, we look at a sentence occurring in our sample corpus of encoded newspaper articles (see sections 2.1 and 2.2.1):

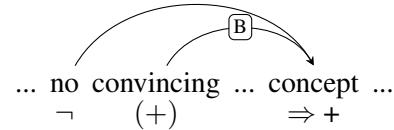
Particularly if no convincing managerial concept is at hand, which the UBS with brutal openness admitted to be the case this week.²

Applying the initial steps of our feature extraction pipeline—dependency parsing and polarity marking—yields the following structure for the first part of the above input sentence:

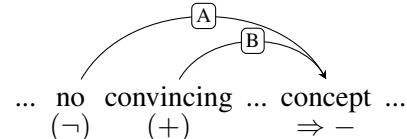


²German original: “Vor allem dann, wenn kein überzeugendes betriebswirtschaftliches Konzept auf dem Tisch liegt, was die UBS diese Woche mit brutaler Offenheit zugab.”

Next, rules (A) and (B) (see section 3.1.1) are applied in turn. As rule (A) cannot be applied in the first iteration (shifters can only invert polar tokens $t_{+/-}$), (B) is fired first:



In the next iteration, (A) can be applied to “kein_” as its *regens* is now polar:



As rules (A) and (B) can only be applied once for each child-head token relationship, the final state is reached for this polar chunk.

3.2 Features for Machine Learning

We form n -grams from polar dependency chunks in order to use our sentiment composition approach for machine learning. To avoid data sparseness we limit n to 2; longer sequences are split into multiple bigrams. In short, our features are constructed as follows:

- (i) *zero gram*: the polarity of the head token,
e.g., NEG
- (ii) *unigram*: the head token and its polarity,
e.g., NEG_concept
- (iii) *bigram*: (ii) plus the child token,
e.g., NEG_concept_convincing

To take actor proximity into account, we append $-S$ to a feature whenever an actor is present in the same sentence, e.g. NEG_concept_convincing-S. As there is no structural information such as title or image caption sections available for the texts in our sample corpus, further contextual information cannot be considered at this point.

We use polar n -grams as feature names (dimensions) and corresponding absolute counts as feature values. For example, encountering “no convincing concept” in an article would raise the value of NEG_concept_convincing-S by 1

(default value: 0). In order to give more weight to long polar chunks, all lower-order n -gram counts are also increased when adding a uni- or bigram.

To sum up, a text consisting of nothing but the sample sentence from section 3.1.2 would result in the following features:

Feature Name	Value
NEG_concept_convincing-S	1
NEG_concept_no-S	1
NEG_concept-S	2
NEG_convincing-S	1
NEG_no-S	1
NEG_openness_brutal-S	1
NEG_openness-S	1
NEG_brutal-S	1
NEG-S	8

4 Evaluation

In order to explore the feasibility of automating media reputation analysis processes, we have implemented a prototype system that classifies the polarity of actors in newspaper articles. In this section, we present the results of our corresponding evaluations.

4.1 Prototype Implementation

We have set up a feature extraction pipeline that processes digital newspaper articles. After handing an article’s full text to a dependency parser (Sennrich et al., 2009), the parser output is converted into CG format (Constraint Grammar; (VISL-group, 2008)) for subsequent enrichment by the *polarity marker*, a compiled VISL constraint grammar that adds prior polarities and polarity functions to single words. This corresponds to the *polarity_marker* function explained in section 3.1.1. Next, the marked CG serialization is handled by the polarity composition component (also VISL-based), and finally, all accordingly derived polarity chunks are converted into a set of features suitable for machine learning algorithms (see section 3.2).

For our prototype implementation, we abstracted all actors from the input texts for training and evaluation, i.e., all occurrences of actor names such as “UBS” or “Crédit Suisse” were replaced by an arbitrary token (“ACTOR”). In this way, we ensure that our classifiers can evaluate

any actor in principle, given their name and an optional list of synonyms. This optimizes flexibility in real-world scenarios, but may lower recall in cases where actor synonyms are not recognized in the replacement process.

Although we did not focus our efforts on optimizing quantitative performance, our prototype pipeline runs reasonably fast with the dependency parser being the only “bottleneck” in speed. On average, parsing an article of the media sample corpus took 4.9 seconds³ in our experiments, while passing it through the remainder of the pipeline (polarity marking, composition, feature extraction, and classification) took another 0.4 seconds.

4.2 Evaluation Method

As mentioned in section 2.2.1, our gold standard consists of 92’822 events in 85’817 newspaper articles (see table 1). It was produced by *various* annotators over the last 14 years. For the purpose of our evaluation, we asked a *single* expert to re-encode a random sample of 200 articles (220 events). Re-encoding took place in the expert’s usual working environment. We used a dedicated web application to collect all annotations and corresponding time stamps.

In this way, the evaluation task was the same for the human expert and our system: reproducing the original gold standard annotations for the random sample of texts. Since there is no inter-annotator agreement known for our gold standard (i.e., the media sample corpus), the performance of the human annotator on the 200 texts may indicate how hard the classification task at hand actually is.

4.3 Experiment 1: 6 Classes

The current revision of the *fög* coding manual (2011) lists six polarity classes (see table 1). Using all but the separated evaluation articles of our sample corpus, we trained a *Naive Bayes* classi-

³All times were measured using a standard Unix server (24x2.3GHz CPU, 128GB RAM). The prototype pipeline runs comparably fast on simple workstation computers.

Accuracy:	Human ₆			System ₆			Accuracy:	Human ₄			System ₄		
	59.5			51.9				66.8			57.4		
Class	P	R	F	P	R	F	Class	P	R	F	P	R	F
neutral	86.3	67.0	75.4	61.9	86.0	72.0	neutral	86.3	67.0	75.4	62.8	81.9	71.1
controversial	33.9	71.4	46.0	26.3	20.0	22.7	controversial	33.9	71.4	46.0	25.3	25.0	25.2
negative, exp.	61.9	78.8	69.3	37.0	30.8	33.6	negative	78.7	72.5	75.5	68.6	52.9	59.8
negative, imp.	40.0	11.1	17.4	25.5	10.7	15.1	positive	58.3	61.8	55.3	42.9	21.1	28.3
positive, exp.	40.0	57.9	50.0	24.6	12.0	16.2							
positive, imp.	33.3	15.8	21.4	8.0	0.5	0.9							

Table 4: Evaluation of Human and System Classification Accuracy on a Test Set of 200 Full Newspaper Articles

fier⁴ that assigns these class labels to all actors in candidate texts, based on our feature pipeline described in section 4.1.

Table 4 (left portion) shows the results for the six class experiment. For each class, we list precision (P), recall (R) and balanced f-score (F). Overall classification accuracy is indicated at the top. We used 200 full newspaper articles containing 220 events for this experiment, i.e., the same events were labelled by *fög* experts (Human₆) as well as our classifier (System₆). On average, classification took 56.3 seconds per event (Human₆) and 4.8 seconds (System₆) respectively.

4.4 Experiment 2: 4 Classes

In a second experiment, we folded *implicit* and *explicit* into one class each for *positive* and *negative* (see table 4, right portion). This addresses the low scores that were obtained especially for *implicit* polarities in both human and system classification. All other conditions were left unchanged.

The findings of our experiments are discussed in the following section.

5 Discussion

Automatically assigning fine-grained classes to actors turned out to be an all but trivial task in our wide domain. This is reflected in the evaluation results of our six class model (System₆), which performs considerably less accurate than a human annotator. The system assigns too much probability to *neutral* events—an obvious drawback of

using a *Naive Bayes* classifier, which elevates the most frequent class in the gold standard because of its high *a priori* probability—resulting in high recall for *neutral*, but lowering precision for this class and recall for all other classes. With f-scores below 35% for all non-neutral classes, classifying actors by use of our proposed feature pipeline is clearly not promising.

Still, our first experiment sheds light on how fragile it is to classify in a fine-grained mode even for experienced annotators. For Human₆, four out of six classes feature f-scores of 50% or lower, hinting that there are ambiguous cases that are difficult to resolve for humans also. This could indicate that “soft” (continuous) boundaries are more suitable than clearly delimitable (nominal) class boundaries when assessing the polarity of an actor in a wide context.

In our second experiment, we have folded the *positive* and *negative* classes. Removing the somewhat “blurry” distinction between *implicit* and *explicit* classes had a positive effect on both human and system classification accuracy, especially in terms of precision. Although System₄ still assigns too much probability to *neutral*, other classes do clearly benefit from the folding. Most remarkably, the f-score of *negative* has nearly reached 60%, which is remarkably higher than the sum of the *negative, implicit* and *negative, explicit* f-scores in System₆ ($\Delta = 11.1\%$). The same holds for *positive* ($\Delta = 11.2\%$), although on a much lower level.

We hypothesize that additional improvements could be gained from including structural information of newspaper articles in the gold standard. As mentioned in earlier sections, such annotations were not available in our sample corpus. How-

⁴Despite preliminary experiments with a number of other learning algorithms using WEKA (Hall et al., 2010), we decided to opt for fast iteration cycles and hence relied on the NLTK framework (Loper and Bird, 2002), which allowed for rapid prototyping.

ever, sentimental ascriptions could be particularly relevant for an actor if they appear in titles, leads or image captions of a newspaper article.

Apart from that, a thorough assessment of more powerful learning algorithms is indispensable for future iterations on our prototype system. This should be accompanied by including additional carefully thought out features, such as polarity class ratios or, as outlined in the previous section, structural information. Also, we will have to separately evaluate our polarity composition component (as illustrated in section 3.1.1) in order to consider possible extensions.

As for polarity class granularity, it is remarkable that *controversial* cases were particularly hard to identify, even in the four class model. The *fög* coding manual says that an event is controversial if “the reputation object is discussed controversially; positive and negative ascriptions are equally balanced.”⁵ Our classifier had no means of assessing the balance of negative and positive ascriptions when trained on our feature set described in section 3.2—there were of course counts for POS and NEG zero-grams, but due to the *Naive Bayes* independence assumption, no positive-negative ratio could be obtained.

From a pragmatic point of view, one could argue that the system’s moderate accuracy could partially be outweighed by quantitative considerations. As mentioned in section 4.1, the pipeline processes an article in 5.3 seconds on average, and in the nature of things, classifier decisions are fully reproducible. Viewed in this light, it does not seem unreasonable to use automated systems for work that is rather tedious for human experts, such as pre-selecting texts or encoding articles where the polarity of an actor is perfectly obvious.

6 Conclusion

In our exploratory work, we sought to assess the feasibility of automating a fine-grained classification process for media reputation analysis. We have trained two prototype classifiers relying on a lightweight approach to lexical sentiment composition, which we evaluated on a set of 200 pre-

⁵“Das Reputationsobjekt wird kontrovers diskutiert; Positiv- und Negativzuschreibungen halten sich die Waage.” (*fög*, 2011)

viously annotated newspaper articles. It clearly turned out that our approach is not suitable for handling a six-class polarity model. However, we gained substantial improvement from folding *implicit* and *explicit* ascriptions into a four-class-model, through which we argue that automated approaches have a promising potential in qualitative media reputation analysis.

In future work, we will consider structural information of newspaper articles for classification, as well as thoroughly examine the impact of more sophisticated machine learning algorithms on classification accuracy. Currently, we are training and evaluating an additional classifier for a three-partite polarity model, as a distinction between *positive*, *neutral* and *negative* is still most important in high-level aggregations such as the *fög* Media Reputation Index (Eisenegger and Imhof, 2008). We consider our present work as a motivating first step towards automating qualitative reputation analysis processes, calling for further collaboration in the intersection between sentiment analysis- and media reputation research.

Acknowledgments

We would like to thank Simon Clematide and Michael Amsler for motivating the use of VISL in our feature extraction pipeline. Also, we are grateful to Corina Tobler for translating the examples provided in section 3.

References

- Chen, Lu, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. 2012. Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM) 2012*, pages 50–57, Dublin.
- Choi, Yejin, and Claire Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2008*, pages 793–801, Honolulu.
- Clematide, Simon, and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA) 2010*, pages 7–13, Lisbon.

- Codebuch Akteursanalyse.* rev. 2011. Center for Research on the Public Sphere and Society (*fög*), University of Zurich. Internal coding manual for media reputation analysis.
- Deephouse, David L. 2000. Media reputation as a strategic resource: An integration of mass communication and resource-based theories. *Journal of Management*, 26(6):1091–1112.
- Eisenegger, Mark, and Kurt Imhof. 2008. The True, The Good and the Beautiful: Reputation Management in the Media Society. In Ansagr Zerfass, Betteke van Ruler, and Sriramesh Krishnamurthy, editors, *Public Relations Research: European and International Perspectives and Innovation*. VS Verlag für Sozialwissenschaften, Wiesbaden, pages 125–146.
- Eisenegger, Mark, Mario Schranz, and Jörg Schneider. 2010. Corporate Reputation and the News Media in Switzerland. In: Craig E. Carroll, editor, *Corporate reputation and the news media. Agenda-setting within business news coverage in developed, emerging, and frontier markets*. Routledge (Communication series), New York, S. 207–220.
- Godbole, Namrata, Manjunath Srinivasiah, and Steven Skiena. 2007. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the First International AAAI Conference on Weblogs and Social Media (ICWSM) 2007*, Boulder.
- Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.
- Hu, Minqing, and Bing Liu. 2004. Mining Opinion Features in Customer Reviews. *Proceedings of the 19th national conference on Artifical intelligence (AAAI) 2004*, pages 755–760, San Jose.
- Klenner, Manfred, Angela Fahrni, and Stefanos Pettrakis. 2009. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of NODALIDA 2009*, pages 235–238, Odense.
- Loper, Edward, and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia.
- Moilanen, Karo, and Stephen Pulman. 2007. Sentiment Composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP) 2007*, pages 378–382, Borovets.
- Polanyi, Livia, and Annie Zaenen. 2004. Contextual valence shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*.
- Springer Verlag (The Information Retrieval Series), Dordrecht, pages 1–9.
- Sennrich, Rico, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of the Biennial GSCL Conference 2009*, pages 115–124, Tübingen.
- VISL-group. 2008. http://beta.visl.sdu.dk/constraint_grammar.html. Institute of Language and Communication (ISK), University of Southern Denmark.
- Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) 2005*, pages 347–354, Vancouver.

Opinion analysis: the effect of negation on polarity and intensity

Lei Zhang, Stéphane Ferrari and Patrice Enjalbert

GREYC lab - CNRS UMR 6072

University of Caen Basse-Normandie

FirstName.LastName@unicaen.fr

Abstract

This paper presents an ongoing work addressing the problem of opinion analysis. It takes part into a collaborative project with industrial partners, aiming at providing a professional with a help for strategic and technical intelligence. Thus, we focus on local semantic analysis rather than text or sentence classification. The purpose of our task-oriented approach is to characterize the properties of opinion statements in an applicative corpus. Inspired by linguistic models, the method we propose is a compositional one, consisting in detecting and analyzing valence shifters such as negation which contribute to the interpretation of the polarity and the intensity of opinion expressions. We describe our model and its first implementation before discussing the results of a proof-of-concept experiment focusing on adjectival expressions.

1 Introduction

Opinion mining and sentiment analysis have been fields of high interest for the NLP community in recent years. There are many different applications in view, such as strategic intelligence, reputation management as well as automatic gathering of customers opinions and expectations. Our work aims at providing a front-end user, a professional in strategic and technical intelligence, with a help for spotting, analyzing and comparing opinion statements in her corpus. This project, *OntOpiTex*, is a collaboration between institutional research labs and industrial partners.¹

¹The *OntOpiTex* project is labelled by the ANR (French Research Agency), ANR-09-CORD-016.

A great part of related works focus on text or sentence classification, proposing methods to spot subjective discourse, and differentiating between positive and negative opinions. Other works studied the constitution of corpus-dependent subjective lexicons. These now well established approaches could have direct applications, but they still need to be completed in order to help a professional user analyze a specific domain. Therefore, the purpose of our project consists in providing a semantic analysis of opinion statements, describing their characteristic properties. In this paper, we focus on two of them, polarity and intensity. We propose a model inspired by linguistic approaches, and present a related experiment.

In the next section, we briefly introduce the generic scope of the *OntOpiTex* project, describing the tasks in view as well as the applicative corpus. In section 2, we present the related works, both in linguistics and in NLP area, paying special attention to the Appraisal theory (Martin and White, 2005), which greatly inspired our approach. In section 4, we describe our method for local semantic analysis of opinion statements and the current implementation of our model. While polarity characteristic is viewed in relation with the ‘Graduation’ property, the intensity of sentiment statements, is described in terms of ‘Force’ and ‘Focus’. The approach is a compositional one, taking into account different kinds of modifiers and negation. Section 5 details a proof of concept experiment restricted to the class of the adjectives. We also present and discuss the results obtained on the applicative corpus and on a similar corpus. In conclusion, we propose a few perspectives, mainly with regard to the integration

of our work in the whole application in view.

2 Scope of the project, task in view

The *OntOpiTex* project is an interdisciplinary one, involving computer scientists and linguists from three different laboratories as well as industrial partners, *Noopsis* and *TecKnowMetrix*. Its purpose consists in designing back-end and front-end tools for a professional use in technical strategic intelligence tasks. The whole architecture can be decomposed in two main parts: a set of tools and resources for the automatic analysis and different graphical user interfaces composing a dashboard for the front-end part. In a broad outline, researchers are mainly involved in the linguistic, NLP models and tools, *Noopsis* partner develops the generic architecture and the front-end application while *TecKnowMetrix* provides the applicative frame and the use case for an extrinsic evaluation.

The line of business of *TecKnowMetrix* concerns competitive intelligence in the technology domain. Their experts have to analyze two kinds of texts: patents for dealing with the legal concepts related to intellectual property, and technical journals, in order to analyze furthermore the activity in a specific domain. The *OntOpiTex* project is limited to the latter, where opinion or evaluation statements may have a better chance to occur. The use case in view is the analysis of the competition in the domain of a client, for comparison or position purposes. The applicative corpus is related to the avionic technologies, w.r.t. Boeing and EADS/Airbus companies. It consists in 377 journalistic texts from economics and technical press in French language, representing around 340 000 words.

As expected, opinion expression is not the main characteristic of such a corpus. However, in our task-oriented application, the professional user knows which targets are of any interest w.r.t. her task. The axiological or evaluative dimension of a statement is mostly a user-centered notion, related to the interpretation process. This is one reason why the resources used for our experiments are currently built to satisfy this constraint, merging generic lexicons (e.g. denoting affect) with very specific ones (e.g. describing technical properties of airplanes). Therefore, local semantic analysis of evaluative statements is proposed to com-

plete any statistical analysis that could be realized on positive and negative tendencies in the corpus. Furthermore, comparing the companies products and activities leads to pay special attention to the intensity of evaluation. For this purpose, we focus on the role of negation and intensity modifiers to detect polarity and intensity variations.

3 Related work

3.1 Opinion mining and sentiment analysis in NLP

Studies in opinion mining and sentiment analysis may be roughly classified in one of the three following tasks: (i) lexicon building, (ii) text or sentence classification and (iii) opinion statements analysis.

(i) The study of Hatzivassiloglou and McKeown (1997) is one of the early works aiming at lexicon building. The authors present a way to determine the orientation of the conjoined adjectives based on constraints on conjunctions. Different approaches use *seeds lists* to initiate a lexicon extraction from corpus, as in (Turney and Littman, 2003). Esuli and Sebastiani (2006) exploited WORDNET to develop the SENTIWORDNET extension, associating two properties (namely subjectivity and semantic polarity) to *Synsets*.

(ii) Text and sentence classification is generally viewed as a binary task: objective *vs* subjective or positive *vs* negative. Most studies are based on data mining and machine learning techniques. Many text genres have been studied. Movie review may be the earliest and most common one (Turney, 2002; Pang and Lee, 2004). Recently, the studies massively focus on customer reviews, weblogs and twitters (Breen, 2012; Singh et al., 2012). Recent approaches such as (Lambov et al., 2010) also propose models for reducing the domain dependence of subjective texts classifiers.

(iii) The last task, opinion statements analysis, consists in determining complementary features. Hu and Liu (2004) study the consumer reviews of product's technical features, pointing the targets of opinion statements as well as their polarities. Our study is included in this third task, based on the *Appraisal* theory (described in 3.3).

3.2 Main approach in linguistics

The French linguistics is currently missing a unified theory for the notion of evaluation. To synthesize, this notion has been considered in studies on subjectivity according to three main points: (i) the study of some modal values (appreciation, evaluation,...); (ii) the analysis and compilation of subjective lexicons and (iii) the study of some related notions in linguistics such as the point of view, the engagement and the endorsement. Different systems of modal values have been proposed by the linguists (Charaudeau (1992), Culoli (1980), Kerbrat-Orecchioni (1999)) with more or less emphasis on those coming under the subjectivity of the speaker (potentially the opinion holder).

The situation in English linguistic is different: opinion has been studied in a more systematic manner. For instance the collective work (Hunston and Francis, 2000) is largely oriented on text analysis. More precise studies (Hunston and Sinclair, 2000) focus on the notion of local grammars of evaluation. Three books propose a complete model of evaluation. The first, Evaluative semantics (Malrieu, 2002) present a model based on a cognitive model of evaluation. The second, Appraisal in media discourse (Bednarek, 2006) present a model based on two main categories: factors and values. The factors are the different appraisal fields (comprehensibility, emotivity, expectedness, importance...). This model is applied in order to classify the evaluations specific of the trustworthy newspapers from the ones specific of the tabloids. The third book, Appraisal in English (Martin and White, 2005) proposes a theory detailed more precisely in next section 3.3.

3.3 Appraisal theory

Appraisal theory is a relatively recent linguistic theory, elaborated for English language.² It proposes a complex system for describing the properties of opinion statements according three main aspects: Attitude, Engagement and Graduation (see figure 1).

Attitude itself divides into three sub-systems: Affect, Judgment (human behavior) and Appreciation (objects and products). Attitudinal mean-

ing can be clearly conveyed by individual words – for example, ‘angry’, ‘brave’ and ‘beautiful’. But most of the times, it is not individual words but words combinations which convey Attitude – for example, ‘his election [is] an affront to the democratic principle’. Therefore, though individual words may be ‘attitudinal’, it is better to see Attitude as a feature or property of complete statements and of stretches of language which present a complete proposition or proposal. In the Appraisal theory, Polarity is part of the Attitude system.

Engagement concerns the intersubjective dimension, linked to linguistic marks which explicitly position texts proposals and propositions inter-subjectively. For example: modals of probability – ‘perhaps’, ‘I think...’, ‘surely’; expectation – ‘predictably’, ‘of course’, etc.

Graduation involves two dimensions : ‘Force’ (variable scaling of intensity) and ‘Focus’ (sharpening or blurring of category boundaries). It can apply to both Attitude (graduating the opinion itself) and Engagement (graduating the endorsement). For example: force – ‘very’, ‘completely’, ‘rather’; focus – ‘a kind of’, ‘effectively’, ‘a true friend’.

In (Whitelaw et al., 2005), Appraisal groups used for sentiment analysis are described in terms of Attitude, Polarity, Force and Focus. The authors use the example of *not very happy* to illustrate the role of negation and intensity modifiers in such groups. This example is discussed further in the next section.

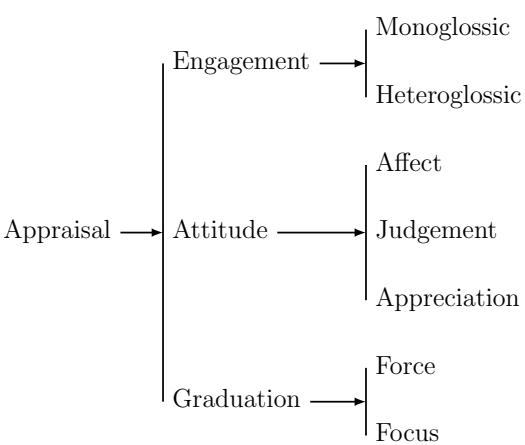


Figure 1: the Appraisal framework (simplified)

²See <http://www.grammatica.com/appraisal/> for an outline.

3.4 Negation and graduation

Valence shifters play a crucial role in sentiment analysis. Among different aspects of valence shifters, three types are frequently considered: negation, intensifiers and diminishers.³

Negatives are the most obvious shifters that affect polarity and force/focus. How ‘not’ can flip the valence of a term has been discussed in several works: (Pang et al., 2002), (Hu and Liu, 2004), etc. Wiegand et al. (2010) propose an interesting survey on the role of negation in sentiment analysis. According to the authors, negation is highly relevant for sentiment analysis. However, negation is a complicated phenomenon, and despite the existence of several approaches to negation modeling, current models are still incomplete (Wilson et al., 2005).

Intensifiers and diminishers are also important valence shifters. They can belong to all open lexical classes. The calculation of polarity modified by negatives or intensifiers/diminishers has often been done separately. Whitelaw et al. (2005) chose to reverse the force and the polarity in the context of conjoined negation and intensifier: ‘very happy’ (polarity:positive, force:high) → ‘not very happy’ (polarity:negative, force:low). In our opinion, for French language at least, this result depends the force level (low, high, extreme), as presented in the following section 4.

4 Model for graduation analysis

Under Graduation, our concerns cover two dimensions: Force and Focus.

Force Force has to do with the intensity of a word or expression – *assez* “rather”, *très* “very”, etc. It includes quantity and proximity modifiers – *un peu* “a few”, *beaucoup* “many”; *quasiment* “almost”, *presque* “nearly”, etc. It should also be noted that this principle of force grading operates intrinsically across values of attitude in the sense that each particular attitudinal meaning represents a particular point along the scale of low to high intensity. For example, some adjectives represent the highest scaling such as *extraordinaire* “extraordinary”, *magnifique* “brilliant”, *énorme* “huge”, etc. In addition, some prefixes can be interpreted as ‘very’: super-, hyper-, extra-, etc.

³In our work, Graduation covers diminishers and intensifiers.

In our system, we consider five different values of Force: **low** – *un peu* “a little”, **moderate** – *moyennement* “fairly”, **standard**, **high** – *très* “very” and **extreme** – *extrêmement* “extremely”. An adjective without intrinsic intensity label is assigned the ‘standard’ value by default.

Figure 2 describes how negation acts on force and polarity of a word or expression. Our model of negation is inspired by French linguists (Charaudeau (1992), Muller (1991), etc.). We chose to present this model with the opposite pair *bon* “good” ↔ *mauvais* “bad”, because the antonymy relation characterizes most gradable words, placing them on a scale of values from a negative extremity to a positive one.



Figure 2: Our model of negation and graduation

The rules for negation and positive polarity are the following:

- when acting on ‘extreme’ values, force is lowered and polarity is preserved (not extremely good ≡ a little good);
- when acting on ‘high’ values, force is lowered and polarity is reversed (due to euphemism in discourse – not very good ≡ a little bad);
- when acting on ‘moderate’ or ‘low’ values, force is raised and polarity is preserved (not a little/fairly good ≡ very good);
- when acting on ‘standard’ values, polarity is reversed but force stays uncertain.

The first rule concerning ‘extreme’ values was established through an experiment.⁴ two experts

⁴Detailed in (Enjalbert et al., 2012).

annotated 125 sentences containing axiological adjectives in a negative context. The disagreement (4 %) over the polarity is mainly due to some adjectives with extreme value in negative context such as “n'est pas catastrophique” *is not catastrophic* (probably in relation to the effect of rhetoric).

These rules also apply on negative polarity with one exception: Muller (1991) points out that when negation acts on ‘standard’ negative values, the degree of value can be anything but ‘standard’ in the positive pole. More precisely, *pas mauvais* “not bad” means whether *moyennement bon* “fairly good”) or *très bon* “very good” but not just *bon* “good”.

For applicative purposes, all values have to be instantiated in the implementation. Therefore, the last rules are not exactly respected, a ‘standard’ value being currently attributed to uncertain cases (in these particular cases, standard is viewed as the mean of the possible values): not good \equiv bad, not bad \equiv good.

Focus In the Appraisal theory, Focus is used to intensify not gradable categories: modifiers such as ‘true’, ‘pure’ or ‘sort of’ sharpen or soften the belonging to a category, therefore intensifying the associated axiological value, if any.

In our model, we also consider focus modifiers with the two possible values ‘sharpening’ or ‘softening’. The following rules are used in order to combine them with the other modifiers:

- when acting with a non-standard Force, a sharpening focus modifier is interpreted as a force intensifier, and a softening focus modifier as a force diminisher, the previous rules for Force are applied;
- when acting with a standard Force, Focus is reversed by negation: SHARPEN (central) \leftrightarrow SOFTEN (peripheral).

For example, in *les résultats ne sont pas vraiment très bons* “the results are not really very good”, ‘vraiment’ is computed as a force modifier: ‘bon’: (force:standard..polarity:pos)
 \rightarrow ‘très bon’: (force:high..polarity:pos)
 \rightarrow ‘vraiment très bon’: (force:extreme..polarity:pos)
 \rightarrow ‘pas vraiment très bon’: (force:low..polarity:pos)
 \equiv “a little positive”

In *les résultats ne sont pas vraiment bons* “the results are not really good”, the same word is computed as a focus modifier:

‘vraiment bon’: (focus:sharpen..polarity:pos)
 \rightarrow ‘pas vraiment bon’: (focus:soften..polarity:pos)
 \equiv “hardly positive”

In *les résultats ne sont vraiment pas bons* “the results are really not good”, the order of the modifiers changes the final interpretation:

‘pas bon’: (focus:standard..polarity:neg)
 \rightarrow ‘vraiment pas bon’: (focus:sharpen..polarity:neg)
 \equiv “truly negative”

5 Implementation and first results

5.1 Implementation: broad outline through an example

The previous model is currently implemented in an IDE, using *Noopsis* plugin to the Eclipse environment. Different built-in modules help preparing texts to analyses: extraction of textual parts from the XML source texts (XSLT module), tokenizers for words and sentences (mostly regular expressions), POS tagging. Different lexicons can be designed via this environment (XML format, associating feature sets to forms or lemma entries, including multi-words expressions) and projected on the texts. A chunking module has also been developed for the purpose of the project, producing simple chunks as well as ‘groups’ of chunks (ideally syntactic groups). The entries for our core analyzer can therefore be described as hierachic structured texts (texts \supset paragraphs \supset sentences \supset groups \supset chunks \supset words), with a feature set associated to each unit.

The core analysis, mainly implemented in prolog, consists in the three following steps:

1. projecting resources on the texts, both lexicons for opinion words or expressions and lexicons for negation and graduation modifiers;
2. filtering opinion sentences (presence/absence of an opinion word);
3. analyzing filtered sentences.

The latter step produces a feature set for each opinion word (i.e. issued from the corresponding resource), activating the opinion analyzer. If word is ambiguous, e.g. potentially both modifier and axiological, this activation depends on the

local context. For instance, in *une belle réussite* “a great success”, though potentially axiological, the adjective ‘belle’ is here considered as a modifier, intensifying the noun ‘réussite’, no opinion analysis will be activated for the adjective itself.

The opinion analyzer is a set of (prolog) rules, adapted to the POS tag of the initiator. When activated, it creates a new feature set added at the sentence level. Three main features are created: **init**, describing the word or expression which activated the analysis, **tgt**, describing the part of the sentence which contains information related to the target of the opinion, and **graduation**, indicating the polarity, force and focus. Lets consider the following example:

« Trouver les ressources nécessaires pour l'A320 NEO n'a pas vraiment été chose facile », a déclaré Tom Enders dans un communiqué.

“Finding the necessary funds for A320 NEO was not really an easy thing”, stated Tom Enders in a press release.

Figure 3 and following show different partial views of the feature sets created. The output is serialized in an XML format where features are represented by XML elements and features values are their textual content.

```
-<annotation type="sentence">
- <text>
  « Trouver les ressources nécessaires pour l'A320 NEO n'a pas vraiment été chose facile », a déclaré Tom Enders dans un communiqué.
</text>
- <features>
  <subjective>true</subjective>
  - <opil>
    + <init></init>
    + <tgt></tgt>
    - <grad>
      <polarity>positive</polarity>
      <force>STD</force>
      <focus>SOFTEN</focus>
      + <from></from>
    </grad>
  </opil>
</features>
</annotation>
```

Figure 3: Example of feature set

The **init** feature (*initiator*) contains information about the lexicon entry. Here, the adjective *facile* “easy” initiated the analysis. The lexicon

gives information only about its polarity (‘positive’) and force (‘standard’) when used in an axiological way. For this specific example, the initiator is not known as a potential modifier, so the opinion analyzer is automatically activated.

The **tgt** feature is an intermediary result, spotting the words related to the opinion target, to be combined with a domain ontology and an anaphora resolver (out of the scope of the present paper). It also indicates which rule has been applied for analyzing the initiator. In figure 4, an adjective activated the *AdjInGN* rule, which looks for a name in the same chunk to spot the target, and which also explores the context in order to find potential negation and modifiers before the chunk.

```
-<tgt>
<rule>AdjInGN1</rule>
<id>chose</id>
<lemma>chose</lemma>
<sg>[ chose facile ]</sg>
</tgt>
```

Figure 4: Information about the target

Each rule has its specific exploration process, a systematic local exploration (inside the chunk), and a contextual exploration which depends on the rule. For adjectives, we currently use 4 specific rules and one default (X_{init} : element (here only adjectives) activating the analysis; X_{focus} , X_{force} , X_{neg} : focus, force or negation modifiers [$G_{context}$: statement explored outside the chunk boundaries]):

- **AdjInGN**: an epithet, inside a nominal phrase – the context is explored to identify a possible attributive verb before. For example: $[it \text{ is } [really]_{focus} \text{ not}_{neg}]_{context} \text{ an } [important]_{init} \text{ contract}[_{target}]_{NP}$.
- **GNGVAttrGAdj**: attribute of a subject, with an attributive verb already identified. For example: $[\text{The model}]_{target} \text{ is } [\text{particularly}]_{focus} \text{ innovative}[_{init}]_{AP}$.
- **GNGAdjAppo**: affixed adjective phrase – no context exploration.
- **GAdjAppoGN**: affixed adjective phrase before the qualified nominal phrase – no context exploration. For example: $[\text{Very force ergonomic}]_{init} \text{ and comfortable}[_{init}]_{AP}, [\text{the new model}]_{target} \dots$

- GAdj: default, no target, no context exploration.

For each rule, the embedding group is explored to build a list of close modifiers (including negation). The context exploration process may produce a second list of modifiers (including negation) found in the verbal group and appended to the previous one.

The **graduation** processed is based on the initiator description and the (possibly empty) list of modifiers. Figure 5 shows the result for the previous example.

```

<-<grad>
  <polarity>positive</polarity>
  <force>STD</force>
  <focus>SOFTEN</focus>
  -<from>
    -<mod_int>
      <lemma>pas</lemma>
      <dir>none</dir>
      <force>false</force>
      <focus>false</focus>
      <neg>true</neg>
    </mod_int>
    <polarity>positive</polarity>
    <force>STD</force>
    <focus>SHARPEN</focus>
    -<from>
      -<mod_int>
        <lemma>vraiment</lemma>
        <dir>SHARPEN</dir>
        <force>false</force>
        <focus>true</focus>
        <neg>false</neg>
      </mod_int>
      <polarity>positive</polarity>
      <force>STD</force>
      <focus>STD</focus>
    </from>
  </from>
</grad>

```

Figure 5: Graduation feature with embedded ‘from’ features

A **from** feature helps understanding the applied modifications (and checking the modifiers list): *pas* (not) applied to *vraiment* (really) applied to the initial adjective.

In the project, this feature is produced to be used by a following module dedicated to discourse level, in order to allow revisions. When two statements are incoherent, negation interpretation may be revised (out of the scope of the present paper). Indeed, the modifiers are combined as proposed in our model, but we had to choose which force value to associate with negation rather than presenting the final user with a

possible choice between multiple values (the two cases of negation acting on standard values, with positive or negative polarity).

Next section presents an experiment applying this implementation for adjectives only.

5.2 First results

The applicative corpus is constituted of 377 texts, mainly economics articles from the French newspaper *Les Échos*. Its size corresponds to the mean size of corpus *TecKnowMetrix* currently deals with to handle other real cases. Texts are all about at least one of the two avionic companies Boeing and EADS/Airbus.

As already observed in many previous works, adjectives are the most frequent forms used for producing opinion statements. Due to the small size of our applicative corpus, this first experiment has been limited to this category. The lexicon was built by experts observing the occurrences of adjectives used in the corpus (Enjalbert et al., 2012). It consists in 283 adjectives, some related to the generic categories of Attitude proposed in Appraisal, some specific to the application domain.

2323 opinion statements have been processed on the whole corpus. Only 1 feature set was created in 1755 sentences, 2 feature sets in 225, 3 feature sets in 34 and 4 feature sets in 4. In other words, 87% of the subjective sentences contains one opinion statement only (involving an opinion adjective). Table 1 shows raw results: for 2323

<i>type of unit</i>	# occ.
analyzed sentences	2018
created feature sets	2323
feature sets with modifier	365 (15.7%)
negation in feature set	85 (3.7%)
feature sets with >1 modifier	6 (0.3%)

Table 1: Raw results

analysis, a modifier list was built 365 times, 85 times involving a negation. Only 6 occurrences of more than one modifier were found. The example described in the previous section is one of them, showing a combination between negation and a focus modifier. The other ones are combinations between a negation and a force modifier, like the following:

« *Il y a encore moins d'un an, Airbus*

n'était pas très favorable au GTF, reconnaît David Hess.»

“Less than one year ago, Airbus was not very disposed to GTF, admits David Hess.”

Our tool produced the following feature set for the previous: (polarity:negative..force:LOW..focus:STD), taking into account *pas*, *très* acting on *favorable*. Due to the size of this applicative corpus, statistics are not relevant.

We built a similar corpus from the French newspaper ‘Le Monde’, extracting all articles (6227 articles) about EADS/Airbus or Boeing from years 1987 to 2006 in order to evaluate our approach.

Recall could not be evaluated, because of the rarity of the studied phenomenon: 6 examples found in a 340 000 words corpus. It must also be noted that the lexicon used in our experiment is not exactly designed for the second corpus. Articles in specialized press are written for professional readers, while a daily paper like ‘Le Monde’ addresses a larger readership. Different opinion adjectives should also be added before evaluating recall in this context.

In order to evaluate precision, we make the assumption that the processing of intensity and polarity does not depend on the context. We observed the opinion statements involving the same adjectives in the new context, focusing on the most elaborated expressions. In a sample of examples combining negation and modifiers (65 segments: sentences or paragraphs), 83.9% are correctly analyzed, w.r.t our model. The main errors are due to tagging or chunking problems (6.5%), difficulty to take punctuation into account when in the context exploration process (4.8%), temporal aspects combining with axiological expression (3.2%) (for exemple, *jamais aussi actif* “never so active”), as well as insufficient resources (*rien de* “nothing”, *ni ... ni* “neither ... nor”) for the remaining.

The adjective roles (epithet, attribute, affixed) allowed rather simple and limited context exploration strategies. We currently are generalizing the model to nouns, verbs and adverbs categories, using the same rules for Graduation computing. Context exploration strategies are not as easy to

design as for adjectives. This work is realized in collaboration with the linguist partners who identified relevant patterns, which may be more specific of the genre of the applicative corpus.

6 Conclusion

In this paper, we addressed the problem of opinion analysis. We first described the application in view, a help for strategic and technical intelligence, and the related use-case corpus: articles related to EADS/Airbus and Boeing companies issued from specialized press. The constraints of our task led us to focus on local semantic analysis of opinion statements, with special attention on their polarity and intensity for comparison purposes.

Relying on the Appraisal theory (Martin and White, 2005) and other linguistic studies, we proposed a model for computing the values of Force and Focus, two Graduation characteristics reflecting intensity, w.r.t. negation and other modifiers. The implementation of this model is a specific module in the whole project application. We detailed its behavior regarding adjectives analysis. The results on a small applicative corpus are not relevant, because the most elaborated rules are barely activated: only 0.3% of the outputs combine at least two modifiers. An experiment on a similar corpus has therefore been realized in order to evaluate the accuracy of these rules. With a current precision of 83.9%, we consider the module not accurate enough for professional purposes.⁵

However, this first experiment allowed us to identify the remaining problems. Tagging errors and wrong contextual analysis are the most frequent errors encountered. Our further works will focus on the integration of our module in the whole application. In this scope, we ought to take advantage of other analysis, which may correct some of the current errors: domain ontology and terminology as well as domain specific patterns established by linguistic partners should improve the tagging and context exploration around opinion statements. We plan an extrinsic evaluation after integration, with returns from the final user on the accuracy of the whole application.

⁵If not totally bad, this precision means more than 1 error every 10 results.

References

- M. Bednarek. 2006. *Evaluation in media discourse: analysis of a newspaper corpus*. Continuum Intl Pub Group.
- J.O. Breen. 2012. Mining twitter for airline consumer sentiment. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, page 133.
- P. Charaudeau. 1992. *Grammaire du sens et de l'expression*. Hachette París.
- A. Culoli. 1980. Valeurs aspectuelles et opérations énonciatives: l'aoristique. *La notion d'aspect*, pages 181–193.
- P. Enjalbert, L. Zhang, and S. Ferrari. 2012. Opinion mining in an informative corpus: Building lexicons. In *PATHOS 2012*.
- A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC'06: Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.
- V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA. Association for Computational Linguistics.
- M.Q. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- S. Hunston and G. Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*, volume 4. John Benjamins Publishing Company.
- S. Hunston and J. Sinclair. 2000. A local grammar of evaluation. *Evaluation in Text: Authorial stance and the construction of discourse*, pages 74–101.
- C. Kerbrat-Orecchioni. 1999. *L'Énonciation. De la subjectivité dans le langage*. Armand Colin, Paris.
- D. Lambov, G. Dias, and J.V. Graça. 2010. Multi-view learning for text subjectivity classification. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 30–35.
- J.P. Malrieu. 2002. *Evaluative semantics: Cognition, language and ideology*, volume 3. Routledge.
- J.R. Martin and P.R.R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave.
- C. Muller. 1991. *La négation en français*. Librairie Droz.
- B. Pang and L. Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278.
- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP'02: conference on Empirical methods in natural language processing*, pages 79–86.
- V.K. Singh, M. Mukherjee, G.K. Mehta, S. Garg, and N. Tiwari. 2012. Opinion mining from weblogs and its relevance for socio-political research. *Advances in Computer Science and Information Technology. Computer Science and Engineering, Part II*.
- P.D. Turney and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.
- P. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL (ACL'02)*, pages 417–424. Philadelphia, Pennsylvania, USA, ACL.
- C. Whitelaw, N. Garg, and S. Argamon. 2005. Using appraisal groups for sentiment analysis. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631, New York, NY, USA. ACM.
- M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68. Association for Computational Linguistics.
- T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.

Domain-specific variation of sentiment expressions: a methodology of analysis for academic writing

Stefania Degaetano-Ortlieb

Saarland University

s.degaetano@mx.uni-saarland.de

Elke Teich

Saarland University

e.teich@mx.uni-saarland.de

Ekaterina Lapshinova-Koltunski

Saarland University

e.lapshinova@mx.uni-saarland.de

Abstract

In this paper, we present work in progress towards a methodology for the analysis of domain-specific sentiment. In our case, we consider highly specialized scientific disciplines at the boundaries of computer science and selected other disciplines (e.g., computational linguistics, bioinformatics). Our approach is corpus-based and comprises the detection, extraction and annotation of features related to sentiment expressions, focusing on opinion targets.

1 Introduction

While many studies have been dedicated to the exploration of sentiment expressions, there is no comprehensive or uniform method of analysis. Studies vary not only in terms of their methodological (text-based, corpus-based, computational) but also in their theoretical approaches, see e.g., appraisal (Martin and White, 2005; Bednarek, 2006; Hood, 2010), stance (Biber and Finegan, 1989; Hyland, 2005), evaluation (Hunston and Thompson, 2001; Hunston, 2011) and sentiment analysis (Wilson, 2008; Somasundaran, 2010; Taboada et al., 2011). This multifaceted picture is also due to the phenomenon itself, which can be realized in various linguistic ways, especially when considering different domains.

We introduce a methodology for a semi-automatic corpus-based analysis of features related to sentiment expressions in academic writing. The methodology comprises: (1) feature detection by a manual annotation of features related to sentiment expressions in a small corpus

of 100.000 tokens, (2) automatic feature extraction for comparative analyses of domain-specific variation in a corpus of 34 million tokens, and (3) automatic feature annotation to enrich the corpus.

2 Theoretical framework and data

Our methodology of analysis is based on the theoretical framework of Systemic Functional Linguistics (SFL) (Halliday, 2004). In SFL each element in a language is explained by reference to its function in the overall linguistic system and components of meaning are seen as functional components. There are three main functional components inherent to language, i.e. metafunctions: ideational (entities involved in a discourse), interpersonal (personal participation) and textual (structure of information). The core entities involved in sentiment expressions are writer and reader as well as target, which relates to any entity evaluated in a discourse. Sentiment expressions and the relation between a writer and a reader belong to the interpersonal metafunction.

With its register theory, SFL also accounts for domain-specific variation, assuming that meaning is realized in language by specific lexico-grammatical features, and different distributions of these features give rise to specific registers. As we are interested in domain-specific variation of sentiment expressions in registers emerged by register contact (e.g., bioinformatics emerged by contact between computer science and biology), we compare the distribution of interpersonal lexis-grammatical features within what we call *contact registers* (such as bioinformatics) and *seed registers* (such as computer science and bi-

core entities	examples of features	realization examples
writer	self-mention epistemic expressions	<i>I, we, our possible, may, suggest</i>
reader	reader pronouns directives	<i>you, your consider, see</i>
target	TARGET_BE_eval-expr TARGET_eval-V	<i>The claim is true. We see that A fails to be a BPP algorithm.</i>

Table 1: Interpersonal lexico-grammatical features

ology). These features can then be related back to the core entities inherent to sentiment expressions. In the case of academic writing, we have writer- and reader-oriented features, which relate to the writer’s presence and the reader engagement in a discourse (Hyland, 2005; Biber et al., 1999), as well as target-oriented features, which relate to the entity evaluated (see Table 1 for examples). So far, studies of sentiment expressions in academic writing have mainly focused on writer-oriented and reader-oriented features (also known as stance and engagement features, see Degaetano and Teich (2011), Hyland (2005) and McGrath and Kuteeva (2012)). In this paper, we focus on target-oriented lexico-grammatical features in academic writing.

ing). The corpus amounts to approx. 34 million tokens and is segmented into sentences, tokenized, lemmatized and part-of-speech tagged.

3 Methods

Our methodology for the analysis of sentiment expressions comprises a threefold semi-automatic process: detection, extraction, and annotation of lexico-grammatical features.

To detect features related to sentiment expressions, we look at a subcorpus of SciTex (approx. 100.000 tokens) and manually annotate interpersonal features related to writer, reader and target for each register. For this purpose, we use the UAM CorpusTool (O’Donnell, 2008), which allows to build an own annotation scheme and to adapt the scheme during annotation. Out of the annotation scheme, we generate a list of writer-, reader-, and target-oriented features.

The list of interpersonal features as well as insights gained from the manual annotation on different linguistic realizations of these features serve to create rules for the automatic extraction of features. For the extraction, we use the Corpus Query Processor (CQP), part of the IMS Corpus Workbench (CWB) (Evert, 2005; CWB, 2010), which allows feature extraction in terms of regular expressions over tokens and their linguistic annotations (e.g., part-of-speech). Especially for the extraction of target-oriented features, macros are built that cover several linguistic realizations of these features based on the insights of the manual annotation. As an example, Table 2 shows an extract of the macro for the *TARGET_BE_eval-expr* feature, which extracts realizations with and without a relative pronoun (compare the realization examples in Table 2). As not every adjective following the sequence *target+verb-BE* is evaluative, we use the lexical constraint *\$eval-adj* (see

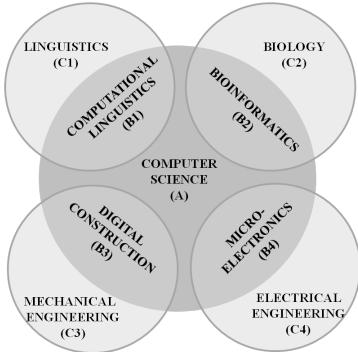


Figure 1: SciTex corpus

For our investigation we use the English Scientific Text corpus (SciTex; see Figure 1), specifically built to analyze register contact (Teich and Fankhauser, 2010; Degaetano-Ortlieb et al., 2012). SciTex contains a subcorpus for the contact registers (computational linguistics, bioinformatics, digital construction and microelectronics) and two subcorpora for the seed registers (one for computer science and one for linguistics, biology, mechanical engineering and electrical engineer-

	query building blocks	comments	realization
1	MACRO TARGET_BE_eval-expr(0) (begin macro and first query	
2	[pos="N.*"]	noun	<i>results</i>
3	[pos="MD"]?	optional modal verb	-
4	[lemma="be"]	verb BE	<i>are</i>
5	[pos="RB"]?	optional adverb	<i>quite</i>
6	[word=\$eval-adj]	evaluative adjective	<i>good</i>
7) (end first and begin second query	
8	[]	any token	<i>terminology</i>
9	[word=";"]?	optional comma	-
10	[pos="WDT"]	relative pronoun	<i>that</i>
11	[pos="MD"]?	optional modal verb	<i>will</i>
12	[lemma="be"]	verb BE	<i>be</i>
13	[pos="RB"]?	optional adverb	
14	[word=\$eval-adj] ...	evaluative adjective	<i>useful</i>

Table 2: Extract of the macro for the target-oriented feature *TARGET_BE_eval-expr*

lines 6 and 14 in Table 2), which restricts the extraction query to evaluative adjectives only, collected on the basis of the manual annotation. The automatic extraction is then performed on the full version of SciTex for comparative analyses of domain-specific variation.

Moreover, we want to automatically annotate the features back into the full version of SciTex to enrich the corpus with information at the interpersonal level. The annotation procedure is derived from the methods used in the YAC recursive chunker (Evert and Kermes, 2003) with CWB/CQP. The algorithm uses CWB perl-modules and perl-scripts to create additional annotation layers with CQP by writing back the query results of the extraction rules in form of new structural attributes into the corpus.

Based on the annotated features, we aim in the long-term at a corpus-based register analysis of interpersonal features. In the following section, we show selected analyses of target-oriented features.

4 Domain-specific variation - a sample analysis on target-specific tendencies

In order to analyze domain-specific variation of target-oriented features, we look at the differences in the targets evaluated across registers in SciTex as well as differences related to register contact by comparing contact registers with computer science or the other related seed register.

Considering the *TARGET_BE_eval-expr* feature, we observe domain-specific variation in

terms of targets evaluated across registers. Table 3 shows a triple comparison of bioinformatics with computer science and biology for the five most frequent targets. The targets seem to be mostly domain-specific, especially for biology. Bioinformatics seems to adopt targets from both seed registers, e.g. *gene* and *algorithm*. The same holds for the other contact registers, e.g. *algorithm* is in the first twenty targets for computational linguistics and under the five most frequent targets for digital construction and microelectronics.

register (size)	target	F	per 1M
computer science (2,612,258)	<i>algorithm</i>	79	30.24
	<i>problem</i>	49	18.76
	<i>result</i>	45	17.23
	<i>P</i>	41	15.70
	<i>lemma</i>	31	11.87
bioinformatics (1,425,237)	<i>method</i>	50	35.08
	<i>model</i>	37	25.96
	<i>gene</i>	29	20.35
	<i>algorithm</i>	24	16.84
	<i>approach</i>	23	16.14
biology (2,161,297)	<i>gene</i>	41	18.97
	<i>protein</i>	38	17.58
	<i>sequence</i>	37	17.12
	<i>region</i>	30	13.88
	<i>site</i>	26	12.03

Table 3: Targets of the *TARGET_BE_eval-expr* feature

When we think of *algorithm* as a domain-specific target of computer science which is adopted by the contact registers, the question arises whether *algorithm* is evaluated in the same way, i.e. are the sentiment expressions also adopted from computer science or is *algorithm*

evaluated differently in the contact registers?

frames	comp. sci.		contact reg.	
	F	%	F	%
accuracy	2	2.82	2	2.56
being_necessary	2	2.82	1	1.28
correctness	6	8.45	2	2.56
desirability	21	29.58	9	11.54
difficulty	15	21.13	14	17.95
importance	2	2.82	8	10.26
likelihood	4	5.63	4	5.13
obviousness	2	2.82	0	0.00
sufficiency	2	2.82	2	2.56
suitability	4	5.63	4	5.13
usefulness	10	14.08	23	29.49
success	0	0.00	4	5.13
fame	0	0.00	2	2.56

Table 4: *TARGET_BE_eval-expr*: eval. frames with *algorithm*

To answer this question, we extract the sentiment expressions used to evaluate *algorithm* with CQP and categorize them semantically to have a basis of comparison. The semantic categorization is done manually according to FrameNet (Ruppenhofer et al., 2010), which provides a dataset of semantic frames of lexical units. The sentiment expressions are inserted in the online FrameNet interface which outputs the respective frames for each expression. Ultimately, we aim an automatic categorization. Table 4 shows semantic frames used in the *TARGET_BE_eval-expr* feature in computer science and the contact registers. The percentages show that computer science uses *algorithm* more frequently with the frames desirability (e.g., *perfect*, *good*) and correctness (by *correct*), the contact registers instead more frequently with importance (e.g., *key element*, *important*) and usefulness (e.g., *helpful*, *useful*).

To test whether these differences are significant, we calculate the fisher's exact test, a univariate method for small raw frequencies, with the R environment (R Development Core Team, 2010) for the relevant frames (desirability, correctness, importance and usefulness). The p-value for this comparison is 0.00119, showing that computer science and the contact registers differ significantly in their use of these frames with *algorithm* in the *TARGET_BE_eval-expr* feature.

However, while the expression of correctness may be an evaluative act in some domains, to call an *algorithm* 'correct' in computer science

is rather a factual attribution. Thus, the concepts used to evaluate may vary according to register. We have to investigate further the semantic diversification related to scientific registers as well as the appropriateness of FrameNet frames for academic concepts.

Nevertheless, in summary we can say that for the *TARGET_BE_eval-expr* feature the contact registers adopt targets from the seed register computer science, but in the case of *algorithm* evaluate it differently.

5 Summary and envoi

In this paper, we have introduced a methodology to analyze sentiment expressions in academic writing. Our focus of analysis has been on specific lexico-grammatical features used to evaluate targets. As our overarching goal is to analyze registers emerged by register contact, we have (1) analyzed whether opinion targets are adopted from the seed register by the contact register, and (2) whether these targets are evaluated in the same way by both seed and contact registers. The analysis shows that there are domain-specific targets adopted by the contact registers. However, the evaluation of the targets can differ, as in the case of *algorithm*. Thus, we were able to detect domain-specific variation in terms of sentiment expressions, i.e. for interpersonal lexico-grammatical features.

In terms of methods, we have applied a corpus-based approach that allows to detect, extract and annotate features related to sentiment expressions semi-automatically in academic writing.

In the future, we will widen the range of target-oriented features moving towards a more comprehensive picture of domain-specific variation in terms of targets. We also have to take into account the semantic diversification related to different registers and investigate further the appropriateness of FrameNet categories to describe concepts used within scientific writing. As our model of analysis accounts also for writer- and reader-oriented features, we investigate these features as well. Doing so, we aim at analyzing domain-specific variation in terms of the strength of the writer's presence in specific domains and the reader's engagement across registers.

References

- Monica Bednarek. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. Continuum.
- Douglas Biber and Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language*, 65(3):487–517.
- Douglas Biber, Stig Johansson, and Geoffrey Leech. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.
2010. The IMS Open Corpus Workbench. <http://www.cwb.sourceforge.net>.
- Stefania Degaetano and Elke Teich. 2011. The lexico-grammar of stance: an exploratory analysis of scientific texts. In Stefanie Dipper and Heike Zinsmeister, editors, *Bochumer Linguistische Arbeitsberichte 3 - Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*. Bochum.
- Stefania Degaetano-Ortlieb, Kermes Hannah, Ekaterina Lapshinova-Koltunski, and Teich Elke. 2012. SciTex A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, Vol. 3. Narr.
- Stefan Evert and Hannah Kermes. 2003. Annotation, storage, and retrieval of mildly recursive structures. In *Annotation, storage, and retrieval of mildly recursive structures*, Lancaster, UK.
- Stefan Evert, 2005. *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.
- M.A.K. Halliday. 2004. *An Introduction to Functional Grammar*. Arnold, London.
- Susan Hood. 2010. *Appraising Research: Evaluation in Academic Writing*. Palgrave Macmillan.
- Susan Hunston and Geoff Thompson. 2001. *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford University Press, Oxford.
- Susan Hunston. 2011. *Corpus approaches to evaluation: phraseology and evaluative language*. Taylor & Francis.
- Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192.
- Jim R. Martin and Peter R.R. White. 2005. *The Language of Evaluation, Appraisal in English*. Palgrave Macmillan, London & New York.
- Lisa McGrath and Maria Kuteeva. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, 31:161–173.
- Michael O'Donnell. 2008. The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA*, Almeria, Spain, 3-5 April.
- R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petrucci, Christopher R. Johnson, and Jan Schefczyk. 2010. FramaNet II: Extended theory and practice. Technical report, ICSI.
- Swapna Somasundaran. 2010. *Discourse-level relations for opinion analysis*. Ph.D. thesis, University of Pittsburgh.
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.
- Elke Teich and Peter Fankhauser. 2010. Exploring a corpus of scientific texts using data mining. In S. Gries, S. Wulff, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*, pages 233–247. Rodopi, Amsterdam and New York.
- Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh.

Creating Annotated Resources for Polarity Classification in Czech

Kateřina Veselovská, Jan Hajíč, Jr. and Jana Šindlerová

Institute of Formal and Applied Linguistics

Charles University

Prague, Czech Republic

{veselovska|hajicj|sindlerova}@ufal.mff.cuni.cz

Abstract

This paper presents the first steps towards reliable polarity classification based on Czech data. We describe a method for annotating Czech evaluative structures and build a standard unigram-based Naive Bayes classifier on three different types of annotated texts. Furthermore, we analyze existing results for both manual and automatic annotation, some of which are promising and close to the state-of-the-art performance, see Cui (2006).

1 Introduction

One of the main subtasks in sentiment analysis is the polarity detection, which aims to classify documents according to the overall opinion they express, see e.g. Pang et al. (2002). Usually, the first step towards polarity detection is the generation of a subjectivity lexicon, i.e. a list of words marked for polarity, see e.g. Jijkoun and Hofmann (2009). Since up to now there was a lack of annotated resources for performing sentiment analysis tasks in Czech, our primary focus was to provide corpora that could serve as a basis for polarity detection systems, and to develop and test sentiment analysis systems on it. In the current study, we describe our annotation scheme, comment on its merits and pitfalls and their relation to the linguistic issues of sentiment analysis or the specific domain of the processed data. We also attempt to suggest possible remedies as a step towards providing a reliable annotation scheme for sentiment analysis in Czech.

2 Related Work

The very first stage of the project has been described in Veselovská (2012). Closely related work on sentence-level polarity classification is done by Wiegand and Klakow (2009), who also consider different linguistic features, e.g. part-of-speech information or negation scope, while building a classifier. Our annotation guidelines were inspired by the work of Wiebe (2002). Contrary to Wiebe, we do not take into account the type of attitude and the onlyfactive attribute yet. Some work on sentiment analysis in Czech has been done so far by Steinberger et al. (2011), who detect subjectivity in Czech news articles using parallel sentiment-annotated corpora. Although the authors only used one annotator for Czech, they noticed the same problems as our annotators did while annotating the news domain. Some of them, e.g. the discrepancy between the author's intention and reader's interpretation, was experienced also by Balahur and Steinberger (2009), who decided to redefine the task of sentiment analysis in the news significantly. They prepared new, more detailed guidelines which increased the final inter-annotator agreement considerably. However, our paper primarily aims at the comparison of the three data domains, and we decided not to explore news articles for the time being. There is a number of papers dealing with polarity detection in movie reviews, mostly using Naive Bayes classifier and the International Movie Database data set, with quite good results (see e.g. Pang et al. (2002)). Therefore, we decided to use Czech Movie Database and a similar method as a starting point.

3 Basic Methodology of Annotation

There are three levels at which polarity can be annotated: the expression level, the sentence (segment)¹ level and the document level. Of these, especially the first one has been widely explored, but there is also a significant number of papers dealing with the problem of sentence-level sentiment analysis, e.g. Meena and Prabhakar (2007). In our subjectivity annotation project, we have decided to start with the sentence level plain text annotation, with the long-term intention to implement the results (namely the subjectivity lexicon items derived from the annotation) in a richly annotated treebank. The sentence level annotation enables us to explore many useful linguistic features in the analysis, which can hardly be explored at the document level, such as part-of-speech information or features derived directly from the sentence structure, as described e.g. in Wiegand (2009). On the contrary, we still need to account for the fact that classifiers trained on a bag-of-words model usually perform worse at the sentence-level than at the document level, since the total number of words within a sentence is rather small and, as a result, feature vectors encoding sentences tend to be much sparser.

In the task of sentence-level polarity classification in Czech, we distinguish three functional evaluative components that need to be identified, following Wiebe (2004):

- the source, i.e. the person or entity that expresses or experiences the private state (the writer, someone quoted in the text etc.),
- the evaluation, expressed by polar elements, i.e. words or phrases inherently bearing a positive or negative value,
- the evaluated target.

In contrast to e.g. Wilson (2008), we restrict our analysis to evaluative opinions/states only. Moreover, we take into account and mark in the annotation some further aspects concerning a fine-grained subjectivity analysis, mainly

¹We use 'sentence' and 'segment' interchangeably in this article. Every sentence is a segment, but not every segment is a sentence as linguistics would have it, as there were items like news headlines or one-word exclamations in the data.

expressions bordering the area of sentiment analysis, such as good/bad news (see Section 4.2), or elusive elements (expressions bearing evaluative power, but such that we cannot describe them in terms of standard polarity values, e.g. "controversy").

The annotation practice is based on the manual tagging of appropriate text spans, and it is performed by two independent annotators (later referred to as A and B).

4 Data Sets

We have trained and tested our classifier on several data sets. The primary motivation for our research was to create a tool for detecting the way news articles influence public opinion. Therefore, we initially worked with the data from the news website Aktualne.cz. However, the analysis of such texts has proven to be a rather difficult task in terms of manual annotation, as well as automatic processing, because there was a strong tendency to avoid strongly evaluative expressions, or even any explicit evaluation. For this reason, we also decided to use review data from Czech movie database, CSFD.cz, since movie reviews have been successfully used in the area of sentiment analysis for many other languages, see e.g. Thet et al. (2009). As both sets of the manually annotated data were pretty small, we also used auxiliary data, namely domestic appliance reviews from the Mall.cz retail server.

4.1 Aktualne.cz

There are approximately 560,000 words in 1661 articles obtained from the Home section of the Czech news website Aktualne.cz. In the first phase, we manually categorized some of the articles according to their subjectivity. We identified 175 articles (89,932 words) bearing some subjective information, 188 articles (45,395 words) with no polarity, and we labelled 90 articles (77,918 words) as "undecided". There are 1,208 articles which have not been classified yet. Most of this data is not intended for manual processing but for various unsupervised machine learning methods in potential NLP applications.

The annotators annotated 410 segments of texts (6,868 words, 1,935 unique lemmas). These segments were gained from 12 randomly chosen

opinion articles from Aktualne.cz. The segments are mostly sentences, but they also contain headlines and subtitles. In the sequel, we refer to annotation items as segments.

At the beginning, we tried to annotate all polar states that elicit a reaction from the reader. The primary instruction for annotators was simple: Should you like or dislike an entity occurring in a segment because of what that segment says, tag the entity accordingly. This choice of annotator perspective was motivated by the desired application: if our goal is for the computer to simulate a reader and thus develop sympathies, then the training data should reflect this process. It would also enable us to bypass the issue of identifying sources and assigning some trust parameter to them. However, combined with the requirement of neutrality towards the protagonists, this choice of perspective did impede the annotators' ability to make judgements about the presence of polarity in segments. The inter-annotator agreement was a little over 0.63 by Cohen's Kappa for all polarity classes. The annotators tagged about 30% of all the segments in total.

4.1.1 Problems in Tagging

Concerning the target annotation, we experienced various problems which can be divided into several categories.

The easily resolvable ones were the problems concerning annotators' instructions and their interpretation, namely the (insufficient) clarity of the instructions (e.g. concerning the question whether the annotators should tag the preposition as a part of the target or not), misinterpretation of the annotator instructions, or misinterpretation of some linguistic properties of the text. Due to the generality of the given task the boundary between the latter two phenomena is not very clear. In general it appeared quite difficult for the annotators to abstain from their personal sympathy or antipathy for the given target, especially because the texts deal with the controversial political situation before Czech parliamentary elections in 2010.

One of the specific problems of our annotation was the fact that all of our annotators have had a linguistic background, so they might have tended to tag sentences with some presupposedly linguistically interesting polarity item, even though the

polarity lay in another expression or the sentence was not subjective at all. See (1):²

- (1) A. Vláda schválila něco jiného, než co sli-bovala.

- B. Vláda schválila **něco jiného**, než co sli-bovala.

The government approved [something else]_B than what it had promised.

Here the target of the negative evaluation is actually "the government".

Further problems were caused by a vague interpretation of targets in polar sentences: in evaluative structures, there are different levels on which we can determine the targets, see (3):

- (2) A. Dům *byl* před sedmi lety neúspěšně dražen, nyní je v zástavě banky.

- B. *Dům* byl před sedmi lety neúspěšně dražen, nyní je v zástavě banky.

Seven years ago, [the house]_B [was]_A unsuccessfully auctioned; now it has been pledged to a bank.

Annotator A apparently felt as negative the fact that the house had been offered in the auction, most likely because the auction was unsuccessful, whereas annotator B perceived the house itself as the evaluated entity because it failed in the auction. Here we prefer the second option, since with respect to the overall topic of the document in question, we suppose that the reader will probably evaluate the house rather than the auction.

The above problems can also be caused by seeing the subjectivity structure *source – evaluation – target* as parallel to the syntacto-semantic structure *agent – predicate – patient*. Although these structures may be parallel (and they very often are), it is not always the case.

We found many discrepancies between the local and global polarity – while a part of the sentence being evaluative, the whole sentence appears rather neutral (or even its overall polarity is oriented in the opposite direction), see (4):

²From here on, the tagged expression is indicated in bold if the identified polar state is oriented positively, or in italic if negative. A and B refers to the decisions of the annotators A and B. In the free translation, square brackets and lower indexing mark the annotator's decision.

(3) A. V případě jeho kandidatury na tento post by **jej** podporovalo pouze 13% dotázaných, a to z řad voličů ČSSD a KSČM.

B. V případě jeho kandidatury na tento post by **jej** podporovalo pouze 13% dotázaných, a to z řad voličů ČSSD a KSČM.

In case of his candidacy for this post, [he]AB would be supported only by 13% respondents, mostly supporters of ČSSD and KSČM.

4.1.2 Possible Annotation Scheme Improvements

In order to improve the annotation scheme, we found necessary to abandon the reader's perspective, and to annotate not only targets, but also sources and expressions. Originally, we hoped that taking the readers perspective could prove advantageous for the identification of those polar indicators which are most relevant for the readers. However, it turned out that it is hard to identify reader-oriented polarity (and its orientation) while keeping the sources and targets anonymous. Therefore we find more useful to separate the task of identifying subjective structures and the assignment of relevance to the reader.

Another option might be to abandon the requirement for neutrality and extend the number of annotators, ideally to a representative number of readership, e. g. by means of the so-called crowdsourcing (such as Amazon's Mechanical Turk, see also Akkaya et al. (2010), or similar large scale human annotation resource).

4.2 CSFD data

In the second phase of the project, we decided to use data more convenient for the task, namely the data from Czech movie database CSFD.cz. In comparison with the previous data set, the language of these reviews was significantly more evaluative, even though it was much more domain-dependent. To compare the results, we again chose 405 segments and let the same two people annotate them. In this case, the results were slightly better, with Cohen's Kappa 0.66. However, we again experienced some problems.

4.2.1 Problems in Tagging

Perhaps the most interesting and most disturbing issue we have encountered when annotating polarity is the annotator inconsistency and mutual disagreement in establishing the borderline between polarity target and polarity expression (evaluation). A substantial part of inter-annotator disagreement in target identification lies in different perception of the extent of polarity expression with respect to the entity evaluated. This happens especially in copular sentences, both attributive and classifying.

(4) Tom Hanks je výborný herec.

Tom Hanks is an excellent actor.

In such sentences, known also as qualification by non-genuine classification, see Mathesius (1975), annotators either tag "Tom Hanks" or "actor" or "Tom Hanks;actor"³ as targets of the polarity expression "excellent". The three alternative solutions show three different, but equally relevant ways of polarity perception. Pragmatically, the real-world entity evaluated is Tom Hanks. Syntactically, it is the headword "actor" that is modified by the qualifying adjective "wonderful". And semantically, it is the professional abilities of T. H. as an actor which are being evaluated.

(5) Kate Winslet je špatně oblečená.

Kate Winslet is poorly dressed.

As in the previous example, the target of the negative evaluation is actually both Kate Winslet and the way she dresses herself. At the beginning we have tried to capture this problem by means of copying, i.e. we kept two separate instances of a polar state, one with "Kate Winslet" as the target and "poorly dressed" as the evaluation, the other as "dressed" as the target and "poorly" as the evaluation. Doubling the polar information though did not appear to be advantageous with respect to annotators' time expenses, moreover the annotators did not succeed in capturing each single instance of the structure in question, therefore we withdrew from such treatment in favour of the more complex variant of keeping the entity as the target and the attributed quality/ability/profession etc. in the evaluation category.

³We use semicolon for appending discontinuous items.

Good News, Bad News During the annotation of news articles we felt the need for a separate category capturing the so-called “good” and “bad news”. It appeared useful to separate sentences involving events commonly and widely accepted as “pleasant” or “unpleasant”, such as triumph, wealth, or death, injury, disease, natural disaster, political failure etc., from individual subjective statements of sentiment. Interestingly it appeared quite difficult for the annotators to identify a clear-cut borderline between subjective positive/negative opinion and good/bad news, perhaps because of generally widespread metaphorical uses of the “(un)pleasant”. With movie reviews, the situation was easier. First, due to the maximally subjective character of the texts, good/bad news did not appear significantly often, were easily identifiable and did not intervene much into the annotators’ decision. Nevertheless, this type of disagreement did occur, e.g. in the sentence “Bůh je krutý. God is cruel.” or “Dialogue jsou nepřípadné. The dialogues are inappropriate.”

Non-trivial and Domain-specific Semantics

As expected, the inter-annotator agreement often fails in places where the subjectivity of the sentence is hidden and embedded in metaphorically complex expressions like

- (6) Všichni herci si zapomněli mimické svaly někde doma.

All the actors have forgotten their mimic muscles at home.

- (7) Slovo hrdina se pro něj opravdu nehodí.

The word “hero” does not really fit him.

Moreover, sometimes the annotated polar expression serves the polarity task only within the given semantic domain. Thus whereas expressions like “špatný herec; bad (actor)” or “špatně (oblečená); poorly (dressed)” can function universally across different text genres and topics, the expressions like “psychologicky propracované (postavy); psychologically round (characters)” or “jsou stříhnuty brzo; are edited too early” take the concrete polar value according to the presupposition whether we are dealing with a movie review or not. In a different text genre they could easily acquire a different polarity value, or even

they could serve as neutral, non-subjective element.

4.2.2 Enhancing the Annotation Scheme

During the annotation of CSFD data we have decided to make two improvements in the annotation scheme. First we added two more polarity values, namely NONPOS and NONNEG, for capturing more fine-grained evaluation of the type “not that good” or “not that bad” respectively.

- (8) Meryl není ani krásná ani výjimečná.

Meryl is neither beautiful, nor exceptional.

NONPOS

- (9) Ironický nadhled v první části vlastně nebyl tak zbytečný.

The ironic detached view in the first part wasn't actually that pointless.

NONNEG

These additional labels do not equal simple “bad” or “good” values, but neither do they refer to a neutral state. Essentially, they describe a situation where the source’s evaluation goes against a presupposed evaluation of the reader’s. By adding additional values we risk a slight rise in the number of points of annotator’s disagreement, on the other hand we are able to capture more evaluative structures and get a more thorough picture of the evaluative information in the text.

The second, rather technical improvement was the addition of a special label TOPIC for cases where the evaluation is aimed at the overall topic of the document and there is no other co-referential item in the context to which the target label could be anchored.

- (10) Skvěle obsazené, vtipné, brutální, zábavné, nápadité...

Excellently casted, witty, brutal, funny, imaginative...

As in the previous case, this label should help us capture more evaluative structures that would otherwise stay unidentified. We are aware of the fact that this label might be helpful only in domains with strong evaluative character (like product reviews), but maybe less useful in case of journalistic texts in general.

4.3 Auxiliary data – Mall.cz

We have obtained 10,177 domestic appliance reviews (158,955 words, 13,473 lemmas) from the Mall.cz retail server. These reviews are divided into positive (6,365) and negative (3,812) by their authors. We found this data much easier to work with, because they are primarily evaluative by their nature and contain no complicated syntactic or semantic structures. Unlike the data from Aktualne.cz, they also contain explicit polar expressions in a prototypical use. Furthermore, they do not need to be tagged for the gold-standard annotation.

The Mall.cz data, however, do present a different set of complications: grammatical mistakes or typing errors cause noise in the form of additional lemmas and some of the reviews are also categorized incorrectly; however, compared to the problems with news articles, these are only minor difficulties and can be easily solved.

We use the Mall.cz data to verify that the automatic annotation method presented in Sect. 5 below is sound, at least on a less demanding data set.

5 Automatic Classification Experiments

In our classification scenario, we attempt to classify individual units of annotations – *segments*. The aim of these experiments is not to build state-of-the-art sentiment analysis applications, but to evaluate whether the data coming from the annotations are actually useful, where are their limits and how to eventually change the annotation guidelines to provide higher-quality data.

5.1 Classifier description

In our experimentation, we use the Naive Bayes classifier. Naive Bayes is a discriminative model which makes strong independence assumptions about its features. These assumptions generally do not hold, so the probability estimates of Naive Bayes are often wrong, however, the *classifications* it outputs can be surprisingly good.

Let \mathcal{C} denote a set of polarity classes $C_1, C_2 \dots C_{|\mathcal{C}|}$. The classified unit is a segment, denoted s_j from a set of segments \mathcal{D} . A segment s_j is composed of n lemmas $s_{j,1}, s_{j,2} \dots s_{j,n}$. (Each lemma actually has three factors: the "real"

lemma itself, its Part of Speech and the Negation tag, see 5.2. However, for the purposes of the classifier, it is important to keep negation with the real lemma, as disposing of it would make e.g. *flattering* and *unflattering* indistinguishable.) The lexicon is then the set of all lemmas in \mathcal{D} and is denoted as \mathcal{L} . The size of the lexicon – that is, the number of distinct lemmas in the lexicon – is M . The classification features $F_i, i = 1 \dots M$ are then the presence of the i -th lemma l_i in the classified segment.

Given that the probability of classifying a segment as belonging to C is

$$p(C|F_1, F_2, \dots, F_M) \propto p(C)p(F_1, F_2, \dots, F_M|C) \quad (1)$$

by the Chain Rule ($p_{CP}(F_1, F_2, \dots, F_M|C) = p(C, F_1, F_2, \dots, F_M)$) and by assuming conditional independence of features $F_1 \dots F_M$ on each other it yields the following formula:

$$\begin{aligned} p(C|F_1, F_2, \dots, F_M) &\propto p(C) \prod_{i=1 \dots M} p(F_i|C) \\ &\propto \log p(C) + \sum_{i=1 \dots M} \log p(F_i|C) \end{aligned} \quad (2)$$

Maximization follows by simply *argmax*-ing over both sides. The model parameters – conditional probabilities of seeing the lemma l_i in each of the classes $C_1 \dots C_{|\mathcal{C}|}$ – are estimated as MLEs $p_T(F_i|C)$ on some training data $T = w_1 \dots w_{|T|}$ with Laplacian smoothing of strength α , computed as

$$p_T(F_i|C) = \frac{\text{freq}(i, C) + \alpha}{\text{freq}(C) + \alpha|\mathcal{C}|} \quad (3)$$

where $\text{freq}(i, C)$ is the number of times lemma l_i was seen in a segment s_j labeled as belonging to the class C .

A special *UNSEEN* lemma was also added to the model, with parameters $p(C|UNSEEN)$ estimated as the marginal probabilities $p(C)$ – the probability of something generating a polarity class should be the general probability of seeing that class anyway.

5.2 Experimental settings

The experiments were carried out across the three datasets described in 4: the small richly annotated Aktualne.cz and CSFD datasets and the larger Mall.cz data which are not annotated below segment level. Exploration of those richer features, however, has not been done extensively as of yet.

When merging the annotations, we used an “eager” approach: if one annotator has tagged a segment as polar and the other as neutral, we use the polar classification; NONPOS and NONNEG are considered NEG and POS, respectively, and segments classified as BOTH and NEG (or POS) stay as BOTH. Varying the merging procedure had practically no effect on the classification.

All data for the classifiers are tagged and lemmatized using the Morce tagger (Votrubec, 2005). We retain Part of Speech and Negation tags and discard the rest.

6 Evaluation and results

In order to judge annotation quality and usefulness, we use two distinct approaches: annotator agreement and classifier performance.

6.1 Annotator agreement

On segment level, we measured whether the annotators would agree on identifying polar segments (unlabeled agreement), polar segments and their orientation (unlabeled agreement) and whether they agree on orienting segments identified as polar (orientation agreement). Additionally, we measured text anchor overlap for sources, polar expressions and targets. We used Cohen’s kappa κ and f-scores on individual polarity classes (denoted f-ntr, f-plr, etc.) for agreement and f-score for text anchor overlap. Orientation was evaluated as BOTH when an annotator found both a positively and negatively oriented polar state in one segment.

For the Aktualne.cz data, out of 437 segments, the annotators tagged:

with the following agreement:

Text anchor overlap:

Annotator	1	2
Neutral	376	358
Polar	61	79
Negative	49	62
Positive	11	16
Both	1	1

Table 1: Annotator statistics on Aktualne.cz

Agreement	κ	f-ntr	f-plr	f-neg	f-pos	f-both
Unlabeled	0.659	0.944	0.714	-	-	-
Labeled	0.649	0.944	-	0.708	0.593	0
Orientation	0.818	-	-	0.975	0.889	0

Table 2: Agreement on Aktualne.cz data

Overlap	f-score
Source	0.484
Polar expr.	0.601
Target	0.562

Table 3: Overlap, Aktualne.cz

On CSFD data, out of 405 segments, the annotators identified (numbers – except for ‘neutral’ – are reported for polar states, thus adding up to more than 405):

Annotator	1 (JS)	2 (KV)
Neutral segs.	171	203
Polar states	348	281
Negative	150	132
Positive	180	135
Nonneg.	10	8
Nonpos.	8	6
Bad News	22	23
ESE	15	56
Elusive	2	22
False	0	10

Table 4: Annotator statistics on CSFD data

with the following agreement (reported for segments; ‘both’ are such segments which have been tagged with both a positive and a negative polar state):

Agreement	κ	f-ntr	f-plr	f-neg	f-pos	f-both
Unlabeled	0.659	0.809	0.850	-	-	-
Labeled	0.638	-	0.806	0.752	0.757	0.371
Orientation	0.702	-	-	0.873	0.876	0.425

Table 5: Agreement on CSFD data

Text anchor overlap:

Overlap	f-score
Source	0.750
Polar expr.	0.580
Target	0.706

Table 6: Overlap on CSFD data

The overlap scores for the Bad News, Elusive, False and ESE labels were extremely low; however, the annotators were each consistent in their reasoning behind applying these labels. A large majority of disagreement on these labels is from mislabeling. We therefore believe that agreement can be improved simply by pointing out such examples and repeating the annotation.

Aktualne.cz generally scores better on neutral segments and worse on polar ones. The similar κ would suggest that this can be put down to chance, though, because the higher prevalence of polar segments in the CSFD data makes it easier to randomly agree on them. However, the text anchor overlap shows that as far as expression-level identification goes, the annotators were much more certain on the CSFD data in what to “blame” for polarity in a given segment.

6.2 Classifier performance

The baseline for all classifier experiments was assigning the most frequent class to all segments. For all classifier experiments, we report f-score and improvement over baseline. The reported f-score is computed as an average over f-scores of individual classes weighed by their frequencies in the true data.

20-fold cross-validation was performed, with the train/test split close to 4:1. The split was done randomly, i.e. a segment had a 0.2 chance of being put into test data. No heldout data were necessary as the Laplace smoothing parameter α was set manually to 0.005; changing it didn't significantly alter results. All data were lemmatized by the Morče tagger (Votrubec, 2005).

On the Mall.cz data:

	Accuracy	f-score
Baseline	0.630	0.286
Naive Bayes	0.827	0.781

Table 7: Classifier performance on Mall.cz data

On Aktualne.cz data, the classifier was not able

to perform any different from the baseline. We hypothesised, however, that this may have been due to the massive imbalance of prior probabilities and ran the experiment again with only the first 100 neutral segments.

	Accuracy	f-score
Baseline	0.787	0.694
Naive Bayes	0.787	0.694
Baseline, 100 ntr.	0.304	0.142
NB, 100 ntr.	0.778	0.531

Table 8: Classifier performance on Aktualne.cz data

On CSFD data:

	Accuracy	f-score
Baseline	0.341	0.173
Naive Bayes	0.766	0.754

Table 9: Classifier performance on CSFD.cz data

7 Conclusion

Comparing the described attempts of annotating subjectivity, we must pinpoint one observation. The success in inter-annotator agreement is dependent on the annotated text type. Unlike newspaper articles, where opinions are presented as a superstructure over informative value, and personal likes and dislikes are restricted, CSFD reviews were written with the primary intention to express subjective opinions, likes, dislikes and evaluation. Both data sets will be available to the research community for comparison.

Possibly the most important finding of the classifier experiments is that the very simple Naive Bayes polarity classifier can be trained with decent performance (at least on the film review data) with only a very modest amount of annotated data.

The fact that annotator agreement exceeded $\kappa = 0.6$ can be, given the considerable subjectivity and difficulty of the task, considered a success.

Acknowledgments

The work on this project has been supported by the GAUK 3537/2011 grant. This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

- C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea. 2010. *Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation*, NAACL-HLT 2010 Workshop on Creating Speech and Lanugage Data With Amazon's Mechanical Turk.
- A. Balahur and R. Steinberger. 2009. *Rethinking Opinion Mining in Newspaper Articles: from Theory to Practice and Back*, Proceedings of the first workshop on Opinion Mining and Sentiment Analysis (WOMSA 2009).
- C. Banea, R. Mihalcea, and J. Wiebe. 2008. *A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources*, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008).
- H. Cui, V. Mittal, and M. Datar. 2006. *Comparative experiments on sentiment classification for online product reviews*, In proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence.
- V. Jijkoun and K. Hofmann. 2009. *Generating a Non-English Subjectivity Lexicon: Relations That Matter*, In: Proceedings of the 12th Conference of the European Chapter of the ACL.
- V. Mathesius. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Walter de Gruyter.
- A. Meena and T. Prabhakar. 2007. *Sentence Level Sentiment Analysis in the Presence of Conjunctions Using Linguistic Analysis*, In: Proceedings of ECIR.
- B. Pang, L. Lee, and S. Vaithyanathan. 2002. *Thumbs up? Sentiment Classification Using Machine Learning Techniques*, In: Proceedings of EMNLP.
- J. Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger, and E. van der Goot. 2011. *Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora.*, Proceedings of the 8th International Conference Recent Advances in Natural Language Processing.
- T.-T. Thet, J.-Ch. Na, Ch. S.-G. Khoo, and S. Shakthikumar. 2009. *Sentiment analysis of movie reviews on discussion boards using a linguistic approach*, In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.
- K. Veselovská. 2012. *Sentence-level sentiment analysis in Czech*, In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics.
- J. Votrubec (Raab). 2005. *Morphological Tagging Based on Averaged Perceptron*, In: WDS06 Proceedings of Contributed Papers.
- J. Wiebe. 2002. *Instructions for Annotating Opinions in Newspaper Articles.*, Department of Computer Science Technical Report TR-02-101 , University of Pittsburgh, Pittsburgh, PA.
- J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. *Learning subjective language*, Computational Linguistics, 30, 3.
- M. Wiegand and D. Klakow. 2009. *The Role of Knowledge-based Features in Polarity Classification at Sentence Level*, In: Proceedings of the 22nd International FLAIRS Conference (FLAIRS-2009).
- T. Wilson. 2008. *Fine-Grained Subjectivity Analysis*. PhD Dissertation, Intelligent Systems Program, University of Pittsburgh.

A Phrase-Based Opinion List for the German Language

**Sven Rill, Sven Adolph,
Johannes Drescher, Dirk Reinel,
Jörg Scheidt, Oliver Schütz,
Florian Wogenstein**

University of Applied Sciences Hof
Hof, Germany

(srill,sadolph,jdrescher,dreinel,jscheidt,
oschuetz,fwogenstein)@iisys.de

Abstract

We present a new phrase-based generated list of opinion bearing words and phrases for the German language. The list contains adjectives and nouns as well as adjective- and noun-based phrases and their opinion values on a continuous range between -1 and $+1$. For each word or phrase two additional quality measures are given. The list was produced using a large number of product review titles providing a textual assessment and numerical star ratings from Amazon.de. As both, review titles and star ratings, can be regarded as a summary of the writers opinion concerning a product, they are strongly correlated. Thus, the opinion value for a given word or phrase is derived from the mean star rating of review titles which contain the word or phrase. The paper describes the calculation of the opinion values and the corrections which were necessary due to the so-called “J-shaped distribution” of online reviews. The opinion values obtained are amazingly accurate.

1 Introduction

The amount of textual data increases rapidly in the World Wide Web, so does the need of algorithms enabling an efficient extraction of information from this data. Namely the extraction of opinions (opinion mining) gained increasing attention in the research community.

Many opinion mining algorithms and applications need text resources like polarity lexicons / opinion lists consisting of words or phrases with

Roberto V. Zicari, Nikolaos Korfiatis

Goethe University Frankfurt
Frankfurt am Main, Germany

zicari@informatik.uni-frankfurt.de
nikos@dbis.cs.uni-frankfurt.de

their opinion values. These opinion values either classify words in positive, neutral or negative words or give opinion values in a continuous range between -1 and $+1$ providing a finer resolution in the measure of their opinion polarities. Polarity lexicons usually are derived from dictionaries or text corpora. The quality of the used lexical resources is of utmost importance for the quality of the results obtained from opinion mining applications. In (Oelke et al., 2009) the authors applied opinion mining on customer feedback data. They analyzed error sources and came to the conclusion that about 20% of the errors occurred due to faults in the opinion list. In addition, most of the other error sources were related to the opinion list.

Using opinion lists in applications raises the question, how to handle phrases containing opinion words plus one or several valence shifters (intensifiers or reducers) or negation words or combinations of both. One could assume that negation words just change the sign of the opinion value and valence shifters change its absolute value by a defined step. In some cases, however, this is not correct. For example, “gut” - ‘good’ and “perfekt” - ‘perfect’ are positive opinion words. The negation of “gut”, “nicht gut” - ‘not good’ can be regarded as a negative phrase, although “nicht perfekt” - ‘not perfect’ cannot.

Similar effects occur for valence shifter words. In the field of sentiment composition, the handling of these valence shifters and negation words is discussed in several papers (Choi and Cardie, 2008; Klenner et al., 2009; Liu and Seneff, 2009; Moilanen and Pulman, 2007).

An alternative way is the inclusion of intensifiers, reducers and negation words directly in the opinion list. We follow this approach and use an algorithm presented in (Rill et al., 2012) to generate a list containing opinion bearing phrases together with their opinion values for the German language.

2 Related Work

The automatic extraction of opinions and sentiments has gained interest as the amount of textual data increases permanently. Thus, a lot of research work has been done in the area of opinion mining. An overview of the whole topic recently has been given in (Liu and Zhang, 2012).

Text resources, namely lists of opinion bearing words together with an assessment of their subjectivity, have been provided for several languages.

For the English language publicly available word lists are SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006a; Esuli and Sebastiani, 2006b), Semantic Orientations of Words (Takamura et al., 2005), the Subjectivity Lexicon (Wilson et al., 2005) and two lists of positive and negative opinion words provided by (Liu et al., 2005).

Also for the German language lists of opinion bearing words already exist. In (Clematide and Klenner, 2010) the authors described a polarity lexicon (PL) listing the opinion values for about 8,000 German nouns, adjectives, verbs and adverbs. The words are classified into negative and positive words with opinion values in six discrete steps. In addition, PL includes some shifters and intensifiers. The list was generated using *GermaNet*, a German lexicon similar to *WordNet*, which already was used to derive English polarity lexicons.

In (Waltinger, 2010) *GermanPolarityClues* (GPC) was introduced. It consists of more than 10,000 German nouns, adjectives, verbs and adverbs classified as positive, neutral or negative opinion words. For a part of these words, also the probabilities for the three classes are given. GPC also lists about 300 negation words. *GermanPolarityClues* was not derived directly from German text data but was produced using a semi-automatic translation approach of English-based

sentiment resources.

PL and GPC will be used as benchmarks for our list (see Section 4.3).

In (Rill et al., 2012), the authors proposed a generic algorithm to derive opinion values from online reviews taking advantage of the fact that both, star ratings and review titles can be regarded as a short summary of the writer's opinion and therefore are strongly correlated. The authors use this algorithm to derive a list of opinion bearing adjectives and adjective-based phrases for the English language. In this work, we use this algorithm to produce a new opinion list consisting of adjectives and nouns as well as adjective- and noun-based phrases for the German language.

In addition and in contrast to the previous work, corrections, which are necessary due to the "J-shaped distribution" of online reviews, are applied. Reasons and implications of this "J-shaped distribution" are discussed in several publications (Hu et al., 2007; Hu et al., 2009).

Online reviews are used for several other research projects, for an overview see (Tang et al., 2009).

3 Generation of the Opinion List

3.1 General Approach

On a typical review platform of an online shop, user-written product reviews include a title and a numerical evaluation among other information. Amazon uses a star rating with a scale of one to five stars. Both, title and star rating can be regarded as a summary of the user's opinion about the product under review. Thus, the opinion expressed in the title, using opinion bearing words and phrases, is strongly correlated to the star rating. This leads to the conclusion that opinion values for words or phrases occurring in the titles of reviews can be generated by taking advantage of this correlation.

The calculation of opinion values is performed in several steps, described in the subsequent sections. Figure 1 depicts the whole process.

3.2 Data Retrieval and Preprocessing

Crawling and Language Detection

Basis of this work are review titles and star ratings crawled from the German Amazon site

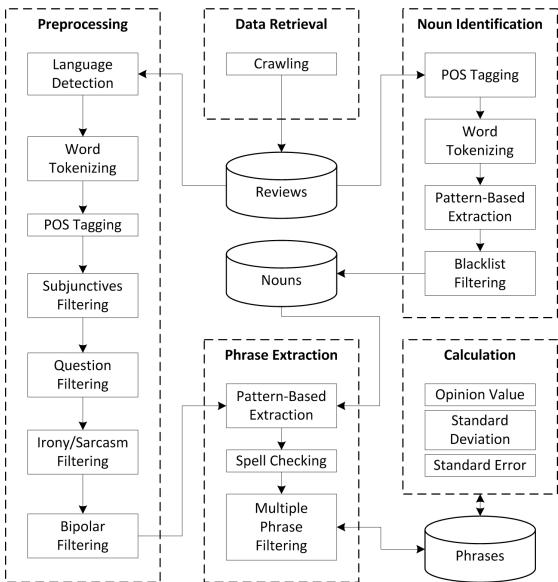


Figure 1: Overview of the opinion list generation.

(Amazon.de). The review texts and the additionally available information like product information, helpfulness count and the comments on the reviews are of no interest for this project. Thus, they were excluded from the crawling. The data set consists of about 1.16 million pairs of review titles and star ratings.

A language detection was performed using the Language Detection Library for Java by S. Nakatani¹.

Word Tokenizing and Part-of-Speech Tagging

The word tokenizing and part-of-speech (POS) tagging are the next preprocessing steps to be performed.

In some cases, the POS tagging is quite difficult for review titles as they sometimes consist only of some words instead of a complete sentence and therefore words are quite often mistagged. Especially, if the title starts with a capitalized adjective, it is often mistagged as a noun or a named entity, e.g., “Gutes Handy!” - ‘Good mobile phone!’ “Gutes” - ‘Good’ has to be tagged as an adjective. We used the *Apache OpenNLP POS Tagger*², with the maximum entropy model which was trained using the *TIGER treebank* (Brants et al., 2002). The POS tags obtained are given using

¹<http://code.google.com/p/language-detection/>

²<http://opennlp.apache.org/>

the *Stuttgart-Tübingen Tagset* (STTS) (Schiller and Thielen, 1995).

To improve the POS tagging in respect to the above-mentioned problem, we converted each first word of a review title to small letters, if it is tagged as a noun, and repeated the POS tagging. If the probability for being an adjective exceeds the noun probability after the conversion to small letters, the word is taken to be an adjective.

Filtering

In some cases star rating and textual polarity are not correlated. Therefore, we perform the same filtering steps like proposed in (Rill et al., 2012):

- Subjunctives often imply that a statement in a review title is not meant as the polarity of the word or phrase indicates, e.g., “Hätte ein guter Film werden können” - ‘Could have been a good film’. In the German language “hätte” - ‘could’, “wäre” - ‘would be’, “könnte” - ‘might be’ and “würde” - ‘would’ are typical words indicating a subjunctive. Hence, review titles containing one of these words are omitted.
- Some titles are formulated as questions. Many of them are not useful as they often express the opposite opinion compared to the star rating, e.g., “Warum behaupten Leute, das sei gut?” - ‘Why do people say that this is good?’’. Therefore, titles are excluded if they contain an interrogative and a question mark at the end.
- Some review titles are meant ironically. Irony cannot be detected automatically in most of the cases (Carvalho et al., 2009) but exceptions exist. Sometimes, for example, emoticons like “;-)” can be regarded as signs of irony. Also, quotation marks are sometimes used to mark a statement as ironic, e.g., “Wirklich ein ‘großartiger’ Film!” - ‘Really a “great” movie!’. Thus, titles containing emoticons or quotation marks are excluded from the data set.
- The words “aber” - ‘but’, “jedoch” - ‘however’, “sondern” - ‘but’ and “allerdings” - ‘though’ are indicators for a bipolar opinion, e.g., “scheint gut zu sein, aber ...” - ‘seems

good, but ...'. Again, the star rating does not correspond to the opinion value of the opinion word so titles containing one of these words are omitted.

3.3 Opinion Word and Phrase Extraction

In this work we extract opinion words and phrases based on opinion bearing adjectives and nouns like "absolut brilliant" - '*absolutely brilliant*', "nicht sehr gut" - '*not very good*', "exzellent" - '*excellent*' (adjective-based) and "totaler Müll" - '*complete rubbish*' (noun-based). Verb-based phrases like "enttäuscht (mich) nie" - '*never disappoints (me)*' are not regarded. Opinion phrases consist of at least one opinion bearing word. In addition, they might contain shifters and/or negation words and/or other words like adverbs or adjectives.

Opinion Bearing Nouns

The first step in the construction of opinion bearing phrases is the identification of candidates for opinion bearing nouns like "Meisterwerk" - '*masterpiece*', "Enttäuschung" - '*disappointment*' or "Schrott" - '*dross*'. To create a list consisting of such words, we look at the review titles considering the nouns of following patterns as candidates for opinion bearing nouns:

1. A single noun or a single noun with an exclamation sign, e.g., "Frechheit" - '*Impudence*' or "Wahnsinn!" - '*Madness!*'.
2. A single noun with an article in front and an exclamation mark (optional), e.g., "Der Allesköninger! - *The all-roundner!*'.
3. A noun with a form of "sein" - '*to be*' or "haben" - '*have*', e.g., "Der Service ist eine Frechheit!" - '*The support is a cheek!*' or "Die Kamera hat ein Problem" - '*There is a problem with the camera*'.
4. A noun with "mit" - '*with*', or "in" - '*in*' in front, e.g., "Karte mit Macken" - '*Card with faults*' or "Trio in Höchstform" - '*Trio in top form*'.
5. A noun with a following "bei" - '*during*', e.g., "Tonstörung bei Wiedergabe" - '*Sound problem during playback*'.

Afterwards, some nouns are removed from the list according to a manually created blacklist. This is

necessary as the list still contains some mistagged words, e.g., adjectives or named entities.

Noun-Based Opinion Phrases

Every opinion bearing noun in a review title is a candidate for an opinion phrase. We start at the end of each review title. For each candidate the phrase is extended to the left as long as it fulfills one of the patterns below.

1. Single noun, e.g., "Enttäuschung" - '*disappointment*'.
2. A noun with an adjective, e.g., "absoluter Mist" - '*absolute rubbish*'.
3. A noun with one or more adverbs, adjectives and/or indefinite pronouns, e.g., "Keine absolute Kaufempfehlung" - '*Not an absolute recommendation to buy*'.

As a last step, the nouns of noun-based opinion phrases have to be lemmatized. This means that each noun has to be reduced to its canonical form. For nouns, the several plural forms as well as forms of several cases are changed to the nominative singular, e.g., "des Meisterwerks" and "Meisterwerke" to "Meisterwerk" - '*masterpiece*'. For the lemmatizing, we used the Web service provided by the *Deutscher Wortschatz* project of the University Leipzig (Quasthoff, 1998).

Adjective-Based Opinion Phrases

In contrast to the construction of noun-based phrases, single opinion bearing adjectives and adjective-based phrases can be retrieved in one step. They are extracted according to the following patterns:

1. Single adjective, e.g., "Großartig!" - '*Great!*'.
2. One or more adverbs, particles or past participles and an adjective, e.g., "Sehr guter (Film)" - '*Very good (movie)*', "Nicht gut (für ein iPad)" - '*Not good (for an iPad)*', "Gewohnt gut" - '*Good as usual*', "Gut verarbeitet" - '*Well processed*'.
3. Like pattern number 2 but with one or more adverbs replaced by adjectives, e.g., "Sehr schöner kleiner (Bildschirm)" - '*Very nice little (screen)*'.

As for noun-based opinion phrases a lemmatizing of the base adjectives has to be performed. For German adjectives, for example, the forms “großer” and “großes” have to be reduced to “groß” - ‘big’.

Filtering of Opinion Bearing Phrases

At this stage of the algorithm a spell checker is applied to identify misspelled words in the phrases. We used the *Hunspell Spell Checker*³ with the *de-DE frami* word list⁴. In cases where the spell checker marks a word as misspelled, this review title is omitted.

In addition, only titles with exactly one opinion phrase are accepted at this point. The reason is that titles containing more than one opinion phrase have the problem that phrases normally have different opinion values. In extreme cases they are contradicting, e.g., “Gute Geschichte, schlecht geschrieben” - ‘Good story, badly written’. Therefore, titles containing two or more opinion phrases are discarded.

3.4 Calculation of Opinion Values

After the preselection steps described in the sections before, the data set consists of about 420,000 review titles each having one opinion phrase and a star rating between one and five. For each phrase occurring frequently in the review titles, the opinion value is calculated by transposing the mean star rating of all review titles having this phrase to the continuous scale $[-1, +1]$, assuming that a three star rating represents a neutral one:

$$OV = \left(\frac{\sum_{s=1}^5 n_s \cdot s}{n} - 3 \right) : 2 \quad (1)$$

Here, s is the number of stars (one to five), n_s the number of review titles with s stars and n the total number of review titles for the given phrase. Frequently at this stage means that a phrase has to occur at least ten times in the preselected review titles.

In addition to the opinion value, two quality measures are calculated. The first one is just the standard deviation σ_{OV} of the opinion value. It

is a measure of how much the star rating spreads for a given opinion phrase. The second one is the standard error (SE) calculated by dividing the standard deviation by the square root of the number n_i of review titles having phrase i . In addition to the spread of the stars, it indicates on how many review titles the opinion value of a given phrase is based.

These quality measures can be helpful in the usage of our list. If the opinion value of a phrase is near zero, the σ_{OV} indicates whether the phrase is really used mainly in neutral reviews (small σ_{OV}) or in both very positive and negative reviews (large σ_{OV}). For example, the word “fassungslos” - ‘stunned’ with an opinion value of 0.05 is used as a positive word for book reviews, and also as a negative word for quality features. This results in a large σ_{OV} .

Table 1 shows some opinion phrases together with their opinion values and the two quality measures at this point of the algorithm.

Phrase	OV	σ	SE
großartig - great	0.95	0.23	0.01
einfach gut - just good	0.93	0.19	0.01
sehr gut - very good	0.90	0.22	0.00
nur Schrott - total dross	-0.85	0.52	0.10
nur schlecht - very bad	-0.97	0.16	0.01

Table 1: Some words and phrases with their opinion values and two quality measures.

3.5 Correction of the Opinion Values

Most opinion values already look reasonable at this point, but some are not yet satisfactory. Especially, the values for some single adjectives, expected to carry no opinion, are shifted to positive values. The reason is that samples of online reviews often show a so-called “J-shaped distribution”. This means that for a big collection of reviews on a one to five scale, the star distribution has a parabolic shape with a minimum at about two stars. In our sample of single adjectives, we find about 7% 1-star, 5% 2-star, 6% 3-star, 17% 4-star and 65% 5-star review titles. For an adjective, expressing no opinion and therefore being distributed equally over all review titles, this means that it will receive an opinion value accord-

³<http://hunspell.sourceforge.net/>

⁴http://extensions.openoffice.org/de/project/dict-de_DE_frami

ing to this “J-shaped distribution”. The mean star rating of this distribution at 4.28 stars corresponds to an opinion value of 0.64, so many neutral adjectives get an opinion value in this region. Thus, a correction is necessary. We proceed in the following way. In a first step, we classify all single adjectives into the two classes “Neutrals Following J-shaped Distribution” and “Others”. For words in the first class we require their star distribution to follow the “J-shaped distribution” quantitatively and qualitatively. The quantitative deviation from this distribution we estimate by calculating the measure

$$S_{J_1} = 1 - \sqrt{\sum_{s=1}^5 \left(\frac{n_s}{n} - a_s \right)^2} \quad (2)$$

where a_s is the relative frequency of s -star review titles in the whole sample. We found out that a value of S_{J_1} greater than 0.85 indicates that the star distribution for a word is quite similar to the “J-shaped distribution” of the whole set of review titles. In addition, we set the measure S_{J_2} to TRUE (T) if $n_1 \geq n_2$ and $n_4 \leq n_5$, otherwise S_{J_2} is set to FALSE (F). In this way, we make sure that the correction of the opinion value is only applied to words also following the “J-shaped distribution” qualitatively.

In the second step, the opinion values get corrected for words fulfilling both conditions by weighting the star rating frequencies n_s with the relative frequencies of the whole sample (a_s):

$$OV_c = \left(\frac{\sum_{s=1}^5 \frac{n_s \cdot s}{a_s}}{\sum_{s=1}^5 \frac{n_s}{a_s}} - 3 \right) : 2 \quad (3)$$

Table 2 lists some words and the effect of the “J-shaped distribution” correction to them.

The same correction procedure has to be applied for single nouns. The corresponding fractions of s -star ratings are 13% for 1-star, 6% for 2-star, 8% for 3-star, 15% for 4-star and 58% for 5-star review titles containing the single nouns. The correction is calculated in the same way as described above. Table 3 gives some examples for nouns with corrected opinion values.

Adjective	<i>OV</i>	S_{J_1}	S_{J_2}	OV_c
schnell - <i>fast</i>	0.69	0.96	F	—
gut - <i>good</i>	0.67	0.75	F	—
hübsch - <i>nice</i>	0.56	0.71	T	—
jung - <i>young</i>	0.66	0.93	T	0.12
schwarz - <i>black</i>	0.67	0.97	T	0.08
andere - <i>other</i>	0.57	0.88	T	0.01

Table 2: Some adjectives and the effect of the “J-shaped distribution” correction. Bold opinion values mark the values entering the final list.

Noun	<i>OV</i>	S_{J_1}	S_{J_2}	OV_c
Traum - <i>dream</i>	0.91	0.67	T	—
Gefühl - <i>feeling</i>	0.69	0.88	F	—
Gehirn - <i>brain</i>	0.59	0.93	T	0.08
Zeit - <i>time</i>	0.57	0.96	T	0.06
Kunst - <i>art</i>	0.56	0.96	T	0.05

Table 3: Some nouns and the effect of the “J-shaped distribution” correction. Bold opinion values mark the values entering the final list.

4 Results and Discussion

4.1 Statistical Summary

After these steps, our list consists of 3,210 words and phrases. Table 4 summarizes the frequencies of adjectives, adjective phrases, nouns and noun phrases as well as the number of weak and strong subjective words and phrases. We regard a word or phrases as weak subjective, if the opinion value lies between 0.33 and 0.67 or -0.33 and -0.67 . Having an opinion value greater than 0.67 or smaller than -0.67 , a word or phrase is assumed to be strong subjective. Words and phrases with an opinion value between -0.33 and $+0.33$ we regard as being neutral.

	total	n	ws	ss
Adjectives	1,277	390	400	487
Adjective phrases	938	135	170	633
Nouns	502	124	112	266
Noun phrases	493	51	88	354
Sum	3,210	700	770	1,740

Table 4: Number of neutral (n), weak subjective (ws) and strong subjective (ss) words and phrases.

More than half of the words and phrases are strong subjective while less than one fourth are neutral.

4.2 Examples of Opinion Values for Words and Phrases

Table 5 lists some adjectives and adjective phrases, Table 6 some nouns and noun phrases together with their opinion values.

Adjectives	OV
großartig - great	+0.95
exzellent - excellent	+0.91
optimal - optimal	+0.87
amüsant - amusing	+0.71
gut - good	+0.67
brauchbar - useful	+0.38
durchschnittlich - average	-0.02
mies - crummy	-0.44
schlecht - bad	-0.56
enttäuschend - disappointing	-0.65
unbrauchbar - unusable	-0.86
grottenschlecht - abysmal	-0.96
Adjective Phrases	OV
extrem gut - extremely good	+1.00
sehr sehr gut - very very good	+0.97
sehr gut - very good	+0.90
sehr praktisch - very useful	+0.87
gewohnt gut - good as usual	+0.78
recht gut - quite good	+0.48
nicht schlecht - not bad	+0.38
nicht optimal - not optimal	-0.05
eher schwach - rather weak	-0.26
ziemlich langweilig - rather boring	-0.61
nicht gut - not good	-0.64
sehr schlecht - very bad	-0.83
ganz mies - just crummy	-1.00

Table 5: Examples of adjectives and adjective phrases with their opinion values.

The opinion values look plausible. The whole range of possible opinion values is covered.

Interesting is a look at the role of valence shifters and negation words.

The valence shifter “sehr” - ‘very’ shifts the opinion values in the expected direction. Also a multiple usage (“sehr sehr gut” - ‘very very good’)

leads to a further increase of the opinion value. Also visible is the fact that a valence shifter can not be described with a common factor which can be applied to all adjectives.

The results for the negation word “nicht” - ‘not’ show that it does not always change the sign of the opinion value (as for “gut” - ‘good’) leaving its absolute value nearly unchanged. In fact, for many words the negation changes a strong polarity to a weak or neutral one, e.g., for “schlecht” - ‘bad’ or “optimal” - ‘optimal’.

Nouns	OV
Weltklasse - world class	+0.98
Superprodukt - super product	+0.97
Meisterwerk - masterpiece	+0.94
Durchschnitt - average	-0.05
Unsinn - nonsense	-0.52
Enttäuschung - disappointment	-0.71
Zumutung - impertinence	-0.77
Frechheit - impudence	-0.89
Zeitverschwendung - waste of time	-0.93
Noun Phrases	OV
absolutes Muss - absolute must	+0.97
sehr gute Qualität - very good quality	+0.94
großer Spaß - great fun	+0.75
nur Durchschnitt - only average	-0.02
mangelnde Qualität - lack of quality	-0.55
absoluter Fehlkauf - absolute bad buy	-1.00

Table 6: Examples of nouns and noun phrases with their opinion values.

Also for the nouns the opinion values obtained look good. Again, the valence shifters change the opinion values in the expected way.

4.3 Comparison to Existing Lists

To compare our list with the two existing polarity lexicons (PL and GPC, see Chapter 2), we compare some single opinion words, which are frequently used in text sources. The result of this comparison is given in Table 7. For both, PL and GPC, the polarity (positive - P and negative - N) is given. For GPC, the attached numbers give the probability of the word having this polarity. For PL, the number indicates the strength of the opinion. We can see that the values in the three lists

agree in the sense that the classification into positive and negative words is consistent in all cases.

Adjectives	OV	PL	GPC
toll - <i>great</i>	+0.87	P 1.0	P 0.43
zufrieden - <i>satisfied</i>	+0.82	P 1.0	P 0.49
spannend - <i>exciting</i>	+0.78	P 1.0	P 0.17
gut - <i>good</i>	+0.67	P 1.0	P 0.32
schwach - <i>weak</i>	-0.18	N 0.7	N 0.61
schlecht - <i>bad</i>	-0.56	N 1.0	N 0.61

Table 7: Opinion values taken from our list (OV) compared to values from the polarity lexicon (PL) and the GermanPolarityClues (GPC).

Quantitatively, each of the two benchmark lists contain more than twice as many words than our list.

However, we see an advantage of our approach in the fact that our list not only contains single words but also phrases, so the treatment of negation and intensification gets much easier. Furthermore, instead of only classifying the words into a few polarity classes, we calculated opinion values in a continuous range between -1 and $+1$ which leads to a more precise assessment when applying the list, e.g., in the field of aspect-based opinion mining.

4.4 Shortcomings and Future Work

In Section 3.5 we discussed a correction necessary due to the “J-shaped distribution” of online reviews. This special distribution could cause another effect on the opinion values of single adjectives carrying a negative opinion. Some of these words sometimes are used ironically or in an idiomatic expression and therefore do not express a negative opinion. If these cases are rare and equally distributed over all review title, this effect would not lead to shifts in the opinion values. However, as these “misusages” cannot be regarded as distributed equally due to the “J-shaped distribution” of reviews, this can lead to a shift of the opinion values to positive values for some opinion words.

As stated in Section 3.4 we assumed a three star rating to be the expression of a neutral opinion about a product. However, this has not to be exactly the case. In fact, reviews with a three star

rating are sometimes regarded as slightly negative ones. This could lead to a shift of opinion values to positive values. If the mean star rating for neutral reviews is shifted by 0.1 stars to 3.1 stars for example, the resulting opinion values would be shifted by 0.05 to the positive side. This systematic uncertainty should be kept in mind for phrases with opinion values having a small statistical error and being close to zero.

A general problem of the approach results from the fact, that online reviews were used as a text resource to derive the opinion values. Thus, the vocabulary used in these reviews determines the content of the opinion list. This leads to the conclusion, that opinion lists produced using this approach may be suitable for analyzing user generated content from Web 2.0 sources but may not be applicable for other text resources.

To measure the quality of the list, we intent to perform benchmark tests using manually created opinion lists. Later we will perform also a quantitative evaluation.

For a later version of the list, we want to enrich it with values for opinion bearing phrases based on verbs.

5 Conclusion

In this paper we calculated opinion values for German adjective- and noun-based phrases with a fine granularity using the titles of Amazon reviews together with the star rating. Necessary corrections were applied.

It seems as if we got astonishingly good results for more than 1,700 single words and 1,400 phrases, which are to evaluate in detail. The list obtained will be made available to the community soon.

Acknowledgments

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-10)*, pages 2200–2204.
- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER

- treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41.
- Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement (TSA-09)*, pages 53–56.
- Yejin Choi and Claire Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pages 793–801.
- Simon Clematide and Manfred Klenner. 2010. Evaluation and Extension of a Polarity Lexicon for German. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA-10)*, pages 7–13.
- Andrea Esuli and Fabrizio Sebastiani. 2006a. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 193–200.
- Andrea Esuli and Fabrizio Sebastiani. 2006b. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 417–422.
- Nan Hu, Paul A. Pavlou, and Jennifer Zhang. 2007. Why do Online Product Reviews have a J-shaped Distribution? Overcoming Biases in Online Word-of-Mouth Communication. *Marketing Science*, 198:7.
- Nan Hu, Jie Zhang, and Paul A. Pavlou. 2009. Overcoming the J-shaped Distribution of Product Reviews. *Communications of the ACM*, 52:144–147.
- Manfred Klenner, Stefanos Petrakis, and Angela Fahrni. 2009. Robust Compositional Polarity Classification. In *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-09)*, pages 180–184.
- Jingjing Liu and Stephanie Seneff. 2009. Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*, pages 161–169.
- Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer US.
- Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International World Wide Web Conference (WWW-05)*, pages 342–351.
- Karo Moilanen and Stephen Pulman. 2007. Sentiment Composition. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-07)*, pages 378–382.
- Daniela Oelke, Ming Hao, Christian Rohrdantz, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug, and Halldór Janetzko. 2009. Visual Opinion Analysis of Customer Feedback Data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST-09)*, pages 187–194.
- Uwe Quasthoff. 1998. Projekt Deutscher Wortschatz. In Gerhard Heyer and Christian Wolff, editors, *Linguistik und neue Medien. Proc. 10. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*, pages 93–99. DUV.
- Sven Rill, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, and Daniel Simon. 2012. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM-12)*.
- Anne Schiller and Christine Thielen. 1995. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens "Lexikon + Text", 17./18. Februar 1994, Schloß Hohentübingen*, Lexicographica Series Maior, pages 193–203. Niemeyer, Tübingen.
- Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 133–140.
- Huifeng Tang, Songbo Tan, and Xueqi Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36:10760–10773.
- Ulli Waltinger. 2010. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-10)*, pages 1638–1642.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Human Language Technology Conference (HLT-05)*, pages 347–354.

Opinion Mining in an Informative Corpus: Building Lexicons

Patrice Enjalbert, Lei Zhang, and Stéphane Ferrari

GREYC lab - CNRS UMR 6072

University of Caen Basse-Normandie

FirstName.LastName@unicaen.fr

Abstract

This paper presents first steps of an ongoing work aiming at the constitution of lexicons for opinion mining. Our work is corpus-oriented, the corpus being of informative nature (related to avionic manufacturers) rather than opinion-oriented (as in current works dealing with social networks). We especially investigate the question of interrelation between factual information and evaluative stance. Another aspect concerns the intensity of expressed opinions. Lexicons for adjectives and adverbs have been built, based on the given corpus, and we present the principles and method used for their construction.

1 Introduction

The present work takes place inside the interdisciplinary project *Ontopitex*¹, involving computer scientists and linguists from three different laboratories as well as industrial partners. The applicative task is provided by the company *TecKnow-Metrix* and related to competitive intelligence.

The corpus concerns avionic technologies and more precisely Boeing and Airbus companies. It consists in 377 journalistic texts from economic and technical press in French language, representing approximatively 340 000 words. As expected, opinion expression is not the main characteristic of such texts. However, even “objective” facts are commonly accompanied with some evaluative stance, either by connotation (“efficient”, “noisy”, “active”...) or denotation (“attractive”, “welcome”, “good”...). Inside this context the

present paper concerns the constitution of a lexicon of adjectives and adverbs that take part in evaluative acts, either because they support by themselves an evaluation (Section 2) or because they contribute to its intensity (Section 3).

While most of current studies relative to lexicon generation for opinion mining focus on automated methods (Hatzivassiloglou and McKeown, 1997; Turney, 2002), our lexicon was built “manually”, by corpus observation. Several reasons motivate such an enterprise: 1) Participate in the conceptual and linguistic study of the phenomenon of evaluation; 2) Take into account corpus specificities (for example, relative to ambiguity); 3) Provide a reference in order to evaluate automated procedures; 4) Provide a bootstrap for automated analyses, with special interest on negation and intensity. Despite interesting efforts (Vernier et al., 2009)² such resources are especially missing in French.

2 Evaluative lexicon

A firm opposition is often drawn between so-called “objective” and “subjective” sentences or terms (Wiebe and Mihalcea, 2006). However we can easily observe that both are often combined: a property (or fact) is presented for its own but in a manner which inherently associates a positive or negative evaluation. This is especially true in our journalistic corpus. We say that such assertions provide an *axiologic* evaluation of the denoted fact or property: the evaluation comes from intrinsic (“objective”) properties of its target together with a specific domain-oriented axiology.

¹Supported by the Agence Nationale de la Recherche.

²See <http://www.lina.univ-nantes.fr/?Ressources-disponibles-sous.html>

Apart from these, we distinguish opinion judgement that are closely related with the speaker's relation with the target, hence expressing directly his/her "subjectivity": we call them *estimations*. These two categories might be associated with the traditional connotation/denotation paradigm, but with a different theoretical background. In the following we will show how these ideas may be applied to help the constitution of lexicons. For sake of brevity we will focus on adjectives but the method applies to adverbs as well, with minor adaptations. We will put stress on linguistic tests that allow to detect such terms inside texts. Note that the proposed method was in fact first elaborated in order to detect all kinds of evaluative segments in texts, for example in order to establish reference annotations.

2.1 Axiologic (intrinsic) evaluation

We call *axiologic-evaluative* an adjective that fulfills the two conditions:

1. It implies a necessary or contingent property;
2. This property can be considered as desirable or on contrary regrettable. This qualification is relative to an axiology (or norm), reflecting some goals in a given situation.

A *linguistic test* readily comes from this definition. In order to know if an occurrence of an adjective is axiologic-evaluative, insert the proposition *comme il est souhaitable/regrettable* "as it is desirable/regrettable" in the text. One of the alternatives should provide a clear contradiction, and the other a strengthening of the evaluative force of the assertion. Besides deciding of the evaluative qualification, the polarity is inferred in the obvious way. Otherwise, the insertion is irrelevant and the term occurrence non evaluative. Examples³:

- (1) *It [Boeing 787] is particularly innovative* [as it is desirable (OK) / regrettable (Contradiction)].
- (2) *But the production of this new plane is so complex* [as it is desirable (Contradiction) / regrettable (OK)] *that the 787 adds problems to problems*.
- (3) *The ins and outs of this contract are complex and varied* [as it is desirable (Irrelevant) / regrettable (Irrelevant)]. *They are not confined to ...*

Note that the formulation of the test may have to be slightly accommodated to the embedding

³All examples come from the French corpus. For sake of brevity, we only mention the English translation.

sentence, for example in terms of tense and verbal mode. Also observe that these definitions primarily apply to occurrences. An application of the test *in abstracto*, without context, is always possible: *X is Adj [as it is desirable/regrettable]* but with attention to the possible ambiguities. For example, *complexe* "complex", that our test reveals as evaluative in (2) but not in (3).

A deeper analysis can show various *axiologic dimensions*. After corpus examination, we propose a set of dimensions closely related with the discourse domain (which is mainly economic/commercial and technical) such as: activity-reactivity (of a person or institution), quality (of a product or device), commercial performance, etc. Such corpus-oriented evaluative dimensions appear more relevant, and easier to assign to each lexical item, than generalist and rather psychologically oriented ones as in (Bednarek, 2009) or (Martin and White, 2005).

2.2 Estimations (extrinsic evaluation)

An *estimation* is a statement expressing a positive/negative appreciation of an entity, but which does not imply any intrinsic property of the qualified entity. Two (major) classes of adjectives seem to fall in that category:

- qualifications that implicitly or explicitly imply a relation between the speaker and the entity. Here we find common terms such as: *acceptable* "acceptable", *bienvenu* "welcome", *décevant* "disappointing" ...
- qualifications expressing an overall judgement of the entity, so-to-speak summing-up a bunch of facts or properties evaluated according to definite axiologies⁴: *bon/mauvais* "good/bad", *beau* "nice, beautiful", *brillant* "brilliant", *célèbre* "famous" ... Hence again, the speaker is present as the one who has made the integration of different features⁵.

The linguistic test has to stage that enunciative characteristic (presence of the speaker). We pro-

⁴Indeed, this kind of judgement is often accompanied in the corpus by informations that provide its motivation.

⁵This class of terms is especially subject to ambiguities and context sensitivity. For example *good* may also be an intensifier or catch a specific aspect of the qualified entity as in *a good dish*. To be more precise if some specific property emerges from the sentence, this results from the combination with the qualified name, not from the adjective alone.

pose to insert the following comment: "To say so reveals a good/bad appreciation of the speaker but does not mention any specific property of it". Examples (we leave the reader fill the sentence with the comment):

(4) *Le titre continue de nous paraître très attractif*
"the share still seems very attractive"

(5) *Les résultats ne sont guère plus brillants pour British Airways* "The perspectives are no more brilliant for British Airways"

2.3 Experiments

Lexicon building The construction of the lexicons was performed according to the following steps, same for evaluative terms and intensity modifiers.

Step 1. Collect the terms tagged as adjective or adverb by our p.o.s. tagger⁶ (1657 and 494 respectively) and clean this list according to tagging errors, removing terms that are certainly non evaluative in any context (ethnonyms, logical connectives ...). 625 adjectives and 140 adverbs remain as "possibly evaluative or intensifier".

Step 2. A concordancer is constituted with these terms and a labeling is performed according to the above mentioned categories, using the linguistic tests.

Step 3. Four lexicons are established in XML format (adjectives or adverbs, evaluative or modifier). Due to ambiguities a same item may appear in different lexicons An item is stored as soon as it possesses one evaluative occurrence.

On the overall we retained 415 adjectives (283 "axiologic", 92 "estimations" and 40 modifiers) and 86 adverbs (36 evaluative and 50 modifiers). An evaluative entry contains the following informations: lemma, p.o.s., subclass (axiologic/estimation), polarity, intensity. The question of "intensity" is addressed in Section 3.2.

Evaluation Two tests were performed as a first attempt to evaluate the reliability of our lexicons. A first one consisted in a projection on a similar corpus. Place lacks for details, but the overall result is a good preservation; notably, 90% of entries presents in both corpora have the same labeling.

The second test took advantage of an experiment concerning negation (cf. Section 3). It relies on a corpus of articles from the French journal *Le Monde* concerning similar topics as in our main

one (about 3.9 M words). 25 thousand occurrences of our adjectives in a negative context were obtained by syntactic patterns, such as: "n'est pas ADJ" (*is not ADJ*). From this result, we selected a sample of 125 sentences (no more than 2 for each lemma). Then two annotators had to decide for each occurrence: (a) if it is an evaluative statement and if so (b) if it is and axiologic one (in the above sense) or an estimation and (c) the polarity (including the effect of negation).

The agreement was about 93% for test (a), 82% on test (b) and 96% on test (c) (respectively raised to 94, 86, and unchanged after coordination). The disagreement over the polarity is mainly due to some extreme adjectives in negative context such as "n'est pas catastrophique" *is not catastrophic* (further studied in Section 3.2). These results (cautiously) advocate for reliability of the lexicon, including polarity; concerning the axiologic/estimation distinction, the notion appears as quite relevant and mostly consensual, with a fuzzy zone as expected.

3 Intensity

Semantic orientation, i.e. polarity and intensity, of evaluative segments is a key issue in opinion mining (Turney, 2002). It is determined by informations in the evaluative lexicon combined with negation and intensity modifiers (sometimes together called *valence shifters* (Zaenen and Polanyi, 2004)). We present here the part of our work devoted to intensity: first the construction of a lexicon of modifiers, and second a procedure to assign an intensity to evaluative terms.

3.1 Intensity modifiers: lexicon building

Information on intensity is notably supported by both adjectives and adverbs. Concerning adverbs, we distinguish a closed list of grammatical items (*très* "very", *peu* "a little"...) and true lexical items, which can be seen as a subclass of manner adverbs, and concentrate on the latter.

Extracting such a lexicon from the corpus is rather easy. Adjectives are applied to nouns denoting a gradable entity, or more generally that possess some gradable feature: in the present context it will be a graded evaluative value, as in *une victoire complète* "a complete victory", *une belle réussite* "a nice succes" ... Hence, we can apply the following simple test : "replace the Adj by *important/petit* "important/small" (or

⁶Due to the second industrial partner Noopsis.

close variants)”. In fact *important/small* appears as a weak synonymous in the sense of implication: *Adj N* implies *important/small N*. The situation is similar for adverbs. They can be replaced with little meaning loss by *très (adj) / beaucoup (verb phrase)* “very/much” or close terms⁷.

Following the same procedure as in Section 2.3 we get 40 “intensity” entries for adjectives and 50 for adverbs (out of 415 and 86 respectively). The XML format codes for the following features: lemma, p.o.s., and two slots to describe the role as intensity operators: a *direction* - “ascending” for intensifiers and “descending” for moderators - and *force* - “standard” or “extreme”. Force was first determined by a test of compatibility similar as the one described now for evaluative adjectives.

3.2 Intensity of evaluative adjectives

The question here is to assign intensity to evaluative lexical terms. A first decision to be taken is the number of such values to be considered. In a first step, we fixed this number to 2, “medium” and “extreme”, for lexical items, leading to 5 values for evaluative statements by combination with negation or modifiers (Zhang and Ferrari, 2012). In our opinion, the concept of “extreme intensity” is quite relevant and useful. First we may observe that in the experiment presented Section 2.3 disagreement over polarity is due to some adjectives with “extreme force”. But the main argument is that it provides a firm, linguistically motivated, ground for assignment of intensity values to lexical items, as described now.

Our test is based on the hypothesis that extreme qualities cannot vary in intensity; in other words, *extreme* adjectives (or adverbs) are *non-gradable*. First we build a lexicon of grammatical intensifiers (proposed by Charaudeau (1992) and Noailly (1999)) which consists in 5 classes: **low** – *un peu* “a little”, **moderate** – *moyennement* “fairly”, **high** – *très* “very”, **extreme** – *extrêmement* “extremely” and **relative** – *trop* “too much”. Then, we count the frequency of co-occurrences respecting the pattern “intensifier + evaluative adjective”⁸ in a big corpus of articles from the french journal *Le Monde* (20 years). For each intensifier class int we gather the set \emptyset_{int} of adjec-

ives of frequency 0 or 1, i.e. adjectives that can hardly or not at all be varied by these intensifiers. We observe the following properties:

- $|\emptyset_{\text{low}}| = 37$, $|\emptyset_{\text{moderate}}| = 78$, $|\emptyset_{\text{high}}| = 22$, $|\emptyset_{\text{extreme}}| = 78$, $|\emptyset_{\text{relative}}| = 42$.
- $\emptyset_{\text{high}} \subset \emptyset_{\text{extreme}}$, $\emptyset_{\text{relative}} \subset \emptyset_{\text{extreme}}$, $\emptyset_{\text{low}} \subset (\emptyset_{\text{extreme}} \cup \emptyset_{\text{moderate}})$.
- $\text{nonGradable} = \bigcap_{\text{int}} \emptyset_{\text{int}}$, $|\text{nonGradable}| = 7$

From these observations, we conclude first that the “non-gradable” criterion is too strict, since it retains a set of only 7 items, clearly too small. We observed then that the words contained in $\emptyset_{\text{moderate}}$ and not in $\emptyset_{\text{extreme}}$ cannot be qualified as “extreme”. Moreover we found adjectives that cannot be varied by extreme intensifiers but do so by all others, such as *terrible* “terrible”, *extraordinaire* “extraordinary”, *catastrophique* “catastrophic”, etc. It appears then that $\emptyset_{\text{extreme}}$ gathers all plausible candidates and we decided to check carefully these 78 words. For semantic-morphologic reasons the subset of words derivated from verbs or nouns are not extreme⁹, and in addition, we found a few moderate words and technical words. Finally we retained about half of this set (34) as genuine extreme adjectives.

We conclude that our method allows to perform an efficient filtering for human validation: in our case 78 candidates were automatically selected out of 283 adjectives.

4 Conclusion

In this paper we proposed a method for the creation of corpus-dependant, manually curated, lexicons for opinion adjectives and adverbs in French. We especially investigated situations where factual information and evaluation are interrelated, and the question of evaluative intensity. The method heavily relies on several linguistic tests. First experiments are encouraging regarding the quality of the created resources, presently in use in the *Ontopitex* project (and freely available on demand from the authors). Future work notably includes an extension of the method to deal with nouns and verbal expressions, and experimentations on new corpora.

⁷The intensity value can be combined with an evaluative one, as in *un beau contrat* “a nice contract”.

⁸60 intensifiers and 283 axiologic adjectives.

⁹suffixes: *-ible*, *-able*, *-ant*, *-é*, *-aire*, *-teur*

References

- M. Bednarek. 2009. Dimensions of evaluation: cognitive and linguistic perspectives. *Pragmatics and Cognition*, 17(1):181–193.
- P. Charaudeau. 1992. *Grammaire du sens et de l'expression*. Hachette París.
- V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA. Association for Computational Linguistics.
- J.R. Martin and P.R.R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave.
- M. Noailly. 1999. *L'adjectif en français*. Editions Ophrys.
- P. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL (ACL'02)*, pages 417–424. Philadelphia, Pennsylvania, USA, ACL.
- M. Vernier, L. Monceaux, B. Daille, and E. Dubreil. 2009. Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des Nouvelles Technologies de l'Information*, pages 45–69, nov.
- J. Wiebe and R. Mihalcea. 2006. Word sense and subjectivity. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072.
- A. Zaenen and L. Polanyi. 2004. Contextual valence shifters. In *Proc. AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 106–111.
- L. Zhang and S. Ferrari. 2012. Opinion analysis: the effect of negation on polarity and intensity. In *PATHOS 2012*.

What is the Meaning of 5 *'s? An Investigation of the Expression and Rating of Sentiment

Daniel Hardt

Copenhagen Business School
dh.itm@cbs.dk

Julie Wulff

Copenhagen Business School
cbs@juliewulff.dk

Abstract

Do user populations differ systematically in the way they express and rate sentiment? We use large collections of Danish and U.S. film reviews to investigate this question, and we find evidence of important systematic differences: first, positive ratings are far more common in the U.S. data than in the Danish data. Second, highly positive terms occur far more frequently in the U.S. data. Finally, Danish reviewers tend to under-rate their own positive reviews compared to U.S. reviewers. This has potentially far-reaching implications for the interpretation of user ratings, the use of which has exploded in recent years.

1 Introduction

There is a persistent stereotype concerning the way sentiment is expressed and evaluated by Scandinavians and Americans, which is illustrated by these two anecdotes. In the first anecdote, a U.S. researcher gives a talk in a Scandinavian country. After the talk, the researcher is approached by an audience member, who says, “the talk was ok”. The U.S. researcher is puzzled by this, until another member of the audience explains to him that this was actually intended to express high praise. The second anecdote: a student at the beginning of his graduate studies at a U.S. university has several meetings with a prominent faculty member, and is repeatedly told that his research ideas are “wonderful”. The student is gratified by this, until he overhears other students talking about how this faculty member seems to

always respond to ideas by calling them “wonderful”.

There is abundant anecdotal evidence that Scandinavians and Americans differ in the way they express and evaluate sentiment: compared to Americans, it seems that Scandinavians downgrade their positive expressions of sentiment. But is this stereotype actually true? In this paper, we investigate this question by analyzing large collections of Danish and U.S. film reviews. These reviews are short pieces of text, combined with a numerical rating which expresses the user's overall evaluation. In our view, such data should provide a meaningful test of the stereotype – if Scandinavians and Americans do indeed differ as we have described, this should be reflected in distributional differences in these datasets.

In particular, the hypothesis concerns distributions of very positive evaluations: compared to U.S. reviewers, we expect a Danish tendency to “downgrade” from very positive to somewhat less positive. We will examine this hypothesis from three different perspectives, in looking at the Danish data vs. the U.S. data:

1. **Ratings:** are there relatively fewer high ratings?
2. **Text:** are there relatively fewer highly positive terms?
3. **Ratings vs. Text:** are there fewer high ratings for texts of a given positivity?

In what follows, we begin with a description of the data sets. Next we examine the distribution of ratings. Then we look at the text positivity: we

develop a metric for positivity of terms, and examine their relative distributions. This is followed by an examination of the relation between ratings and texts in the two data sets. We show that the hypothesis is strongly confirmed in all three of its variants. Finally, we observe that these results could have far-reaching implications for the interpretation of recommender systems and user ratings, the use of which has exploded in recent years.

2 Data

The Danish data was downloaded from the Danish movie website scope.dk and contains rated user reviews from 829 films and has a total size of 1,624,049 words. The U.S. data was downloaded from The Internet Movie Database (imdb.com) and contains rated user reviews from 678 films and has a total size of 34,599,486 words.

A search function on www.imdb.com was used to create a list of films and matching IMDb ID tags for films produced in the years 1920-2011. 678 films on the list had a match in the Scope data on title and production year . The IMDb ID tags was used to find the page containing data for each of the films and all reviews which had a correlated rating were downloaded for those 678 films. The U.S. IMDb reviews are rated on a scale of 1 to 10, while the Danish Scope reviews are rated on a scale of 1 to 6.

3 Ratings

Figure 1 gives the number of reviews in each category for *IMDb*.

For IMDb, the top category of 10 has by far the most reviews. For the most part the number of reviews decreases from category 10, with a modest increase in the number of reviews for the lowest category, 1. This distribution makes intuitive sense – it’s not surprising that people would be most motivated to write reviews of films they are most enthusiastic about, and, to a lesser extent, also be motivated in cases where they have strong negative feelings. This has been noted in the literature: (Wu and Huberman, 2010) point out that the so-called “brag and moan” view of ratings is fairly typical (as also mentioned by (Hu et al., 2006; Dellarocas and Narayan, 2006)). The tendency of the top category to be the most frequent

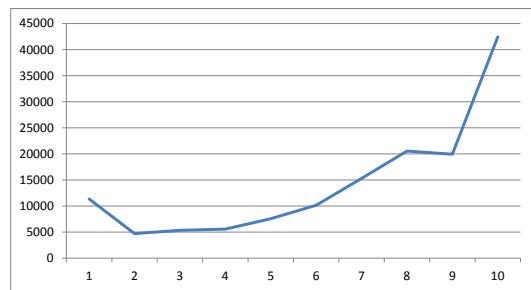


Figure 1: IMDb reviews per category

is also mentioned on the yelp.com site, where the top category of 5 is the most frequent: “The numbers don’t lie: people love to talk about the things they love!” (FAQ, 2012).

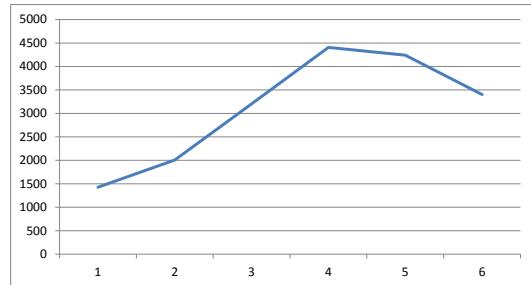


Figure 2: Scope reviews per category

There is a very different distribution in the Danish Scope data, as shown in Figure 2. Here, category 4 (out of 6) is the most frequent. This supports the general prediction that highly positive evaluations are over-represented in the U.S. data compared to the Danish data.

4 Text

We turn now to a second version of our hypothesis: that highly positive terms are over-represented in the U.S. data. We consider highly positive terms to be those that tend to occur in the most positive category and tend not to occur in the other categories.

For each category, we follow (Constant et al., 2009) in defining what they call a *log-odds* distribution for each term, as follows:

$$\text{log-odds}(x_n, R) = \ln\left(\frac{\text{count}(x_n, R)}{\text{count}(n, R) - \text{count}(x_n, R)}\right)$$

Here, n is 1, 2 or 3, denoting terms consisting of one, two or three words (i.e., unigrams, bigrams and trigrams). R is a rating category (1-6 in Scope and 1-10 for IMDb). $\text{Count}_n(R)$ is the number of occurrences of all ngrams of length n in Category R , while $\text{count}(x_n, R)$ is the number of occurrences of a particular ngram x_n in Category R . Thus we take the log of the number of occurrences of a given ngram in a category, divided by the number of occurrences of all other ngrams in that category.

Intuitively, highly positive terms are those most frequent in the top category and most infrequent in the other categories. Thus we determine positivity as follows:

$$\text{positivity}(x_n) = \text{log-odds}(x_n, R_{\text{pos}}) - \text{log-odds}(x_n, R_{\text{other}})$$

For Scope, R_{pos} is category 6, and R_{other} is categories 1 through 5, while for IMDb R_{pos} is categories 9 and 10, and R_{other} is 1 through 8.

Negativity of terms is defined in a symmetrical fashion:

$$\text{negativity}(x_n) = \text{log-odds}(x_n, R_{\text{neg}}) - \text{log-odds}(x_n, R_{\text{other}})$$

Here, R_{neg} is 1 for Scope and 1 and 2 for IMDb, while R_{other} is 2 through 6 for Scope and 3 through 10 for IMDb.

Tables 1 through 4 give the top 25 most negative and positive terms for both IMDb and Scope. For the negative terms, the most negative terms are at the top of the list, while for the positive terms, the most positive are at the bottom.

Our point of departure is that all terms with positivity greater than 0 are positive terms, while those with negativity less than 0 are negative terms. This gives the ratios of positive to negative terms as shown in Table 5.

There are somewhat more positive than negative terms in IMDb, and slightly more negative

Negativity	Term
-5.579750143176	absolutely the worst
-5.47055003096302	the worst piece
-5.47055003096302	or money on
-5.38977979264451	10 worst
-5.30349485263692	money back !
-5.20818412600157	awful movie !
-5.10282306351303	absolutely no redeeming
-5.04493752542859	of worst
-4.98431669202047	! complete
-4.88660293269565	worst piece of
-4.88587585595205	worst piece
-4.85150754157158	horrible waste of
-4.85150754157158	. * from
-4.85150754157158	no redeeming features
-4.85150754157158	the worse movies
-4.85078074773538	... avoid
-4.85078074773538	beyond bad
-4.77740634497507	this is awful
-4.77740282048268	horrible film .
-4.77739929600291	i wasted on
-4.77739929600291	this horrible film
-4.77739929600291	piece of c
-4.6973563149145	what a pile
-4.6973563149145	misfortune of seeing
-4.6973563149145	utter crap </s>

Table 1: 25 most negative terms IMDb

Positivity	Term	Negativity	Term
3.66887212985741	gets better every movie . 10	-4.870965702	elendig ! (<i>terrible</i>)
3.68657177169013	sterling hayden top ten movies	-3.867670635	ret elendig (<i>really terrible</i>)
3.70287245596558	direction is flawless	-3.666983304	min tid (<i>my time</i>)
3.70396174809963	. outstanding ! film . 9	-3.666958989	noget bras (<i>some junk</i>)
3.70396174809963	masterpiece of film	-3.577451105	skodfilm (<i>trash film</i>)
3.73786336450852	best gangster movie	-3.549191951	ringe ! (<i>bad</i>)
3.73786336450852	see movie ! . greatest	-3.531008767	lorte (<i>crap</i>)
3.77065325206472	... 10 /	-3.484669428	elendig </s> (<i>terrible</i>)
3.80131270948848	this masterpiece . . (9	-3.484669428	liggeyldig film (<i>meaningless film</i>)
3.80240201511252	movie changed my . 9.5	-3.418380646	ikke engang kan (<i>can't even</i>)
3.80240809083371	. a 10 . 10 /	-3.415652241	skod </s> (<i>trash</i>)
3.83317373851249	. 10 out . 10 out	-3.398851791	bras (<i>junk</i>)
3.8630267663954	++ </s>	-3.356843736	elendig film (<i>terrible film</i>)
3.89092504888785	favorite movies ! ! 10	-3.264221321	<s> anonym kedelig (<i>anon. boring</i>)
3.89201436800188	! 10 ! ! 10	-3.261493243	anonym kedelig (<i>anon. boring</i>)
3.92751010266618	! 10 /	-3.163755010	spilde (<i>waste</i>)
3.94759374427238		-3.149240635	stinker (<i>stinks</i>)
4.02554608429261		-3.141251329	crap
4.07433637792856		-3.112599209	elendigt (<i>terrible</i>)
4.20381518291327		-3.076666572	uudholdelig (<i>unbearable</i>)
4.41420530401696		-3.038365052	blandt min (<i>among my</i>)
4.43932293273085		-3.030337065	skod (<i>junk</i>)
4.5851638142275	outstanding ! </s>	-2.973857889	en elendig (<i>a terrible</i>)
		-2.973834211	ret nej (<i>really no</i>)
		-2.942324421	elendig (<i>terrible</i>)

Table 2: 25 most positive terms IMDb

Table 3: 25 most negative terms Scope

Positivity	Term
2.87692577130	elsker den (<i>love it</i>)
2.87848113547	film den er (<i>film it is</i>)
2.88763018980	fantastisk ! </s> (<i>fantastic !</i>)
2.89096615230	fantastisk film ! (<i>fantastic film !</i>)
2.92051716735	mest geniale (<i>most genius</i>)
2.92728613305	kan se igen (<i>can see again</i>)
2.95568635350	ret kanon (<i>really great</i>)
2.98294109871	jeg elsker den (<i>i love it</i>)
3.00279792930	genial </s> (<i>genius</i>)
3.02076226510	bedste film jeg (<i>best film i</i>)
3.05406651227	mega god (<i>mega good</i>)
3.06084014079	6 stjerner . (<i>6 stars</i>)
3.11470908673	<s> 6
3.28394697268	bedste film der (<i>best films that</i>)
3.40913662108	bedste film nogensinde (<i>best films ever</i>)
3.45951064085	geniale film (<i>genius</i>)
3.45951064085	film overhovedet (<i>films at all</i>)
3.61366505188	fortjener 6 (<i>deserves 6</i>)
3.62043477181	ret fantastisk ! (<i>really fantastic</i>)
3.75397394578	fed ! ! (<i>great</i>)
3.75397394578	ret den bedste (<i>really the best</i>)
3.86498694263	simpelthen fantastisk (<i>simply fantastic</i>)
3.97713305996	elsker den film (<i>love the film</i>)
4.06566510061	6 /
4.94095107482	6 / 6

Table 4: 25 most positive terms Scope

	Positive Terms	Negative Terms	Ratio
IMDb	50,304,859	46,642,846	1.0785
Scope	1,017,939	1,027,940	0.9903

Table 5: Ratio of positive to negative terms

terms than positive in Scope. However, it is not clear if such a comparison is meaningful. Furthermore, our hypothesis does not concern the total positivity of terms in Danish vs. English, but rather, a difference in the distribution of terms in the most positive categories. To focus our investigation on this issue, we define thresholds very close to zero such that the ratio of positive to negative terms in both data sets is 1.0.

We now can measure the number of occurrences of positive occurrences in each category. As discussed above, our hypothesis is that there should be a difference in distribution of positive terms, especially in the most positive categories. Figures 3 and 4 show that there is indeed a striking difference in distribution.

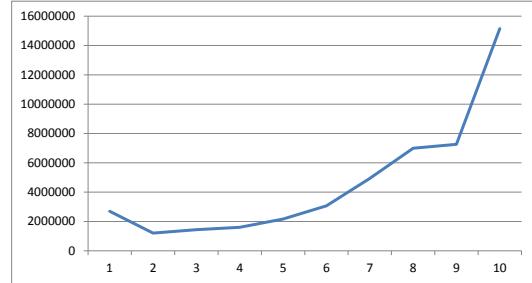


Figure 3: IMDb positive terms per category

5 Ratings vs. Text

We have shown that the hypothesis has been confirmed in two ways: first, there are proportionately more top rated reviews in the U.S. data compared to the Danish data. Second, there are proportionately more occurrences of positive terms in the top categories in the U.S. data vs. the Danish data. We now wish to tease apart these two factors, and pose the question: does the numerical

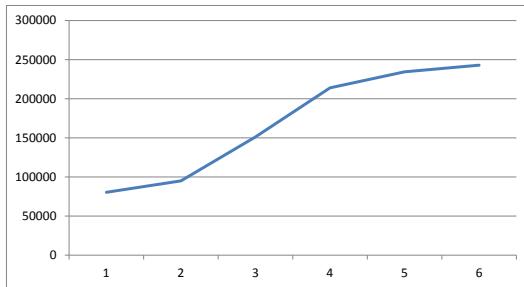


Figure 4: Scope positive terms per category

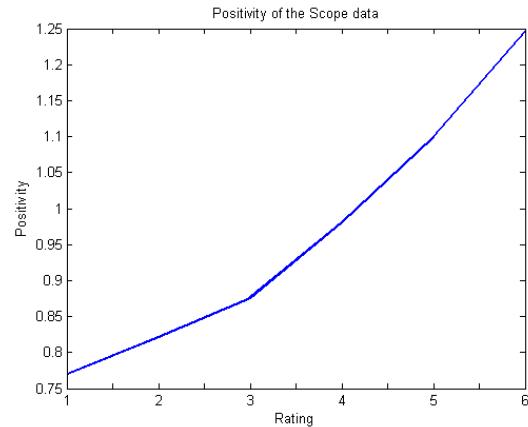


Figure 6: Scope positivity

rating correspond to the positivity of the review?

We define the positivity of a text as the ratio of positive occurrences to negative occurrences in that text. This can be used to assess the positivity of a given review, or the positivity of the complete collection of reviews in a given category. Figures 6 and 5 show the positivity of reviews in each rating category, for Scope and IMDb.

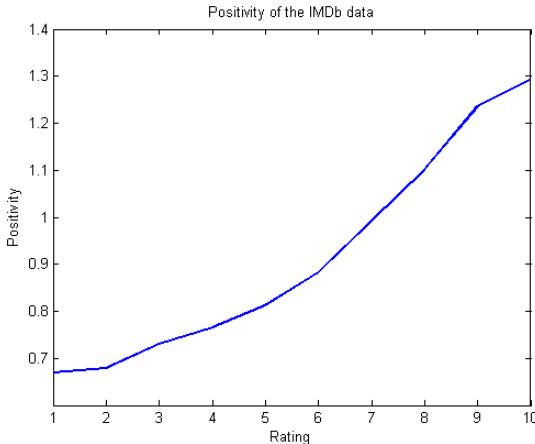


Figure 5: IMDb positivity

Our interest is in the increase in positivity in the highest categories: in IMDb this increase is relatively modest, while it is quite steep for Scope. To assess this difference, we compare the average increase in positivity per category both before and after a category of interest. For Scope, the category of interest is 4: the hypothesis is that reviewers would tend to resist giving ratings higher than 4, even in the face of very positive review text.

	Category of Interest	Rate Below	Rate Above	Ratio
Scope	4	.16	.28	.57
IMDb	8	.22	.19	1.15

Table 6: Positivity - Rate of Increase

This is indeed what we find: the rate of change per category above 4 nearly doubles from .16 to .28. We perform a similar analysis with the IMDb data, selecting 8 as the category of interest. Here we find a striking contrast: the rate of change actually drops above 8 (see Table 6).

This analysis strongly supports the third version of our hypothesis: the difference in positivity of U.S. and Scandinavian reviews reflects a difference in the relation of text positivity to rating, for very positive texts. For such texts, Danish reviewers, when compared to U.S. reviewers, have a tendency to “downgrade” a text of a given positivity.

6 Conclusion

There is a widely-held belief that Americans and Scandinavians differ in the way they express and rate positive sentiment. To our knowledge this paper represents the first attempt to test such a belief in a systematic way. Using large collections of film reviews, we have found strong confirmation of the hypothesized difference, defined from three different points of view: ratings, text, and text-rating relations.

In recent years, the use of rating systems have exploded, to the point where they are relied on every day for millions of decisions about everything from where to eat to what film to see, or where and how to take a vacation. The present work, while limited in Scope, suggests a potentially far-reaching conclusion; namely, it points to the possibility that there are systematic differences in rating systems, that we ignore at our peril. As we have seen, Danes differ sharply from Americans in the positivity of ratings and text: they give far fewer top ratings; and the frequency of highly positive terms in the top categories is quite a bit less. One natural conclusion is that there are cultural differences leading Danes to produce reviews and ratings in a rather different way than Americans. In our experience, those familiar with Danish and American culture find this quite plausible and readily suggest numerous potential explanations – perhaps the most compelling of which concerns the traditional grading system in Danish schools¹, where the top grade of “13” was given in only the most exceptional of circumstances, and was always far less frequent than the top grade of “A” in U.S. schools.

There is an obvious alternative explanation for these differences, namely, that Danes are simply less enthusiastic about the films they see. This might seem somewhat paradoxical – since Danes and Americans are both free to choose which films they see, one might expect that they are equally enthusiastic about the films they choose to see and review. However, it has often been suggested that the film industry in many European countries is subject to U.S. cultural imperialism, which would hold that, because of its economic and cultural power, the U.S. film industry is able to substantially alter the film-going options of the Danish public.

We don't discount the possibility that our data in part reflects a general lack of enthusiasm for the films on offer in Denmark, either due to U.S. cultural dominance or perhaps some other factors. This explanation would be rather uninteresting in terms of the general issues concerning the rating and expression of sentiment in different popula-

¹The Danish grading system was revised in 2006, in part to make it more in line with grading systems in other countries.(Wikipedia, 2012)

tions, although it ought to be of interest to the producers and distributors of film in Denmark. In any case, we are convinced this is not the complete explanation, because of our third finding, concerning the relation of ratings to text. This shows that there are systematic differences between Danes and Americans for texts expressing a similar level of positivity – Danes tend to move many of these from a top category to a less positive one. In our view this constitutes clear evidence of a systematic difference in how sentiment is treated in the two populations.

We have argued that these differences point to a potentially important problem with the use of rating systems, especially if such differences are widespread. In future work, we intend to examine reviews in other domains, to see if the difference we have found is limited to certain domains or is one that is generally found when comparing Danes and Americans. We are also exploring ways to address the problem these differences pose: one natural hypothesis is that, when there is a systematic mismatch between text and rating, the text positivity is a better guide to the true sentiment. We would like to see if an automatic sentiment analysis might reduce systematic mismatches in these cases.

Acknowledgements

Thanks to Bernardo Huberman for suggesting this line of work, and for lots of feedback and creative input. Chris Zimmerman also gave important feedback and suggestions.

References

- Noah Constant, Christopher Davis, Christopher Potts, and Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung*, 1(2):5–21.
- C. Dellarocas and R. Narayan. 2006. What motivates consumers to review a product online? a study of the product-specific antecedents of online movie reviews. In *In Proceedings of the International Conference on Web Information Systems Engineering*.
- FAQ. 2012. Yelp.com. <http://www.yelp.com/faq>.
- N. Hu, P. A. Pavlou, and J. Zhang. 2006. Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online

- word-of-mouth communication. In *In Proceedings of the ACM Conference on Electronic Commerce*.
- Wikipedia. 2012. Academic grading in Denmark — Wikipedia, the free encyclopedia. [Online; accessed 27-August-2012].
- Fang Wu and Bernardo Huberman. 2010. Opinion formation under costly expression. *ACM Transactions on Intelligent Systems and Technology*, 1(1).

KONVENTS 2012

**LThist 2012: First International Workshop on Language Technology
for Historical Text(s)**

September 21, 2012
Vienna, Austria

Preface

The interaction between Human Language Technology (HLT) and Digital Humanities (DH) at large has been of interest in various projects and initiatives during the last years, aiming to bring forward language resources and tools for the Humanities, Social Sciences and Cultural Heritage.

The specific focus of LThist 2012 lies on the development of technology and resources required for processing historical texts. Workshop contributors and participants discuss ways and strategies for shaping HLT resources (tools, data and metadata) in ways that are maximally beneficial for researchers in the Humanities. The necessity for a strong interplay between proponents from language technology and from the Humanities is also reflected in the invited talks. While Caroline Sporleder takes a language technology perspective, Sonia Horn addresses the needs and requirements from a medical historian's point of view. A major aspect of the workshop is the exchange of experiences with and comparison of tools, approaches, and standards that make historical texts accessible to automatic processing. Moreover, LThist encourages the interchange of historical data and processing tools.

In the present workshop, historical texts are understood in two ways: i) texts as documents of older forms of languages, and ii) texts as documentations of historical content. Accordingly, the contributions comprise a broad range of topics, genres and diachronic language varieties, including scientific prose, narratives, folk tales, riddles etc., as well as trade-related documents and marriage license books with the latter being a valuable resource for demography studies. The presented papers address various aspects of data preparation and (semi-)automatic processing for a number of languages including Old Swedish, Late Middle English, Middle English, Early Modern English and Modern English, diachronic varieties of German, Dutch and Spanish, and Old Occitan. The proposed approaches and technical solutions center around problem areas such as improving the OCR quality of historical texts, orthography harmonization and mapping historical to modern word forms, as prerequisites for automatic mining of historical texts. Also, the possibilities of cross-language transfer of morphosyntactic and syntactic annotation from resource-rich source languages to under-resourced target languages are examined. Technical infrastructures, specifically tailored for historical corpora, are discussed, including mark-up languages for historical texts and representation formats for diachronic lexical databases, processing tools and architectures.

Overall, LThist 2012 well reflects the current discussions regarding automatic processing of historical texts where OCR errors and the lack of harmonization in orthography are still major practical issues, but where also machine learning and cross-language transfer are coming more and more into focus.

Thierry Declerck, Brigitte Krenn and Karlheinz Mörtl
Workshop Organizers

Saarbrücken & Vienna, September 2012

Organisation Committee:

Thierry Declerck (DFKI GmbH, Saarbrücken, Germany & ICLTT-ÖAW, Vienna, Austria)
Brigitte Krenn (ÖFAI, Vienna, Austria)
Karlheinz Mörtz (ICLTT-ÖAW, Vienna, Austria)

Program Committee:

Ulrike Czeitschner (Austrian Academy of Sciences)
Thierry Declerck (DFKI GmbH & Austrian Academy of Sciences)
Stefanie Dipper (University of Bochum)
Martin Durrell (University of Manchester)
Anette Frank (University of Heidelberg)
Jost Gippert (University of Frankfurt)
Vangelis Karkaletsis (NCSR "Demokritos")
Brigitte Krenn (ÖFAI, Vienna)
Anastasia Krithara (NCSR "Demokritos")
Piroska Lendvai (Hungarian Academy of Sciences)
Anke Lüdeling (Humboldt University, Berlin)
Karlheinz Mörtz (Austrian Academy of Sciences)
Hélène Pignot (Université Paris I)
Michael Piotrowski (University of Zürich)
Claudia Resch (Austrian Academy of Sciences)
Silke Scheible (University of Stuttgart)
Eiríkur Rögnvaldsson (University of Iceland)
Antal van den Bosch (Radbout University Nijmegen)
Martin Volk (University of Zürich)
Kalliopi Zervanou (University of Tilburg)

Invited Speakers:

Caroline Sporleder (Saarland University)
Sonia Horn (Medical University of Vienna)

Sponsors:

We are very grateful to the AMICUS (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts) Network for funding the invited talk by Caroline Sporleder: <http://ilk.uvt.nl/amicus>

AMICUS

Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts



And to the Project ABaC:us (Austrian Baroque Corpus):
<http://corpus3.aac.ac.at/showcase/index.php/emg>



Workshop Program

Friday, September 21, 2012

09:00–09:10 Welcome and opening

09:10–09:45 First invited talk: Caroline Sporleder

09:45–10:30 First session: Normalisation

Rule-based normalization of historical text – A diachronic study

Eva Pettersson, Beáta Megyesi and Joakim Nivre

Manual and semi-automatic normalization of historical spelling – Case studies from Early New High German

Marcel Bollmann, Stefanie Dipper, Julia Krasselt and Florian Petran

10:30–11:30 Coffee/tea break and poster/demo session

The Sketch Engine as infrastructure for historical corpora

Adam Kilgarriff, Milos Husak and Robyn Woodrow

Evaluating a post-editing approach for handwriting transcription

Verónica Romero, Joan Andreu Sánchez, Nicolás Serrano and Enrique Vidal

bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa ... Towards lexical link-up for a corpus of Old Swedish

Yvonne Adesam, Malin Ahlberg and Gerlof Bouma

Semantic change and the distribution in lexical sets

Karin Cavallin

Automatic classification of folk narrative genres

Dong Nguyen, Dolf Trieschnigg, Theo Meder and Mariët Theune

11:30–12:15 Second session: Textual and linguistic annotation

The DTA ‘base format’: A TEI-subset for the compilation of interoperable corpora

Alexander Geyken, Susanne Haaf and Frank Wiegand

Building an old Occitan corpus via cross-language transfer

Olga Scrivner and Sandra Kübler

12:15–13:45 Lunch break

13:45–14:30 Second invited talk: Sonia Horn

14:30–15:15 Third session: OCR

Digitised historical text: Does it have to be mediOCRe?

Bea Alex, Claire Grover, Ewan Klein and Richard Tobin

Comparison of named entity extraction tools for raw OCR text

Kepa Joseba Rodriguez, Michael Bryant, Tobias Blanke and Magdalena Luszczynska

15:15–15:45 Tea/coffee break

15:45–16:30 Fourth session: Middle English

From semi-automatic to automatic affix extraction in Middle English corpora:

Building a sustainable database for analyzing derivational morphology over time

Hagen Peukert

The reference corpus of Late Middle English scientific prose

Javier Calle-Martín, Laura Esteban-Segura, Teresa Marqués-Aguado and Antonio Miranda-García

16:30–17:15 Panel discussion: Lexicon (Chair: Brigitte Krenn)

Panelists: Karlheinz Mörtz, Hélène Pignot, Adam Kilgariff, Yvonne Adesam and Thierry Declerck

17:15–17:30 Creative break

17:30–17:45 Networking

17:45–18:00 Closing session

Rule-Based Normalisation of Historical Text – a Diachronic Study

Eva Pettersson[†], Beáta Megyesi and Joakim Nivre

Department of Linguistics and Philology

Uppsala University

[†]Swedish National Graduate School

of Language Technology

firstname.lastname@lingfil.uu.se

Abstract

Language technology tools can be very useful for making information concealed in historical documents more easily accessible to historians, linguists and other researchers in humanities. For many languages, there is however a lack of linguistically annotated historical data that could be used for training NLP tools adapted to historical text. One way of avoiding the data sparseness problem in this context is to normalise the input text to a more modern spelling, before applying NLP tools trained on contemporary corpora. In this paper, we explore the impact of a set of hand-crafted normalisation rules on Swedish texts ranging from 1527 to 1812. Normalisation accuracy as well as tagging and parsing performance are evaluated. We show that, even though the rules were generated on the basis of one 17th century text sample, the rules are applicable to all texts, regardless of time period and text genre. This clearly indicates that spelling correction is a useful strategy for applying contemporary NLP tools to historical text.

1 Introduction

Digitalised historical text is a rich source of information for researchers in humanities. However, there is a lack of language technology tools adapted to old texts, that could make the information more easily accessible. As an example, the work presented in this paper has been carried out in close cooperation with historians in the *Gender and Work* project (Ågren et al., 2011). This project aims at exploring how men

and women supported themselves in the Early Modern Swedish period (1550–1800). Information on this is manually searched for in documents from this period, and stored in a database. The project has shown that work activities are most often described as a verb and its complements. Automatically extracting all the verbs in a text would be a rather trivial task for contemporary text, using standard NLP tokenisation and tagging. However, for historians and other researchers in humanities, the manual way of searching for information is still often the only alternative.

Developing NLP tools for historical text is a great challenge, since old texts vary greatly in both spelling and grammar between different authors, genres and time periods, and even within the same text, due to the lack of spelling conventions. The texts that we have studied are generally written in a spoken language fashion, with less distinct boundaries between sentences, and a spelling that to a larger extent reflects the phonetic form of the word. Sentences with 50–100 words or more are not unusual. The Swedish language was also strongly influenced by other languages at this time. Evidence of this is the placement of the finite verb at the end of subordinate clauses in a German-like style not usually found in modern Swedish texts.

Another problem in developing language technology tools specifically aimed at handling historical text, is the shortage of linguistically annotated historical data. One way of getting around this problem is to use existing NLP tools trained for handling contemporary text, and normalise the historical input text to a more modern spelling be-

fore applying the tools. Pettersson et al. (2012) showed that a relatively small set of hand-crafted normalisation rules had a large positive impact on the results of tagging and parsing of historical text using contemporary NLP tools. We are interested in how well this method works for texts from different time periods and genres. In this paper, we present a case study, where we evaluate the impact of these normalisation rules on 15 different Swedish texts ranging from 1527 to 1812, within the genres of court records versus church documents. Our hypothesis is that the normalisation rules will work the best for texts from the same time period as the text sample used for developing the normalisation rules, and within the same text genre, i.e. court records from the 17th century. However, we hope that the rules will be of use also for both older and younger texts, as well as for other text types. Furthermore, we assume that error reduction as regards normalisation will be larger for older texts, since these differ from contemporary language to a larger extent, whereas the overall normalisation accuracy probably will be higher for younger texts.

From the evaluation we see that the rules are indeed applicable to all texts in the corpus, substantially increasing the proportion of normalised words in the text, as well as improving tagging and parsing results. Part of the explanation for this is that the same types of spelling changes occur in texts from all time periods studied, only with a slightly different frequency distribution.

2 Related Work

Sánchez-Marco et al. (2011) tried a strategy for adapting an existing NLP tool to deal with Old Spanish. Adaptation was performed on the basis of a 20 million token corpus of texts from the 12th to the 16th century, and included expansion of the dictionary, modification of tokenisation and affixation rules, and retraining of the tagger. The dictionary expansion had the highest impact on the results, and was performed by automatically generating word forms through mapping old spelling variants to their contemporary counterparts. The tagger was then retrained based on a gold standard of 30,000 tokens, where the tokens were first pre-annotated with the contemporary tagger, and then manually corrected. The evaluation of the

final tagger showed an accuracy of 95% in finding the right part of speech, and 90% accuracy in finding the complete morphological tag.

Bollmann et al. (2011) tried a rule-based approach to normalisation of texts from Early New High German (14th to 16th century). In this approach, rules were automatically derived from word-aligned parallel corpora, i.e. the 1545 edition of the Martin Luther Bible aligned with a modern version of the same Bible. Since Bible text is rather conservative in spelling and terminology, approximately 65% of the words in the old Bible version already had an identical spelling to the one occurring in the modern version. To cope with non-identical word forms, normalisation rules were generated from the word alignment results, by means of Levenshtein edit distance, and the rules were sorted by frequency of occurrence. In order not to generate non-words, only word forms that could be found in the modern version of the Bible were accepted. Other word forms produced by the rules were discarded, leaving the old spelling preserved. This method proved successful, increasing the proportion of words with a modern spelling from 65% to 91%.

Oravecz et al. (2010) included a standardisation/normalisation step in their work on semi-automatically annotating a corpus of Old Hungarian. Normalisation was performed using a noisy channel model combined with morphological analysis filtering and decision tree reranking. Combining these methods, they reached a normalisation precision of 73%.

Pettersson et al. (2012) used a relatively small set of hand-crafted normalisation rules for improving tagging and parsing of historical text. The aim of this study was to automatically extract verbs and their complements from Early Modern Swedish texts (1550–1800). A part-of-speech tagger trained on contemporary Swedish text was used for verb identification, whereas verbal complements were extracted based on output from a dependency parser. Spelling variation in the input text was handled by a set of 29 hand-crafted rules, produced on the basis of a sample text from the 17th century. The method was evaluated on a corpus of 400 sentences extracted from the period 1527–1737. For verb extraction based on tagging, recall increased from 60% to 76% when normal-

isation rules were applied to the input text before running the tagger. Likewise, the proportion of fully or partially extracted verbal complements based on parsing increased from 53% to 56%.

3 Normalisation Method

In our experiments, we will use the setup described in Pettersson et al. (2012), where a set of 29 hand-crafted normalisation rules are used for transforming the spelling into a more modern version. The normalisation rules were developed on the basis of two sources:

1. Spelling changes documented in the context of the reformed Swedish spelling introduced in 1906 (Bergman, 1995).
2. A text sample from *Per Larssons dombok*, a selection of court records from 1638 (Edling, 1937).

Since 1906, no larger spelling changes have been employed in the Swedish language. With this reform, the spelling for the *t* sound was simplified from *dt* to a single *t*, as in *varidt*→*varit* ("been"). Likewise for the *v* sound, the superfluous letters *h* or *f* were dropped, as in *hvar*→*var* ("was") and *skrifva*→*skriva* ("write"). Also, the use of an *f* for denoting a *v* sound was abandoned, turning for example *af* into *av* ("of/off").

As for the empirical rules based on the 17th century text sample, these include:

- substitution of letters to a phonologically similar variant, such as *q*→*k* in *qvarn*→*kvarn* ("mill") and *z*→*s* in *slogz*→*slogs* ("were fighting")
- deletion of repeated vowels, as in *saak*→*sak* ("thing")
- deletion of mute letters, such as *j* in *vijka*→*vika* ("fold"), and *b* in *dömbdes*→*dömde* ("was sentenced")
- normalisation of spelling influenced by other languages, mainly German, as in *schall*→*skall* ("shall")

The current normalisation scheme has no mechanism for handling word forms that are no longer in use. These will be treated like any other word by the normalisation algorithm.

4 Hypothesis

Our initial hypothesis is that the set of normalisation rules described in section 3, will work better for 17th century court records than for texts from other centuries and within other genres, since most of the rules have been developed on the basis of a text sample from 1638. Since there were no standardised spelling conventions during this period, we also believe that there will be differences in the applicability of the rules between texts from different authors and within different text genres.

A closer look at the gold standard however shows that the most frequent types of spelling changes between the old and the modern text are more or less the same, regardless of time period and text type. Table 1 shows the 10 most frequent changes (at a character level) between the raw historical text and the manually normalised gold standard. In this table, a minus sign means that a letter needs to be removed to transform the spelling into the modern spelling given in the gold standard. For example *-h* is performed for transforming *åhr* into *år* ("year"). In the same way, a plus sign denotes that a letter is inserted to create the modern spelling. For example, *+dbl* means that a letter is duplicated, as when transforming the spelling *alt* into the contemporary spelling *allt* ("everything"). The *"/* sign means that a letter needs to be replaced by another letter, most commonly a phonologically close letter. This can be illustrated by the change *w/v*, transforming for example *beswär* into *besvär* ("trouble").

As seen in the table, the most frequently occurring changes are present in texts from all centuries represented in the corpus, and in both court records and church documents. The frequency distribution differs only slightly, so that for example the deletion of *-h* is the most common transformation seen in texts from the 16th and 17th century, whereas this deletion is only the second most frequent in 18th century text. This information makes us believe that the normalisation rules may have a significant impact on all texts, not only 17th century court records as stated in our initial hypothesis.

16 cent	17 cent	18 cent	Court	Church
-h	-h	w/v	-h	-h
w/v	-dbl	-h	-dbl	w/v
e/ä	w/v	+dbl	w/v	-dbl
-f	-f	-e	-e	e/a
-dbl	-e	-f	-f	-f
e/a	e/a	e/ä	+dbl	e/ä
-e	+dbl	f/v	e/ä	-e
+dbl	e/ä	e/a	e/a	+dbl
u/v	f/v	-dbl	-i	c/k
c/k	i/j	c/k	f/v	i/j

Table 1: Top 10 changes at a character level, for transforming raw historical text into the manually normalised gold standard spelling. 16 cent = 16th century texts (1527–1593). 17 cent = 17th century texts (1602–1689). 18 cent = 18th century text (1728–1812). Court = Court records. Church = Church documents. Dbl = Double letter.

5 Experimental Setup

The normalisation rules described in section 3, and their impact on tagging and parsing results, are to be evaluated taking different text types and time periods into account. The corpus used for evaluation consists of 15 texts, ranging from 1527 to 1812. The text types covered are court records and church documents. 10 out of these 15 texts are the same as the ones used for evaluation by Pettersson et al. (2012). In total, there are 787,122 tokens in the corpus. From this corpus, a gold standard has been extracted, comprising 40 randomly selected sentences from each text, i.e. in total 600 sentences. The proportion of tokens in each text as a whole, and in the gold standard part of the text, is illustrated in table 2.¹

For each text in the gold standard corpus, evaluation includes 1) the proportion of correctly normalised tokens, and the error reduction achieved through normalisation, 2) the accuracy of verb identification (tagging), before and after normalisation, and 3) the accuracy of verbal complement extraction (parsing), before and after normalisation. Error reduction has been calculated by the following formula:

¹The number of tokens differs slightly from those presented in Pettersson et al. (2012), since the current numbers are calculated on tokenised text, whereas the previous numbers were calculated on the "raw" source text.

Court Records			
Name	Year	Total	Sample
Östra Härad	1602–1605	38,477	2,069
Vendel	1615–1645	64,977	2,509
Per Larsson	1638	12,864	2,987
Hammerdal	1649–1686	75,143	1,859
Revsund	1649–1689	113,395	2,328
Stora Malm	1728–1741	458,548	1,895
Vendel	1736–1737	61,664	3,450
Stora Malm	1742–1760	74,487	2,336
Stora Malm	1761–1783	66,236	1,825
Stora Malm	1784–1795	58,738	1,378
Stora Malm	1796–1812	47,671	1,683
Church Documents			
Name	Year	Total	Sample
Västerås	1527	14,149	3,709
Kyrkoordning	1571	60,354	2,246
Uppsala Möte	1593	34,877	1,184
Kyrkolag	1686	35,201	2,086
Total	1527–1812	787,122	33,544

Table 2: Corpus distribution, given in number of tokens in the documents. Total = Number of tokens in the whole corpus. Sample = Number of tokens in the gold standard sample.

$$\frac{\text{CorrectAfterNormalisation} - \text{CorrectBeforeNormalisation}}{\text{IncorrectBeforeNormalisation}}$$

where *CorrectAfterNormalisation* is the percentage of tokens with an identical spelling to the modern version *after* the normalisation rules have been applied, *CorrectBeforeNormalisation* is the percentage of tokens with an identical spelling to the modern version *before* the normalisation rules have been applied, and *IncorrectBeforeNormalisation* is the percentage of tokens differing in spelling from the modern version before the normalisation rules have been applied.

To be able to evaluate the proportion of correctly normalised tokens, each token in the gold standard part of the corpus was manually assigned its modern spelling equivalent. For tagging and parsing, there is currently no gold standard available for the texts used in the gold standard corpus. However, since the overall aim of the research project (*Gender and Work*, described in section 1) is to extract verbs and their complements from historical text, all the verbs and their comple-

ments in the gold standard have been manually annotated. Therefore we may indirectly evaluate tagging accuracy by comparing the verb tags produced by the tagger to the words marked as verbs in the gold standard. Likewise, parsing accuracy may be evaluated by comparing the verbal complements assigned by the parser to the verbal complements given in the gold standard.

6 Results and Discussion

6.1 Normalisation

Table 3 presents the proportion of tokens in the historical texts with the same spelling as in the manually modernised gold standard, before and after normalisation.² This table also shows the error reduction for each test text, i.e. the percentage of correctly normalised tokens that were not originally identical to the modern spelling (see further section 5 for a definition of the calculation of error reduction).

Since the normalisation rules are based on a text sample from 17th century court records, it could be expected that the rules would not be applicable for other time periods and text types. However, the results presented in table 3 show that this set of normalisation rules has a positive effect on all texts. In fact, the largest error reduction for previously unseen text, 30.7% as compared to the average 21.5%, is achieved for the church text from 1593, where the proportion of tokens with a correct modern spelling increases from 47.8% to 63.8% after normalisation. This shows that the normalisation rules are successful for texts from other centuries and genres compared to the text used for developing the rules.

This table also shows the results for part of the text *Per Larssons dombok* (1638). As noted earlier, a sample from this text was used for developing the empirically based normalisation rules. Even though a disjoint sample is used for evaluation, we see a larger error reduction for this text than for all the other texts (38.6% as compared to the average 21.5%). This indicates that the normalisation rules are somewhat biased towards the text on which the rules were based, even though

²It should be noted that the part of *Per Larssons dombok* that is used for evaluation, is disjoint from the sample used for developing the normalisation rules.

Development Text (Court Records)			
Text	Orig.	Norm.	ErrRed.
1638 [†]	56.7%	73.4%	38.6%
Court Records			
Text	Orig.	Norm.	ErrRed.
1602–1605	64.6%	69.8%	14.7%
1615–1645	62.4%	71.8%	25.0%
1649–1686	64.2%	74.9%	29.9%
1649–1689	67.3%	74.7%	22.6%
1728–1741	69.2%	72.8%	11.7%
1736–1737	73.7%	78.8%	20.5%
1742–1760	67.9%	74.4%	20.2%
1761–1783	74.5%	77.1%	10.2%
1784–1795	80.8%	83.1%	19.2%
1796–1812	78.8%	81.2%	11.3%
Church Documents			
Text	Orig.	Norm.	ErrRed.
1527	53.5%	64.4%	23.4%
1571	51.9%	63.9%	24.9%
1593	47.8%	63.8%	30.7%
1686	64.7%	71.5%	19.3%
Average			
1527–1812	65.2%	73.0%	21.5%

Table 3: Normalisation results, given in percentage of tokens with the same spelling as in the gold standard, before and after normalisation. Orig = Proportion of words in the original text that are identical to the modern spelling. Norm = Proportion of words in the normalised text that are identical to the modern spelling. ErrRed = Error reduction. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a † sign.

the results are very promising for the other texts as well.

Naturally, the normalisation rules have a larger impact on the oldest texts, since these differ from the modern spelling to a larger extent, meaning that there are more words that would need normalisation. Hence, the smallest error reduction (11.3%) is observed for the youngest text (*Stora Malm*, 1796–1812), where the proportion of correctly normalised tokens increases from the original 78.8% to 81.2%. Averaging over the numbers in table 3, we see that 16th century text (1527–1593) yields an error reduction of 26.3% as compared to 22.3% for 17th century text (1602–1689)

and 18.6% for 18th century text (1728–1812). Still, the overall percentage of tokens in the normalised text that are identical to the gold standard spelling is generally higher for younger texts.

Interestingly, the effect of the normalisation rules seems not to be dependent only on time period and/or the proportion of original tokens that need normalisation. For example, the court records text from 1602–1605, the court records text from 1649–1686 and the church text from 1686, all have a proportion of 64–65% of words that do not need normalisation. However, for the 1649–1686 text, error reduction is twice as high (29.9%) as for the older 1602–1605 text (14.7%), and also much higher than for the church text from 1686 (19.3%). This could indicate that for texts with equal prerequisites, the normalisation rules work better for texts that are close in time and genre to the text used for developing the normalisation rules.

6.2 Tagging and Verb Extraction

The main goal of the normalisation process is to improve accuracy for language technology tools applied after normalisation, i.e. tagging and parsing. As argued in section 5, there is currently no gold standard for evaluating the automatically tagged test texts. However, all the verbs in the gold standard have been manually assigned a verb label, and we may indirectly evaluate tagging accuracy by comparing the verb tags produced by the tagger, to the words marked as verbs in the gold standard. The tagger used for this purpose is the HunPOS tagger (Halácsy et al., 2007), a free and open source reimplementation of the HMM-based TnT-tagger by Brants (2000). The tagger is used with a pre-trained language model based on the Stockholm-Umeå Corpus (SUC), a balanced, manually annotated corpus of different text types representative of the Swedish language in the 1990s, comprising approximately one million tokens (Gustafson-Capkova and Hartmann, 2006).

The precision and recall measures for the verb extraction comparison are presented in table 4. The results are also compared to the results of verb identification for contemporary Swedish text from the Stockholm-Umeå Corpus. Since the tagger used in the experiments on historical texts is

trained on the whole of SUC, we trained a new model for the tagger in order not to evaluate on the same data as the tagger has been trained. The tagging model used for annotating contemporary text hence includes all tokens in SUC except for the tokens reserved for evaluation.

Development Text (Court Records)				
	Orig.		Norm.	
Text	Prec.	Rec.	Prec.	Rec.
1638 [†]	68.8%	51.8%	82.3%	85.5%
Court Records				
	Orig.		Norm.	
Text	Prec.	Rec.	Prec.	Rec.
1602–1605	72.5%	61.2%	73.1%	76.3%
1615–1645	78.8%	53.9%	78.9%	68.7%
1649–1686	74.5%	65.6%	85.2%	76.7%
1649–1689	77.2%	62.1%	83.7%	82.0%
1728–1741	84.8%	74.0%	85.1%	83.6%
1736–1737	82.1%	68.3%	85.9%	73.7%
1742–1760	85.3%	76.7%	86.0%	85.5%
1761–1783	81.7%	79.3%	85.0%	83.7%
1784–1795	90.6%	88.0%	90.4%	90.4%
1796–1812	84.4%	83.3%	85.1%	87.8%
Church Documents				
	Orig.		Norm.	
Text	Prec.	Rec.	Prec.	Rec.
1527	70.4%	51.7%	67.4%	70.0%
1571	72.7%	66.8%	73.2%	78.1%
1593	63.8%	45.8%	66.1%	68.4%
1686	86.7%	66.7%	84.4%	71.3%
Average				
	Orig.		Norm.	
1527–1812	78.3%	66.3%	80.8%	78.8%
Contemporary Text (SUC)				
	Orig.		Norm.	
1990s	99.1%	99.1%	–	–

Table 4: Verb extraction results after normalisation. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a [†] sign.

From the verb extraction results, it is noticeable that especially recall improves to a great extent for all texts. On average, recall improves from 66.3% for raw input text to 78.8% for the normalised version of the text. This is still not very close to the 99.1% recall noted for verb identifi-

cation in the contemporary SUC sample, but high enough to be useful for our purposes. The best results are achieved for the text from 1784–1795, with a 90.4% recall.

As expected, the lowest results are generally observed for the oldest texts, even though there are exceptions. For example the church document from 1571 has a recall of 78.1%, whereas the substantially younger court document from 1736–1737 yields the lower score of 73.7%. This is remarkable also in the sense that the 1571 text has a lower recall before normalisation (66.8%) than the younger text (68.3%). One reason could be that the 1571 text is a little closer in time to the text used for generating the normalisation rules (from 1638).

In most cases, precision also improves slightly with normalisation. For the 1527 text however, precision drops from 70.4% to 67.4% when adding the normalisation rules. The leap in recall from 51.7% to 70% is however worth the slight decrease in precision. On average, precision improves from 78.3% to 80.8%.

6.3 Parsing and Complement Extraction

As explained in section 5, there is currently no gold standard for evaluating the automatically parsed test texts. However, verbal complements in the gold standard have been marked up manually. Parsing accuracy may thus be indirectly evaluated by comparing the complements assigned by the parser to the complements given in the gold standard. For this purpose, we use a string-based evaluation method first described in Pettersson et al. (2012), where all labels and brackets are removed before comparing the segments extracted from the parser to the segments extracted in the gold standard. Each extracted instance is classified as falling into one of four mutually exclusive categories:

- Fully correct complement set
- Partially correct complement set
- Incorrect complement set
- Missing complement set

A complement set is regarded as fully correct if the output string generated by the system is identical to the corresponding gold standard string.

Furthermore, a complement set is regarded as partially correct if the output string generated by the system has a non-empty overlap with the corresponding gold standard string. A (non-empty) complement set is regarded as incorrect if the output string has no overlap with the gold standard string. Finally, a complement set is regarded as missing if the output string is empty but the gold standard string is not.

Complement extraction results are shown in table 5, where for each text, the results for the "raw" input text is given on top, and the results after normalisation is given at the bottom. A striking difference is that with normalisation, even though the number of fully or partially correctly extracted complements stay more or less the same, the number of incorrectly assigned complements in general drops significantly. On average, the number of incorrectly assigned complements drops from 27.4% to 22.7%. Accordingly, the proportion of verbs that are not assigned any complements at all even though the gold standard states that there should be complements, increases to a similar extent, from 23.2% o 26.2%.

Even though the proportion of correctly extracted complements do not increase much, one should bear in mind that these numbers are calculated based on the words that have been correctly identified as verbs by the tagger, meaning that the actual number of extracted complements has increased by normalisation. It is also interesting to note that for all historical texts, the proportion of fully extracted complements is higher than for the contemporary text (38.1% vs 30.3%). This may be partly explained by different annotation conventions, since the historical gold standard corpus was annotated by us, whereas the contemporary gold standard was annotated by other people. Due to this, the results may not be directly comparable, but are still an indication of parsing performance for historical versus contemporary text. It is also worth mentioning that spelling correction captures differences in surface form at a word level, improving for example tagging based on dictionaries. Grammatical differences however include changes in word order and sentence structure, i.e. differences that need to be handled by some kind of retraining of the parser.

Development Text (Court Records)				
Text	Fully	Part.	Inc.	Unass.
1638 [†]	32.3%	9.9%	35.4%	22.4%
	34.0%	10.3%	25.3%	30.3%
Court Records				
Text	Fully	Part.	Inc.	Unass.
1602–1605	41.6%	14.6%	24.8%	19.0%
	41.8%	12.9%	21.8%	23.5%
1615–1645	39.7%	12.4%	35.9%	12.0%
	36.7%	14.2%	30.7%	18.4%
1649–1686	34.3%	16.9%	27.9%	20.9%
	33.3%	13.4%	26.4%	26.9%
1649–1689	36.8%	14.2%	26.3%	22.6%
	41.4%	11.6%	24.7%	22.3%
1728–1741	34.6%	16.0%	24.7%	24.7%
	34.4%	15.3%	23.0%	27.3%
1736–1737	43.5%	16.1%	21.7%	18.7%
	45.8%	14.6%	18.3%	21.4%
1742–1760	40.7%	14.8%	24.7%	19.8%
	42.1%	12.5%	18.5%	26.9%
1761–1783	34.2%	18.0%	22.4%	25.5%
	33.5%	19.4%	17.1%	30.0%
1784–1795	39.9%	18.6%	16.4%	25.1%
	40.4%	17.0%	16.0%	26.6%
1796–1812	41.3%	17.9%	19.0%	21.7%
	40.7%	19.6%	19.6%	20.1%
Church Documents				
Text	Fully	Part.	Inc.	Unass.
1527	37.3%	11.8%	33.8%	17.1%
	36.0%	11.3%	33.4%	19.3%
1571	30.1%	6.1%	34.9%	28.8%
	39.2%	9.0%	23.1%	28.7%
1593	37.0%	11.1%	25.9%	39.5%
	39.7%	5.0%	18.2%	37.2%
1686	31.8%	10.4%	27.9%	29.9%
	32.6%	9.8%	24.2%	33.5%
Average				
1527-1812	37.0%	13.9%	27.4%	23.2%
	38.1	13.1%	22.7%	26.2%
Contemporary Text (SUC)				
SUC	30.3%	54.2%	9.1%	6.4%

Table 5: Complement extraction results after normalisation. Results for the "raw" input text is given on top, and results after normalisation is given at the bottom. Fully = Fully identical match. Part = Partial match. Inc = Incorrect. Unass = Unassigned. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a [†] sign.

7 Conclusion

In this paper, we have presented a rule-based approach to normalisation of historical text, with the aim of making NLP tools developed for modern language applicable for analysing historical text. We have shown that a relatively small set of hand-crafted normalisation rules based on one single text, had a large positive impact on the usefulness of contemporary NLP tools also for texts from other time periods and genres than the text from which the rules were developed. Our results thus show that existing NLP tools can be successfully used for analysing historical text, even for languages without access to a large amount of linguistically annotated historical data for training the tools. The fact that rules generated from one single text proved useful for other time periods and text types as well, means that we do not even need a large, balanced corpus as a basis for rule development. The choice of text used for developing the rules may however be important for success, since older texts generally contain a higher number of instances of differently spelled words. If we choose a too modern text, the rules generated may not be very useful for older texts. For the texts studied in this paper however, we saw that the same spelling variation occurs in texts from all centuries and genres (16th, 17th and 18th century court records and Church documents), only with a slightly different frequency distribution. It would be interesting to do the same kind of evaluation for other time periods and text types as well. Furthermore, we would like to evaluate the general applicability of the method by testing the same approach on texts from various time periods and in other languages. This would of course require a different set of normalisation rules, but the general method for rule development could still be the same.

Even though the presented normalisation method works well for a range of documents, it is a time-consuming and knowledge-intense manual work to formulate the normalisation rules. Future work includes more sophisticated methods for normalisation, including automatic rule generation. It would also be interesting to explore methods for improving parsing performance, where normalisation alone may not be a sufficient ap-

proach.

References

- Maria Ågren, Rosemarie Fiebranz, Erik Lindberg, and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review*, 59(3):271–291. Forthcoming.
- Gösta Bergman. 1995. *Kortfattad svensk språkhistoria*. Prisma Magnum, Stockholm, 5th edition.
- Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42, Hissar, Bulgaria, September.
- Thorsten Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, Seattle, Washington, USA.
- Nils Edling. 1937. *Uppländska domböcker*. Almqvist & Wiksell.
- Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0. Technical report, December.
- Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, Prague, Czech Republic.
- Csaba Oravecz, Bálint Sass, and Eszter Simon. 2010. Semi-automatic normalization of Old Hungarian codices. In *Proceedings of the ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 55–59, Faculty of Science, University of Lisbon Lisbon, Portugal, August.
- Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Parsing the Past - Identification of Verb Constructions in Historical Text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 65–74, Avignon, France, April. Association for Computational Linguistics.
- Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, OR, USA, June. Association for Computational Linguistics.

Manual and Semi-automatic Normalization of Historical Spelling — Case Studies from Early New High German

Marcel Bollmann and Stefanie Dipper and Julia Krasselt and Florian Petran

Department of Linguistics

Ruhr University Bochum

44780 Bochum, Germany

bollmann, dipper, krasselt, petran@linguistics.rub.de

Abstract

This paper presents work on manual and semi-automatic normalization of historical language data. We first address the guidelines that we use for mapping historical to modern word forms. The guidelines distinguish between *normalization* (preferring forms close to the original) and *modernization* (preferring forms close to modern language). Average inter-annotator agreement is 88.38% on a set of data from Early New High German. We then present *Norma*, a semi-automatic normalization tool. It integrates different modules (lexicon lookup, rewrite rules) for normalizing words in an interactive way. The tool dynamically updates the set of rule entries, given new input. Depending on the text and training settings, normalizing 1,000 tokens results in overall accuracies of 61.78–79.65% (baseline: 24.76–59.53%).

1 Introduction¹

It is well-known that automatic analysis of historical language data is massively hindered by the fact that such data shows large variance with regard to spelling. Characters and symbols used by the writer of some manuscript reflect impacts as different as dialect influences or spatial constraints. This often leads to inconsistent spellings, even within one text written up by one writer.

In this paper, we present guidelines for manual normalization, which define mappings from historical spellings to modern equivalents. Differ-

ences between historical and modern word forms mainly concern graphemic or dialect-specific phonetic/phonological divergencies—which are straightforward to map. The differences further include inflectional and semantic divergencies. In such cases, it is less clear what the goal of normalization should be: We can either stay close to the original and, e.g., keep historical inflection even if it violates modern morpho-syntactic constraints; we call this approach *normalization*. Or else, we can adjust inflection to modern constraints, which we call *modernization*. Of course, (close) normalization is much easier to generate automatically. However, further processing of the data is usually done by taggers and parsers that have been trained on modern data. Hence, data that is maximally similar to modern data would be preferred.

Rather than opting for one of the two forms, we argue for guidelines that serve both camps by providing two levels of normalization.

As already mentioned, historical spelling depends to a large extent on the dialect of the author or printer (or the assumed audience). As a consequence, spelling of historical texts can differ considerably between texts, too. We therefore think that normalization systems should be easily adaptable to specific texts. Our tool *Norma*, presented in this paper, implements such a system, in the form of a semi-automatic normalization tool.

The paper is organized as follows. Sec. 2 addresses related work, Sec. 3 describes the corpora that our studies are based on. Sec. 4 presents the guidelines for manual normalization. In Sec. 5, the tool *Norma* is introduced, followed by an evaluation in Sec. 6. Sec. 7 presents the conclusion.

¹The research reported here was financed by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4-1.

2 Related Work

There has been considerable work on historical corpora for some time (e.g., the Penn Corpora of Historical English, the Perseus and TITUS corpora, or ARCHER²), and increasingly so in the past few years, with the advent of historical corpora for many languages (such as Dutch or Portuguese). Still, guidelines for normalization of spelling variants are often not an issue—e.g., because the corpora are based on editions that standardized spelling to a sufficient extent—or they are not (or not yet) published. This leads to unnecessary duplication of work.

The guidelines developed in the GerManC project (Scheible et al., 2011; GerManC Project, 2012) provide a modern lemma form, using the spelling of the modern dictionary *Duden* or, for obsolete words, the leading forms in *Deutsches Wörterbuch* by Grimm, a historical dictionary. An inflected normalized form is created by attaching suitable modern inflection to the lemma. This form corresponds closely to the modernized form as defined in our guidelines (see Sec. 4.2).

An interactive tool that aids the normalization process is VARD (Baron and Rayson, 2008). It uses a lexicon to detect spelling variants and tries to find modern cognates by a combination of user-defined replacement rules, phonetic matching, and Levenshtein distance. Similarly to the *Norma* tool described in Sec. 5, VARD can be trained by the user confirming or correcting the suggested word forms. However, it is less flexible in that the normalization methods are mostly fixed; e.g., phonetic matching is hard-coded for (Early Modern) English and therefore not suited for German texts, but cannot be turned off or modified. Also, different normalization methods cannot be added.

3 Corpora

To evaluate both applicability of our guidelines and performance of the interactive tool, we created a corpus containing different types

²Penn Corpora: <http://www.ling.upenn.edu/histcorpora>; Perseus: <http://www.perseus.tufts.edu>; TITUS: <http://titus.uni-frankfurt.de>; ARCHER: <http://www.llc.manchester.ac.uk/research/projects/archer>.

of texts: they are written in different dialects, are manuscripts or prints, and from different domains. One part of the texts are fragments of the *Anselm Corpus* (Sec. 3.1), another part comes from the *LAKS Corpus* (Sec. 3.2).

3.1 Anselm Corpus

The *Anselm Corpus* consists of all German manuscripts and prints of the text “Interrogatio Sancti Anselmi de Passione Domini” (‘Questions by Saint Anselm about the Lord’s Passion’). In the 14th–16th centuries, this text was written down in various German dialects (from Upper, Central, and Low German) and transformed into long and short prose and lyric versions. In total, there are more than 50 German manuscripts and prints, which makes the text an exceptionally broadly-documented resource. The texts are transcribed in the context of an interdisciplinary project.³ The transcriptions are diplomatic, i.e., they stay maximally close to the original.

For our study, we selected fragments of 1,000 tokens of four manuscripts and two prints from different dialectal regions, and a 4,500-token fragment of a manuscript. The vast majority of the texts are written in Upper German. There are only two Anselm texts in Central German in our collection, and one of them (COL_p) in fact shows many characteristics of Low German. All texts are from the 15th century, i.e., from the period of Early New High German (ENHG). Table 1 provides more information about the texts that we used in our study.⁴

3.2 LAKS Corpus

The LAKS corpus (*Leipziger und Amberger Kanzleisprache*) consists of texts compiled in the chanceries of the medieval German municipalities Leipzig and Amberg. Leipzig is located in the dialectal area of Eastern Central Germany, Amberg belongs to the Eastern Upper German dialectal area. Like the Anselm texts considered in our

³Project partners (and responsible for the transcriptions) are Simone Schultz-Balluff and Klaus-Peter Wegera, Ruhr-University Bochum.

⁴ BER_m – Berlin, NUR_m – Nuremberg, SAR_m – Sarnen, WEI_m – Weimar, MEL_m – Melk, AUG_p – Augsburg, COL_p – Cologne, AMB_l – Amberg, LEI_l – Leipzig. These locations indicate the depository in the case of manuscripts, and the place of printing in the case of prints.

Corpus	Text	Size	Type	Dialect	MSTTR±SD	MTLD	MSTTR±SD	MTLD
Anselm	BER _m	1,028	ms	ECG	0.701±0.051	73.9		
	NUR _m	1,003	ms	NUG	0.694±0.053	72.2	0.701±0.053	78.1
	SAR _m	1,022	ms	WUG	0.712±0.047	84.7		
	WEI _m	1,003	ms	EUG	0.669±0.035	68.8		
	MEL _m	4,536	ms	EUG	0.693±0.044	74.5		
	AUG _p	1,022	pr	UG	0.704±0.047	85.4		
LAKS	COL _p	984	pr	ECG	0.767±0.042	131.2	0.740±0.052	110.9
	AMB _l	1,013	ms	EUG	0.729±0.081	101.7	0.729±0.063	105.7
LEI _l		1,027	ms	ECG	0.733±0.051	99.5		

Table 1: Information on the components of the test corpus; ms = manuscript, pr = print. The dialect abbreviation mainly consists of three letters: 1st letter: E = East, W = West, N = North; 2nd letter: C = Central, U = Upper; 3rd letter: G = German.

study, the LAKS texts were written in the 15th century.

Medieval urban chanceries were concerned with the production of official documents. That included adjudications on disputes between townsmen, settlements on, e.g., inheritance disputes, and further administrative affairs. The terms and arrangements were put down in writing by a municipal clerk: a person with a university degree, excellent writing skills, and a high reputation within the municipality.

The basis of the LAKS corpus are printed editions of the original manuscripts created by Steinführer (2003), Laschinger (1994), and Laschinger (2004). The editions were aimed at an audience of medieval historians; therefore they made minor adjustments, e.g. regarding punctuation, capitalization, and the representation of special graphemes. For our study, we selected two 1,000 token fragments of the Amberg and Leipzig subcorpora, see Table 1.

3.3 Lexical Diversity of the Texts

As measures of lexical diversity, Table 1 reports the Mean Segmental Type-Token Ratio (MSTTR), and Measure of Textual Lexical Diversity (MTLD, McCarthy and Jarvis (2010)) for each individual text as well as for each subcorpus. MSTTR is the average TTR of all segments of 100 tokens in a text or corpus. MTLD is the average number of consecutive tokens it takes to reach a TTR stabilization point of 0.72.

Anselm prints and LAKS texts have higher scores for both metrics than the Anselm manuscripts, which indicates higher lexical diver-

sity. This poses a challenge for the normalization system since higher diversity means that the system cannot generalize the acquired data as well as for the less diverse texts.

4 Normalization Guidelines

For the normalization of historical data, we developed a set of guidelines (Anselm Project, 2012). The main principle is the distinction between *normalization* (Sec. 4.1) and *modernization* (Sec. 4.2). Normalization maps a given historical word form to a close modern cognate, modernization adjusts this form to an inflectionally or semantically appropriate modern equivalent, if necessary.

4.1 Normalization

Normalization as defined here is the transformation of a historical form into its modern equivalent by implementing sound as well as spelling changes. This step presupposes that the word is still part of the modern language's lexicon.

A common sound change from ENHG to modern New High German (NHG) is, e.g., monophthongization of diphthongs. For instance, ENHG /ue/ (often spelled <ü>) became /u:/ (<u>).

Furthermore, historical word forms often show a high degree of spelling variation due to the lack of a standardized orthography. This variation needs to be mapped onto the modern language's orthography. A common example is the ENHG letter <v>, which is realized as either <w>, <u> or <v> in NHG orthography.

For normalizing proper nouns such as *Judas*, exhaustive lists of modern standard forms are pro-

vided to ensure that they are normalized in the same way. Ex. (1) shows an extract of AUG_p, along with its normalization (in the second line).

- (1) *Do giench Judas zv meinem chind
da ging Judas zu meinem Kind
then went Judas to my child*
“Then Judas went to my child”

4.2 Modernization

Not all instances are as easy to normalize as in Ex. (1). Sometimes, the modern cognates created by normalization need to be adjusted inflectionally and semantically, to adhere to present-day syntax and semantics rules. An adjustment of inflection is necessary if inflectional paradigms change over time. Possible changes are the loss of inflectional affixes or changes in inflection class assignment; another example is the change of a word’s grammatical gender. We call this type of adjustment ‘modernization’ to distinguish it from the “pure” (conservative) normalization process described above.

Particularly difficult examples are “false friends”: ENHG word forms that look like a modern equivalent but need to be adjusted nevertheless. An example is provided in (i) in Table 2. ENHG *kyndt* ‘child(ren)’ refers to multiple children in that context, but looks like a singular form from a NHG perspective. The modern equivalent would be *Kinder* (plural) rather than *Kind* (singular). Normalizing tools that operate on individual word forms—as most tools do (but see Jurish (2010))—would probably produce the singular word form. For further processing, the plural word form is to be preferred (otherwise subject–verb agreement is violated). Hence, we decided to retain two forms: the close normalization *Kind* in column NORM (i.e., the modern singular) and an adjusted modernization *Kinder* in column MOD (with the modern plural suffix).

Changes in a word’s meaning occur, e.g., due to widening and narrowing, or amelioration and pejoration. To determine the meaning, the context of a given word is crucial (which, again, poses problems for most tools).

Moreover, modernization of a given word form can have an immediate effect on the surrounding word forms, e.g., if the semantically modern-

	ENHG	NORM	MOD	Translation
(i)	alle schult die die fraw und ire kyndt schuldig sint	alle Schuld die die Frau und ihre Kind schuldig sind	✓ ² Schulden ✓ ✓ ✓ ✓ ² Kinder ✓ ✓	all debts that the woman and her children owing are
(ii)	vnd er gab Ioseph das vrlaub	und er gab Joseph das Urlaub	✓ ✓ ✓ ✓ ² die ¹ Erlaubnis	and he gave Joseph the permission
(iii)	vnd zuhant nam yn pilatus	und zehant nahm ihn Pilatus	✓ ³ sofort ✓ ✓ ✓	and immediately took him Pilatus

Table 2: Examples for both types of normalization. Columns *NORM* and *MOD* represent (close) normalization and modernization, respectively. If both forms are equivalent, column *MOD* is marked by ✓. Super-scribed numbers in column *MOD* indicate semantic (1) and inflection (2) conflicts, and extinct word forms (3).

ized form has a different gender than the historical and normalized forms. Thus, adjacent word forms might need to be adjusted, too. An example is given in (ii) in Table 2. ENHG *vrlaub* has changed its meaning from ‘permission’ (NHG *Erlaubnis*) to ‘vacation’ (NHG *Urlaub*). Because *Urlaub* and *Erlaubnis* have different genders, the preceding determiner has to be adjusted, too.

4.3 Extinct Word Forms

In some cases, no close cognate exists for a historical word form. In that case, we decided to annotate a “virtual” historical word form as the normalized form, along with a suitable NHG translation as the modernized form.

To determine the virtual word form, annotators are asked to consult, in a given order, a range of printed dictionaries⁵ to look up the standardized

⁵In our case:

lemma forms. Suitable modern inflectional endings are added to these lemmas.

An example is provided in (iii) in Table 2. The ENHG word form *zuhant* ‘immediately’ has no NHG equivalent. Hence, it is first normalized by the Lexer ENHG lemma *zehant*. Second, it is translated to the modern equivalent *sofort* (also meaning ‘immediately’).

4.4 Annotation Results

In our test corpus, 10% of the words were assigned different normalized and modernized word forms in the manual annotation. Among these, 50% are inflection conflicts, 24% semantic conflicts, and 26% concern extinct forms.

For the evaluation of the guidelines, four samples from the Anselm corpus were normalized and modernized manually by two annotators trained on our guidelines. The four samples have a length of 500 tokens each, represent four different dialectal areas, and differ regarding their content. We calculated percent agreement between the two annotators, see Table 3.

Text	Agreement	
	NORM	MOD
BER _m	91.47%	92.06%
NUR _m	85.51%	84.49%
SAR _m	88.02%	89.02%
WEI _m	88.42%	87.82%
Avg.	88.38%	88.38%

Table 3: Inter-annotator agreement on Anselm manuscripts

Average agreement for both tasks is 88.38%. This shows that normalizing is a nontrivial task. However, frequent errors include inflectional adjustments erroneously marked in the NORM rather than the MOD column by one of the annotators.

5 Normalization Tool *Norma*

Norma is a tool for automatic or semi-automatic normalization of historical texts. It is intended to

1. Lexer: <http://woerterbuchnetz.de/Lexer>
2. Deutsches Wörterbuch by Jacob and Wilhelm Grimm: <http://woerterbuchnetz.de/DWB>
3. Deutsches Rechtswörterbuch: <http://drw-www.adw.uni-heidelberg.de/drw>

be flexible, as it can be used with any normalization method or a combination of such methods, and can be used both interactively and non-interactively. It supports normalization methods with a trainable set of parameters, which can be dynamically retrained during the normalization process, thereby implementing an incremental learning approach. At the time of writing, only a command-line interface is available.

5.1 Description

Normalization in the *Norma* tool is a modularized process, where each module—or “normalizer”—represents an automatic normalization method. The actual normalizing (and training) methods are implemented individually within each normalizer; *Norma* does not provide this functionality by itself, but rather invokes the respective methods of its normalizer modules.

A normalizer takes a historical word form as input and outputs a suggested modern equivalent along with a confidence score. Confidence scores are numerical values between 0 and 1; a confidence score of 0 is taken to mean that the normalizer failed to find any normalization. In order to normalize a given input word form, multiple normalizers can be combined to form a “chain”; i.e., if the first normalizer fails to find a modern equivalent, the second normalizer in the chain will be called, and so on. An example configuration is presented below. As soon as a normalizer finds an acceptable modern equivalent, this is considered the final normalization, and the chain is stopped. If no modern equivalent is found at all, the original word form is left unchanged.

This method is comparatively simple when compared to VARD (Baron and Rayson, 2008), which chooses the best candidate by calculating an f-score from all normalizers’ suggestions. Further extensions to *Norma* are conceivable to allow for more sophisticated combinations of normalizers, but are currently not implemented.

Each normalizer may also utilize a set of parameters and implement a method to dynamically train them. The training method is given both a historical word form and its modern counterpart, which can then be used by the normalizer to adjust its parameters accordingly. This way, normalizers can be adapted to different types of texts, di-

alects, or even languages.

Norma can be used in three different modes: batch, training, and interactive mode. In batch mode, texts are normalized as described above without any user interaction. For training mode, an input file containing both historical and modern word forms must be given, which will then be used to train the normalizers. In interactive mode, the input text is processed step-by-step: for each historical word form, the user is presented with the suggested normalization, which can then be either confirmed or corrected. In either case, the resulting pair of historical and modern word form is passed on to the normalizers' training methods. This represents an incremental learning approach and should ideally improve the results over time, gradually reducing the amount of manual corrections the user has to make.

5.2 Example Configuration

For our own experiments, we used a configuration with two normalizers: a *wordlist substitution* engine; and a *rule-based normalizer*. All input texts were pre-processed to plain alphabetic characters (cf. Bollmann et al. (2011)) and converted to lower case.

Wordlist substitution is one of the simplest approaches to normalization: historical word forms are looked up in a wordlist, where they are directly mapped to one or more modern equivalents. Mappings can trivially be learned from training data; additionally, each mapping is enriched with information on how many times it was learned. This way, whenever a word form is mapped to more than one modern word form, a decision can be made based on which mapping is the most frequent one. Consequently, the confidence score is calculated by dividing the frequency of the chosen mapping by the summarized frequency of all mappings for the given word form.

When a historical word form cannot be found in the wordlist, the *rule-based normalizer* is invoked. The rule-based approach is described in detail in Bollmann et al. (2011). Its main idea is to apply a set of character rewrite rules derived from training data to a historical word form, thereby producing the modern spelling of the word. These rewrite rules operate on one or more characters and also take their immediate context into ac-

count. Ex. (2) shows a sample rule.

- (2) $v \rightarrow u / \# - n$
(‘v’ is replaced by ‘u’ between the left word boundary (‘#’) and ‘n’)

Input word forms are processed from left to right, with one rewrite rule being applied at each position according to a probability score, which also determines the confidence score of the generated word form. One additional restriction is imposed on the final output word form: to prevent the generation of nonsense words, each generated word form is checked against a (modern) *dictionary*. Word forms not found in the dictionary are discarded, so that only words contained in the dictionary can ever be generated by this method.

Learning rewrite rules from training data is done via a modified algorithm for calculating Levenshtein distance, which—instead of simply counting the number of edit operations—keeps track of the exact edit operations required to transform the historical wordform into its modern equivalent.

Note that neither method currently takes token context into account; word forms are only considered in isolation. Due to the sparseness of our data, it is unclear whether including context information can actually improve overall accuracy. However, Jurish (2010) has used token context with promising results, so this is a possible line of future research.

6 Evaluation

To evaluate the performance of *Norma* with its current modules, we manually normalized and modernized six Anselm text fragments and two LAKS fragments of around 1,000 tokens each, and one Anselm text of 4,500 tokens (see Table 1). For the evaluation, we tested different scenarios:

(i) Normalization vs. modernization: Full modernization often needs contextual information, e.g., to adjust inflection to modern usage (see Sec. 4.2). Since our tool currently operates on individual word forms only, we expect considerably higher accuracy with normalized data as compared to modernized data.

(ii) Retraining vs. training from scratch: We investigated whether building upon replacement

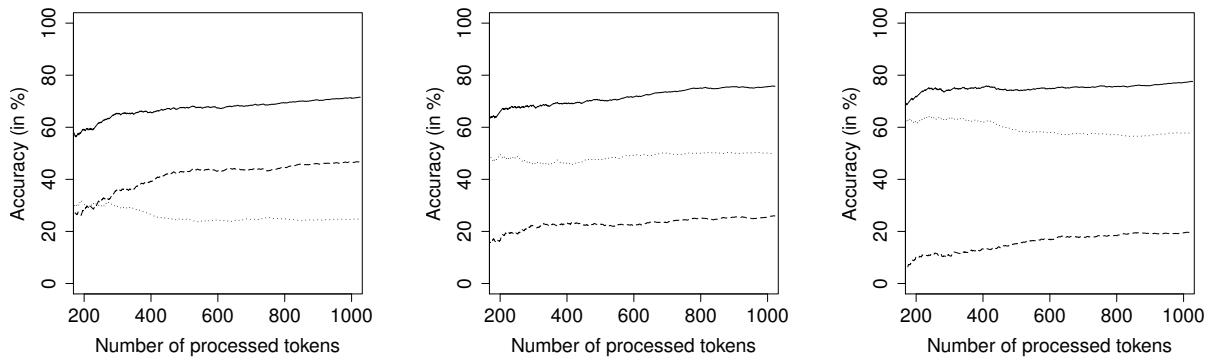


Figure 1: Learning curves from different text types and scenarios: left: BER_m (Anselm manuscript), center: AUG_p (Anselm print), right: LEI_l (LAKS). The solid line (on top) indicates accuracy, the dotted line (in light gray) is the baseline (identity mapping), and the dashed line shows the net learning curve (= accuracy minus the baseline).

rules that were derived from Luther’s bible from 1545 (see Bollmann et al. (2011) for details) would facilitate the normalization task.⁶ Since the language used by Luther is already quite close to modern German, these rules might not be of much help, since they depend on the idiosyncrasies of the evaluated texts.

(iii) Modern dictionary: We also experimented with different dictionaries that the generated word forms would be checked against. In one scenario, the complete vocabulary of a modern Luther bible was used. In the second scenario, the bible wordlist was complemented by a full-form lexicon, consisting of all simplices that can be generated by the German morphology DMOR (Schiller, 1996). Finally, as a kind of “upper bound”, we added all modern word forms of the test texts to the bible-DMOR lexicon, so that the final lookup would never (or rarely) fail.

Table 4 lists the five overall best and worst results (without using the upper-bound dictionary). The COL_p text is the hardest one, showing lots of characteristics of Low German.

The differences between training from scratch and retraining on bible rules could be used as an indication as to how close the text’s language is to Luther’s language: if accuracy improves a lot

⁶5.5 million rule instances of about 10,000 types have been learned from Luther, while each of our short text fragments yields only between 8,500–11,000 instances of 1,200–1,500 types.

Corpus	Norm	Training	Dict	Acc.
LEI _l	norm	retrain	bible	79.65%
LEI _l	norm	retrain	b+d	79.23%
NUR _m	norm	retrain	bible	78.13%
NUR _m	norm	retrain	b+d	77.83%
AUG _p	norm	retrain	bible	77.79%
SAR _m	mod	scratch	bible	64.05%
COL _p	mod	retrain	bible	63.43%
COL _p	mod	retrain	b+d	62.50%
COL _p	mod	scratch	bible	62.09%
COL _p	mod	scratch	b+d	61.78%

Table 4: Overall best and worst results, with different texts and settings; “b+d” refers to the bible dictionary augmented by DMOR forms.

thanks to retraining as opposed to training from scratch, the text must be close to Luther. It turns out that WEI_m and NUR_m profit most from retraining, whereas COL_p and BER_m show small improvement only. This suggests that Luther’s language is more similar to Upper than Central German.

For each text, we created *learning curves* that show how much manual input is needed to arrive at an accuracy of n percent in these different scenarios. Fig. 1 shows the learning curves for one selected text of each subcorpus (Anselm manuscript, Anselm print, LAKS manuscript). The setting in all three cases was: (i) normal-

ization (rather than modernization); (ii) training from scratch; (iii) use of the bible dictionary.

Accuracy (displayed by a solid line) is computed as the ratio of correctly normalized tokens divided by the total number of tokens treated so far (punctuation marks were excluded from the set of tokens). The plots include a simple baseline (displayed by a dotted line): accuracy of a normalizer that does not modify the input word form at all. This is equal to the number of historical word forms that are identical with the corresponding modern word form (and shows how close the text’s language is to modern German). Finally, a dashed line indicates the *net* learning curve, which is accuracy minus the baseline. Only results from 200 tokens onward are plotted.⁷

The plot to the left (BER_m) shows a rather difficult example of the Anselm manuscripts. The baseline is very low, around 25%. Final overall accuracy (after having normalized 1,000 tokens) is 71.5%. The plot in the center (AUG_p) shows the curve of one of the Anselm prints. Here, the baseline is rather high, around 48%. Final overall accuracy is 75.7%. Finally, the plot to the right (LEI_l) displays the results of one LAKS manuscript. The baseline is extremely high, around 60%. Final accuracy is 77.5%. In all three cases, overall accuracy as well as net learning curve show a clear steady growth.

Such individual test runs show rather diverse results. In the following, we try to highlight tendencies, by summarizing our main findings.

(i) Normalization vs. modernization As expected, generating normalized rather than modernized word forms is considerably easier. Accuracy improves by 5.0 ± 0.7 (Anselm prints) to 6.1 ± 0.9 (LAKS manuscripts) percentage points for the 1,000-token texts, averaged over all settings. Interestingly, the gap is smaller when more text is available: improvement with the Anselm 4,500-token manuscript is 4.6 ± 0.1 .

(ii) Retraining vs. training from scratch All texts profit from the rules derived from the Luther bible. If we compare texts of the same size, the average improvement is smallest with the Anselm prints (accuracy improves by 1.0 ± 0.5 percentage

⁷Below 200, the accuracy fluctuates too much to be meaningfully plotted.

points). Improvements are more pronounced with the Anselm 1,000-token manuscripts (2.2 ± 1.0) and the LAKS manuscripts (2.3 ± 0.9). As can be expected, the differences become less important if the text size is increased: improvement with the Anselm 4,500 text is 0.5 ± 0.2 .

(iii) Modern dictionary We observe that the choice of dictionary has less impact on accuracy than the other factors. Average differences are between 0.4 ± 0.3 percentage points (with Anselm manuscripts) and 1.8 ± 1.3 (with LAKS texts). As expected, the “upper-bound” dictionary, which contains all target words, is found most often among the top-10 settings of each subcorpus (in roughly 75% of the cases). However, availability of such a dictionary is certainly not a realistic scenario.

Comparing the original bible dictionary with its DMOR-augmented version, it turns out, surprisingly, that in 69% of the scenarios, the bible dictionary performs better than the augmented version. With the LAKS corpus, however, the augmented version is clearly preferable. This can be attributed to the fact that LAKS, being a corpus of administrative texts, contains many out-of-domain words, which are not covered by the bible dictionary.

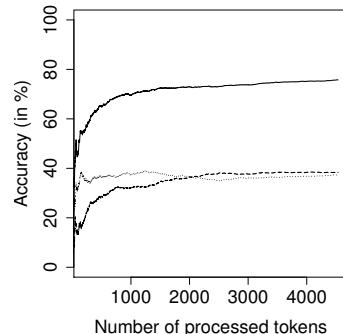


Figure 2: Learning curve for the 4,500 tokens Anselm text (MEL_m); lines as in Fig. 1.

The plots in Fig. 1 show that accuracy still improves considerably when more data is added. According to Baron and Rayson (2009), the first 2–3,000 tokens yield a steep increase in performance (for recall). We therefore normalized one entire Anselm text, MEL_m , with 4,500 tokens, see Fig. 2. The plot seems to suggest that the “turn-

ing point”, after which performance increases less rapidly, is already reached after 1,000 tokens.

Correlating accuracy with lexical diversity (Sec. 3.3), it turns out that more diverse texts (Anselm prints and LAKS manuscripts) achieve higher accuracies. However, this can be attributed to the fact that their baselines are also higher in general. In fact, less diverse texts (Anselm manuscripts) show larger increases of accuracy over their baselines (Anselm manuscripts: 35.4 ± 6.5 percentage points; Anselm prints: 31.8 ± 5.7 ; LAKS manuscripts: 17.9 ± 3.4).

7 Conclusion

In this paper, we presented normalization guidelines, and a semi-automatic normalization tool. We have argued for a two-step approach to normalization: one level (“normalization”) which stays formally close to the original, and another level (“modernization”) which approximates modern language. Automatic generation of normalized forms is considerably easier, with improvements between 5–6 percentage points in accuracy.

In an ideal setting, both levels of normalization would be generated. The second level would serve further processing like morphological tagging. Mismatches between the first and the second level could provide important hints for inflectional and semantic changes between ENHG and modern German.

The tool *Norma* integrates lexicon lookup and rewrite rules to generate modern word forms that can be corrected in an interactive way. Corrections are used to retrain the methods, improving further normalization suggestions. An evaluation showed this approach to be promising, as accuracy increases considerably with even small amounts of training data. However, accuracy was also found to depend to a great extent on the specific text and the setting.

Possible further research includes a more exhaustive evaluation of different normalization methods and combinations of such methods in particular, for which the *Norma* tool provides an ideal framework. Furthermore, we showed that instead of training normalizers from scratch, it is often preferable to build upon previously learned data, even if it stems from a slightly different do-

main. How to best combine data from texts of different lengths, types, and/or dialects in order to improve the results on texts for which no special training was available is still an open question.

References

- Anselm Project. 2012. Guidelines for Manual Normalization of Historical Word Forms. Manuscript, Ruhr University Bochum.
- Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.
- Alistair Baron and Paul Rayson. 2009. Automatic standardization of texts containing spelling variation. How much training data do you need? In *Proceedings of the Corpus Linguistics Conference*.
- Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Applying Rule-Based Normalization to Different Types of Historical Texts — An Evaluation. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 339–344, Poznan, Poland.
- GerManC Project. 2012. GerManC annotation guidelines. Manuscript, University of Manchester.
- Bryan Jurish. 2010. More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.
- Johannes Laschinger. 1994. *Denkmäler des Amberg Stadtrechts. 1034–1450*, volume 1. Beck, München.
- Johannes Laschinger. 2004. *Denkmäler des Amberg Stadtrechts. 1453–1556*, volume 2. Beck, München.
- Philip M. McCarthy and Scott Jarvis. 2010. MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.
- Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of Early Modern German. In *Proceedings of the Fifth Linguistic Annotation Workshop*, pages 124–128.
- Anne Schiller. 1996. Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO. In Roland Hausser, editor, *Proceedings of the first Morpholympics*. Tübingen: Niemeyer.
- Henning Steinführer. 2003. *Die Leipziger Ratsbücher 1466–1500. Forschung und Edition*, volume 1–2. Leipziger Universitätsverlag, Leipzig.

The Sketch Engine as infrastructure for historical corpora

Adam Kilgarriff

Lexical Computing Ltd
Brighton
UK

adam@lexmasterclass.com

Miloš Husák

Lexical Computing Ltd. &
Faculty of Informatics
Masaryk University, Brno, Cz

xhusak@mail.muni.cz

Robyn Woodrow

Lexical Computing Ltd.
Brighton
UK

robyn@sketchengine.co.uk

Abstract

A part of the case for corpus building is always that the corpus will have many users and uses. For that, it must be easy to use. A tool and web service that makes it easy is the Sketch Engine. It is commercial, but this can be advantageous: it means that the costs and maintenance of the service are taken care of. All parties stand to gain: the resource developers both have their resource showcased for no cost, and get to use the resource within the Sketch Engine themselves (often also at no cost). Other users benefit from the functions and features of the Sketch Engine. The tool already plays this role in relation to four historical corpora, three of which are briefly presented.

A premise of historical corpus development is that a corpus, once created, will be widely used. If it is not easy to use it, this will not happen. In 2012, this means making it available to search over the web. You might do this by developing your own tool, or installing and using someone else's, or getting someone else to handle that whole side of things for you.

1 The Sketch Engine

The Sketch Engine is a well-established corpus query tool with a nine-year track record. It is fast, responding immediately for most queries for billion-word corpora, and offers all standard functions (concordancing, sorting and sampling, wordlists, collocates, subcorpora) and some non-standard ones. It takes its name from word sketches, one-page summaries of a word's grammatical and collocational behaviour, as in Fig 1. It is in daily use for lexicography at Oxford University Press, Cambridge University Press, Collins,

Cornelsen and Le Robert, for language research at seven national language institutes, and for linguistic and language technology teaching and research at over 100 universities worldwide.

The Sketch Engine is offered as a web service, with 200 corpora for sixty languages already loaded. Users may also upload and install their own corpus, and then use the Sketch Engine to study it. Many of the corpora in the tool are provided by their creators, often in exchange for free access for them and their colleagues. The resource developer benefits in three ways:

- access to their own corpus in the Sketch Engine, which supports them in their own research on it (including maintaining and developing it)
- an easy way to show their corpus to others, in a way that allows those others to explore it in detail
- access to other corpora already in the Sketch Engine.

The tool uses input and query formalisms developed at the University of Stuttgart for their corpus system in the early 1990s, as widely adopted across corpus and computational linguistics. There have also been extensions to the formalisms, for example for improved querying of parsed data (Jakubíček et al., 2010).

1.1 Maintenance and motivation

The maintenance of resources has often been a bone of contention for those left in charge of them. Resource developers become the victims of their own success: the more successful the resource, the greater the level of expectation that errors will be corrected and upgrades provided, yet

machen (verb) GerManC freq = 1403 (1752.0 per million)

AttrY	VerbX	630	3.3	VerbX+VerbY	1988	2.2	VerbX PräpY	166	2.0	VerbX+AdvY (Vorsilben)	370	1.9
ein	187	8.05		wollen	75	8.62	aus	12	9.17	selbst	14	8.8
d	418	7.4		haben	192	8.56	zwischen	4	9.16	ausfindig	4	8.45
vier	3	6.94		werden	155	8.17	zu	20	8.9	noch	20	8.2
tausend	3	6.86		sein	176	8.01	zur	6	8.89	mehr	9	8.09
eine	4	6.57		können	25	7.85	zum	7	8.89	also	11	8.04
3	2	6.22		sollen	52	7.8	ohne	6	8.76	so	65	8.03
				lassen	33	7.77	durch	10	8.7	nur	12	8.0
				sagen	28	7.7	von	22	8.61	schon	8	7.95
				kan	19	7.46	auf	11	8.41	nun	9	7.93
				sehen	21	7.3	mit	21	8.41	auch	29	7.93
				kommen	21	7.29	gegen	3	8.21	viel	6	7.89
				müssen	19	7.19	im	5	7.91	fort	3	7.86

Figure 1: Word sketch for *machen* in the GermanC corpus of early modern German.

research funding bodies are rarely willing to fund them, since the projects have already had their funding, and maintenance is not part of the research funders' mission. So the host organisation struggles to meet users' requests for little credit or recompense. Nor does resource maintenance offer many opportunities to publish.

Lexical Computing, the Sketch Engine company, depends for its income on the quality of its resources, and on users finding the system works well so they renew their licences. It is motivated to maintain and upgrade the hardware, software and corpora. There is an income stream to fund it, from customers.

For resource management and maintenance, there is much to be said for a market model in which the people who are maintaining a resource are motivated to do it well because their income depends on it.

1.2 Local vs. remote

One of the biggest questions about software, in the age of the web, is: should it be local or remote? Should we download and install, or interact through browsers and APIs? For a growing number of applications, 'remote' is gaining ground. More and more people manage their documents and photos, and read their email, on remote servers. When I want to convert a document

from .ps to .pdf, I do it at <http://ps2pdf.com>. Corpus research is an area where 'remote' is a very appealing answer, as:

- corpora are large objects which are often awkward to copy
- copying them to other people can be legally problematic
- there are many occasional and non-technical potential corpus users who will not use them if it involves software installation
- the software is more easily maintained and updated
- the user does not need to invest in hardware, or expertise for support and maintenance.

For all of these reasons, the preferred model for most corpus use is the remote one. To support users who want robot access there is a web API.

2 Historical Corpora in Sketch Engine

There are currently historical (pre-20th-century) resources publicly available in the Sketch Engine for three languages: Latin (McGillivray and Kilgarriff, 2011), English and German.¹

¹The Sketch Engine is also being used in the ChartEx project (<http://www.chartex.org>) which is applying text min-

2.1 The Corpus of English Dialogues Corpus

The Corpus of English Dialogues 1560-1760 (Culpeper and Kytö, 2010) was created to explore how English pragmatics developed by gathering historical speech and speech-like data. It comprises 1.2 million words of trial proceedings, witness depositions, play-texts, dialogue in prose fiction and didactic dialogues, including ones from language teaching textbooks. Fig 2 shows a concordance for *prihee*, sorted by date, with the genre of the text also shown. Here we are showing changes of speaker turn by adding the name of the speaker between or-bars, in green and italics (other options are easily set up). Note also the facility for navigating to a particular date.

2.2 Penn Historical Corpora (PHC)

The Penn Historical Corpora are the Penn-Helsinki Parsed Corpus of Middle English (second edition; PPCME2), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME), and the Penn Parsed Corpus of Modern British English (PPCMBE). They all comprise texts and text samples of British English prose from the earliest Middle English documents up to the First World War.² Fig 3 shows the ascent of *should* over the past 500 years.

A business issue arose when a user asked if they could access the PHC in the Sketch Engine. Penn have been selling the PHC, on CD-Rom, and this has been funding ongoing research and maintenance. So its creator was keen to make the PHC available in the Sketch Engine - but not at the expense of the income stream. The solution we have adopted is that only those users who have bought the CD-Rom will get access to the PHC in the Sketch Engine, and purchasers of the CD-Rom will receive a year's free access to the Sketch Engine so they can look at PHC (and all the other corpora) there.

2.3 GermanC

GermanC (Durrell et al., 2007) is a corpus of 800,000 words of 17th and 18th century German.

ing methods to medieval Latin charters. It will make the corpora it prepares publicly available through the Sketch Engine as the project proceeds.

²<http://www.ling.upenn.edu/hist-corpora/>

We demonstrate what can be done with GermanC in the Sketch Engine by looking at 'key-word lists', the words with the biggest contrast between frequencies in one corpus (or subcorpus) and another. Here we focus on the more frequent words (by adjusting the 'simplemaths parameter' (Kilgarriff, 2009)).

The fifty top mid-to-high frequency lemmas³ of the 17th century subcorpus of GermanC, in contrast to the 18th century part, include:

ach allhier also Artikel auch begehrn
berichten Christus damit dann darauf der-
selbige dito etlich Feind Fürst gar Gnade
Gott haltenjenige Kapitel Komet König
Leib lieb mit mögen oder Ort Pferd Rat
solch sollen sonder sonderlich sonst statt
Tod Türke und vom wann weil wider
wiederum wohl Zeche

The corresponding 18th century items are:

Absicht Art Begriff besonders denken der-
jenige dies eben ein Erde finden Freund
für Gegend Gegenstand Geschlecht Graf
hier ich immer jeder klein können Körper
machen Mann mein Mercurius Mutter
Natur nur nötig scheinen schon Seite Sie
suchen Teil Ton um Umstand Vater ver-
schieden wahr weit wenigstens wenn wirk-
lich zeigen

The word with the highest frequency contrast, with over double the relative frequency in 17c vs. 18c, was *wann*. At the top of the 18c list was *wenn*.⁴ Both words are of similar frequency. This strongly suggests that people were making a choice between *wann* and *wenn*, and in the 17th century they more often chose *wann*, and in the 18th, *wenn*. While the changes affecting these two words are already familiar (Wright, 1907; Abraham, 1978), with GermanC in the Sketch Engine we can directly explore exactly what the changes are and when and how they took place.

Several other words in the 17c list (*allhier*, *der-selbige*) are marked in dictionaries as old, or obsolete.

³Lemmatisation was by a version of TreeTagger trained on GermanC data (Scheible et al., 2012).

⁴Here the lists are alphabetised. In the tool, *wann* and *wenn* appear at the tops of the two lists.

Corpus: English Dialog Corpus
Hits: 67 (49.3 per million)

Page 1 of 4 (Go) [Next](#) | [Last](#) Concordance is sorted. Jump to: 1676

1602,Drama comedy	And in that vault my bodie I will lay, I prithee leue me, thither is my way. I am sure
1602,Drama comedy	Wanting a rope, I well could hang my selfe: I prithee Mistris, for all my long seruice, For all
1602,Drama comedy	are my seuen yeares Prentiship of woe. I prithee be patient, had some occasion That did
1641,Miscellaneous text types	my feete. Me thinks you're very eloquent: Prithee tell me, Don't Suada , and the Jove-begotten-br
1641,Miscellaneous text types	you produce such eloquent speeches now. Prithee what ist? How? Cause of alacrity? Sfoot
1641,Miscellaneous text types	And dost thou tell me of that? Fie, fie! Prithee why? I did but conjecture out of your sweet
1641,Miscellaneous text types	I acted a Comedy, 'tis so long agoe. But Prithee how comes it to passe that you act Tragedies
1647,Miscellaneous text types	be abortive. How doest thou live then, I prithee ? By my wit, I tell thee, and hold it a
1647,Miscellaneous text types	thee an Inch thou'l take an Ell: but I prithee don't vaunt before the victory: suppose
1652,Fiction	better : and said to the Horsecourser I prithee put my saddle on the horse: that I may
1653,Language teaching handbook: French	goe thither. Lead us in Sir, 'tis yours. Prithee honest friend, draw us a pint of the best
1653,Language teaching handbook: French	up into this Room. This wine is naught, prithee give us better. Sir, the Wine is good,
1661,Miscellaneous text types	church lane , said her husband did her. Prithee Gusman keep thy breath to cool thy pottage
1669,Drama comedy	tol de re tol de re la . Claps their backs Prithee be not so rude Trice . Huswife Constance
1669,Drama comedy	be very glad to see thee at my Lodging; prithee let's not be such strangers to one another
1676,Drama comedy	with the Jewel in his hand he promis'd me; prithee leave me alone with him. Speed the Plough
1676,Drama comedy	a man of Wit, and understands The Town: prithee let thee and I be intimate, There is no
1676,Drama comedy	so improper For such a business as I am. Prithee ! why hast thou so modest an Opinion of
1676,Drama comedy	cannot pitch on a better for your purpose. Prithee ! what is she? A person of Quality, and
1676,Drama comedy	affectionat: You will see Her i'th' Mail to night. Prithee , let thee and I take the Air together.

Page 1 of 4 (Go) [Next](#) | [Last](#) Concordance is sorted. Jump to: 1676

Figure 2: Concordance for *prithee* in the English Dialogues corpus, sorted by date, showing genre and speaker-turns.

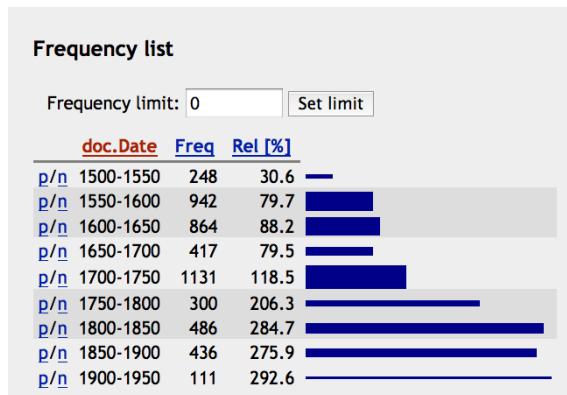


Figure 3: Analysis by time period of *should* in Penn Historical Corpora. Bars show frequency normalised according to the quantity of data available for each time period.

Other notable contrasts are that 17c has more formal texts, talking about religion and royalty, whereas 18c talks more about ordinary life: mothers and fathers and friends. Whether this is an artefact of the selection of texts to go into the corpus, or a reflection of changes in the language or of the role of writing in society, would require investigation beyond the scope of this paper.

An advantage of working in the Sketch Engine is that there are many corpora available for comparison. So we can also make a keyword list for all of GermanC, in contrast to a different corpus, for example deWaC, a web-crawled corpus of contemporary German (Baroni et al., 2009).⁵ The keywords of GermanC are:

aber alle allein also anderer bald da daß
dein derselbe doch du eben einig er gar
gemacht geschehen gewiß gleich Gott hal-
ten Herr Herz ich Ihr König lassen man
mein mögen nichts nun Ort Sache sehen Sie
so solch sollen sonst tun unser wann was
weil Welch wo wohl woollen

While some items (*alle, einig, er, Ihr, Sie*) appear in the list owing to lemmatisation differences and others (*daß*) owing to spelling differences, others are linguistic, owing either to the differences in texts included, or to some other differences in German society or language over the last three centuries. GermanC makes far more use of first and second person pronouns, short time adverbials (*bald, doch*), and some conjunctions (*aber, da*). These seem to be indicative of GermanC being, overall, a corpus of less formal texts than DeWaC.

All three lists contain a number of pronouns. We find pronouns at the top of keyword lists time and time again. Pronouns are the litmus paper of text type.

We may suspect that the 18c-17c comparison is very different to the GermanC-DeWaC comparison since the components of GermanC will be, overall, much more similar to each other than

⁵ As different TreeTagger models were used for the two corpora, there will be slight differences in lemmatisation. With this in mind we also explored the keyword list of word forms. But this was dominated by spelling variants, which had been addressed by giving the normalised form of the lemma, so the lemma list was more informative.

GermanC is to DeWaC. For ways to explore this topic see (Kilgarriff, 2001; Kilgarriff, 2012).

3 Further publishing possibilities

A scenario currently under discussion with one corpus developer involves the Sketch Engine taking on a publisher role, including collecting payments and passing them on to the developer. While some universities have departments that could undertake this role, their costs are often high, and there are benefits to working with a flexible small company with expertise –technical, commercial and legal– in corpora.

4 Conclusion

Two problems often confronting corpus developers are:

1. how to make it easy for everyone to use the corpus
2. how to maintain it and continue to make it available, over a number of years.

We have shown one solution to these problems: subcontract to a commercial company with appropriate tools and expertise. We have shown how this works in several cases, showing a range of the functions and display options that the Sketch Engine offers that are of particular relevance for historical data, and demonstrating how we can immediately make interesting findings using the Sketch Engine.

Acknowledgements

With thanks to Laura Giacomini for assistance with the analysis for German and to Ravi Kiran for the encoding of the Corpus of English Dialogues.

References

- Werner Abraham. 1978. The role of fallacies in the diachrony of sentence connectives. *Studies in Second Language Acquisition*, 1(1):95–134.
Marco Baroni, Silva Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

- Jonathan Culpeper and Merja Kytö. 2010. *Early Modern English dialogues: spoken interaction as writing*. Cambridge University Press.
- Martin Durrell, Astrid Ensslin, and Paul Bennett. 2007. The GerManC project. *Sprache und Datenverarbeitung*, 31:71–80.
- Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. Fast syntactic searching in very large corpora for many languages. In *Proceedings of PACLIC 24*, pages 741–747.
- Adam Kilgarriff. 2001. Comparing corpora. *International journal of corpus linguistics*, 6(1):263–276.
- Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*.
- Adam Kilgarriff. 2012. Getting to know your corpus. In *Text, Speech and Dialogue*. Springer.
- Barbara McGillivray and Adam Kilgarriff. 2011. Tools for historical corpus research, and a corpus of latin. *New Methods in Historical Corpora*.
- S. Scheible, R.J. Whitt, M. Durrell, and P. Bennett. 2012. GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German. *LREC*.
- Joseph Wright. 1907. *Historical German Grammar*, volume 1. OUP.

Evaluating a Post-editing Approach for Handwriting Transcription

Verónica Romero, Joan Andreu Sánchez, Nicolás Serrano, Enrique Vidal

Departamento de Sistemas Informáticos y Computación

Universitat Politècnica de València, Spain

{vromero, jandreu, nserrano, evidal}@dsic.upv.es

Abstract

Marriage license books are documents that were used for centuries by ecclesiastical institutions to register marriage licenses. These books, that were handwritten until the beginning of the 20th century, have interesting information, useful for demography studies and genealogical research. This information is usually collected by expert demographers that devote a lot of time to manually transcribe them. As the accuracy of automatic handwritten text recognizers improves, post-editing the output of these recognizers could be foreseen as a possible alternative. Unluckily, most handwriting recognition techniques require large amounts of annotated images to train the recognition engine. In this paper we carry out a study about how the handwritten recognition system accuracy improves with respect to the amount of training data, and how the human efficiency increases during the transcription of a marriage license book.

1 Introduction

In the last years, huge amounts of handwritten historical documents residing in libraries, museums and archives have been digitalized and have been made available to scholars and to the general public through specialized web portals. Many of these documents are collections of historical documents containing very valuable information in the form of records of quotidian activities. One example of this kind of handwritten documents are the marriage license books considered in this

paper. In many cases, it would be interesting to transcribe these document images, in order to provide new ways of indexing, consulting and querying them.

Transcribing handwritten images manually is a very laborious and expensive work. This work is usually carried out by experts in palaeography, who are specialized in reading ancient scripts, characterized, among other things, by different handwritten/printed styles from diverse places and time periods. How long experts take to make a transcription of one of these documents depends on their skills and experience.

The automatic transcription of these ancient handwritten documents is still an incipient research field that in recent years has been started to be explored. Currently available OCR text recognition technologies are very far from offering useful solutions to the transcription of this sort of documents, since usually characters can by no means be isolated automatically. Therefore, the required technology should be able to recognize all text elements (sentences, words and characters) as a whole, without any prior segmentation. This technology is generally referred to as “*offline Handwritten Text Recognition*” (HTR) (Marti and Bunke, 2001). Several approaches have been proposed in the literature for HTR that resemble the noisy channel approach that is currently used in Automatic Speech Recognition. Thus, the HTR systems are based on Hidden Markov Models (HMM) (Toselli and others, 2004), recurrent neural networks (Graves et al., 2009) or hybrid HMM and neural networks (España-Boquera et al., 2011). These systems have proven to be

suit for restricted applications with very limited vocabulary or constrained handwriting achieving in these kind of tasks relatively high recognition rates. However, in the case of transcription applications of unconstrained handwritten documents (as old manuscripts), the current HTR technology typically achieves results which are far from perfect.

Therefore, once a full recognition process of the document has finished, human intervention is required in order to produce a high quality transcription. The human transcriber is, therefore, responsible for verifying and correcting the mistakes made by the system. In this context, the HTR process is performed off-line: First, the HTR system returns a full transcription of all the text lines in the whole document. Then, the human transcriber reads them sequentially (while looking at their correspondence in the original page images) and corrects the possible mistakes made by the system. Whether the HTR system accuracy is good enough, this post-editing approach can be foreseen as a possible alternative to manually transcription.

In the above mentioned HTR approaches, the character and language models are stochastic models whose parameters are automatically learned from annotated data. One of the bottlenecks of these approaches is the need of annotated data in order to automatically train the models. These annotated data is not usually available at the beginning of the transcription of new document, but as the document is being transcribed, new transcribed material is available for training the HTR models.

In this paper we carry out a study about how the performance of an HTR system varies as the amount of data that is available to train the models increases. First, an HTR system provides automatic transcriptions for a few pages. Second, these transcriptions are post-edited by expert palaeographers, and the models are retrained with the post-edited transcriptions. Then, new pages are transcribed and manually reviewed, and the models are retrained. This process goes on until the complete document is transcribed. We also study how the improvements in the system accuracy produced by retraining the HTR models with the new transcribed material affect to the

human efficiency following the post-editing approach. This study has been carried out during the real transcription of a historical book compiled from a collection of Spanish marriage license books.

This task is described in detail in the following section, and the main problems that arise in these documents are explained. Then, the HTR technology used in this work is shown in section 3. The evaluation methodology and the obtained results are reported in sections 4 and 5. Finally, conclusions are drawn in Section 6.

2 Task description

Marriage license books are documents that were used for centuries to register marriages in ecclesiastical institutions. These demographic documents have already been proven to be useful for genealogical research and population investigations, which renders their complete transcription an interesting and relevant problem (Esteve et al., 2009). Most of these books are handwritten documents, with a structure analogous to an accounting book.

In this paper we have used a book from a collection of Spanish marriage license books conserved at the Archives of the Cathedral of Barcelona. In this book each page is divided horizontally into three blocks, the *husband surname's block*, the *main block*, and the *fee block* and vertically into individual license records. Fig. 2 shows a page of marriage licenses from the book used in this paper. Each marriage license (see Fig. 1) typically contains information about the marriage day, husband's and wife's names, the husband's occupation, the husband's and wife's former marital status, and the socio-economic position given by the amount of the fee. In some cases, additional information is given as well, viz. the father's names and their occupations, information about a deceased parent, place of residence, or geographical origin. The fiscal marker, as well as the exhaustive nature of the source and the variety of types of the parishes involved – from the city centre to the most rural villages – allows researching multiple aspects of demography, specially the chronology and the geography in the constitution of new social classes and groups.

Compared to modern printed documents, the

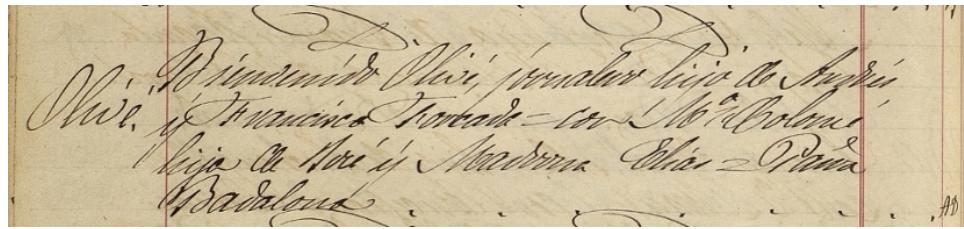


Figure 1: Example of a marriage licenses in Spanish.

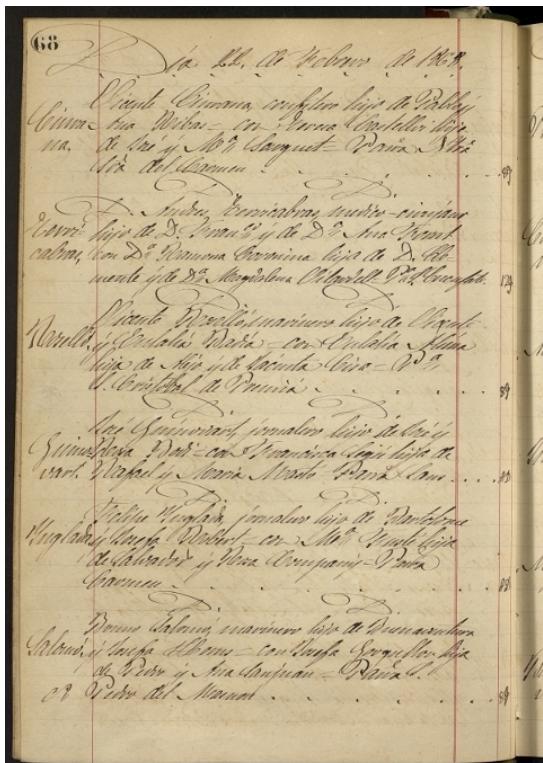


Figure 2: Example of a marriage licenses page.

analysis and recognition of these handwritten historical documents has many additional difficulties. Firstly, the typical paper degradation problems encountered in this kind of documents, such as presence of smear, significant background variation, uneven illumination, and dark spots, require specialized image-cleaning and enhancement algorithms. Secondly, show-through and bleed-through problems can render the distinction between background and foreground difficult (Drida, 2006). Thirdly, document collections spanning several centuries usually do not follow a strict standard notation, but differ from one century to another. The text contains a variety of

special symbols and other recognition challenges. Among those are abbreviations and superscripts, crossed out words with inserted corrections, Roman numerical notation and added words written between the lines.

3 Handwritten text recognition system

The handwritten text recognition (HTR) problem can be formulated as the problem of finding the most likely word sequence, $\mathbf{w} = (w_1 \ w_2 \ \dots \ w_l)$, for a given handwritten sentence image represented by a feature vector sequence $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_m)$, i.e., $\mathbf{w} = \arg \max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x})$. Using the Bayes' rule we can decompose this probability into two probabilities, $P(\mathbf{x} \mid \mathbf{w})$ and $P(\mathbf{w})$:

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) \approx \arg \max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w})P(\mathbf{w}) \quad (1)$$

$P(\mathbf{x} \mid \mathbf{w})$ can be seen as a morphological-lexical knowledge and it is typically approximated by concatenated character HMMs (Jelinek, 1998). On the other hand, $P(\mathbf{w})$ represents a syntactic knowledge and it is approximated by a word language model, usually n -grams (Jelinek, 1998).

The HTR system used here follows the classical architecture composed of three main modules: a document image preprocessing module, in charge to filter out noise, recover handwritten strokes from degraded images and reduce variability of text styles; a line image feature extraction module, where a feature vector sequence is obtained as the representation of a handwritten text line image; and finally a model training/decoding module, which obtains the most likely word sequence for the sequence of feature vectors (Bazzi et al., 1999; Toselli and others, 2004).

3.1 Preprocessing

As previously said, it is quite common for handwritten documents, and particularly for ancient documents, to suffer from degradation problems (Drida, 2006). In addition, there are other kinds of difficulties appearing in these pages as different font types and sizes in the words, underlined and/or crossed-out words, etc. The combination of all these problems contributes to make the recognition process difficult, and hence, the preprocessing module quite essential.

Concerning the preprocessing module used in this paper, the following steps take place: skew correction, background removal and noise reduction, line extraction, slant correction and size normalization. We understand as “skew” the angle between the horizontal direction and the direction of the lines on which the writer aligned the words. Skew correction is carried out on each document page image, by aligning their text lines with the horizontal direction. Then, a conventional noise reduction method is applied on the whole document image (Kavallieratou and Stamatatos, 2006), whose output is then fed to the text line extraction process which divides it into separate text lines images. The method used is based on the horizontal projection profile of the input image. Finally, slant correction and size normalization are applied on each separate line. The slant is the clockwise angle between the vertical direction and the dominant direction of the written vertical strokes. This angle is determined using a method based on vertical projection profile, and used then by the slant correction process to put the written text strokes in an upright position. On the other hand, the size normalization process tries to make the system invariant to character size and to reduce the areas of background pixels which remain on the image because of the ascenders and descenders of some letters. More detailed description can be found in (Toselli and others, 2004; Romero et al., 2006).

3.2 Feature Extraction

The feature extraction process approach used to obtain the feature vectors sequence follows similar ideas described in (Bazzi et al., 1999). First, a grid is applied to divide the text line image into $N \times M$ squared cells. In this work, $N = 20$

is chosen empirically and M must satisfy the condition that N/M is equal to the original line image aspect ratio. Each cell is characterized by the following features: *average gray level*, *horizontal component of the grey level gradient* and *vertical component of the grey level gradient*. To obtain smoothed values of these features, an 5×5 cell analysis window, centred at the current cell, is used in the computations (Toselli and others, 2004). The smoothed cell-averaged gray level is computed through convolution with two 1-d Gaussian filters. The smoothed horizontal derivative is calculated as the slope of the line which best fits the horizontal function of column-average gray level in the analysis window. The fitting criterion is the sum of squared errors weighted by a 1-d Gaussian filter which enhances the role of central pixels of the window under analysis. The vertical derivative is computed in a similar way.

Columns of cells (also called *frames*) are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in their constituent cells. Hence, at the end of this process, a sequence of M 60-dimensional feature vectors (20 normalized gray-level components and 20 horizontal and vertical gradient components) is obtained.

3.3 Training and Recognition

Characters are considered here as the basic recognition units and they are modelled by left-to-right HMMs, with 6 states and a mixture of 64 Gaussian densities per state. This Gaussian mixture serves as a probabilistic law to the emission of feature vectors on each model state. The number of Gaussian densities as well as the number of states were empirically chosen after tuning the system. Character HMMs are trained from images of continuously handwritten text (without any kind of segmentation and represented by their respective observation sequences) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called forward-backward or Baum-Welch re-estimation (Jelinek, 1998).

Each *lexical entry (word)* is modelled by a stochastic finite-state automaton which represents

all possible concatenations of individual characters that may compose the word. By embedding the character HMMs into the edges of this automaton, a *lexical HMM* is obtained.

Finally, the concatenation of words into text lines or sentences is usually modelled by a bigram *language model*, with Kneser-Ney back-off smoothing (Kneser and Ney, 1995), which uses the previous $n - 1$ words to predict the next one:

$$P(\mathbf{w}) \approx \prod_{i=1}^N P(w_i | \mathbf{w}_{i-n+1}^{i-1}) \quad (2)$$

This n -grams are estimated from the given transcriptions of the trained set.

Once all the *character*, *word* and *language* models are available, the recognition of new test sentences can be performed. Thanks to the homogeneous finite-state (FS) nature of all these models, they can be easily *integrated* into a single *global* (huge) FS model. Given an input sequence of feature vectors, the output word sequence hypothesis corresponds to a path in the integrated network that produces the input sequence with highest probability. This optimal path search is very efficiently carried out by the well known Viterbi algorithm (Jelinek, 1998). This technique allows for the integration to be performed “on the fly” during the decoding process.

4 Evaluation methodology

The handwriting recognition techniques used here require annotated images to train the HMM and the language models. In order to assess how the system accuracy varies with respect to the amount of training data available for training the HTR models and how post-editing the output of the HTR system can save human effort, in this paper we have conducted a study during the real transcription of a marriage license book. In the next subsections the assessment measures, the information of the corpus and the procedure are explained.

4.1 Assessment Measures

Different evaluation measures were adopted to carry out the study. On the one hand, the quality of the automatic transcription can be properly

assessed with the well known Word Error Rate (WER). The WER is also a reasonably good estimate of the human effort needed to *post-edit* the output of a HTR recognizer at the word level. It is defined as the minimum number of words that need to be substituted, deleted or inserted to convert a sentence recognized by the HTR system into the reference transcriptions, divided by the total number of words in these transcriptions. On the other hand, in our subjective test with a real user we measured the time needed to fully transcribe each license of the book following the post-editing approach.

4.2 Corpus

The corpus used on the experiments was compiled from a single book of the marriage license books collection conserved at the Archives of the Cathedral of Barcelona. Fig. 2 shows an example of a marriage license page.

The corpus was written by only one writer in 1868 and it was scanned at 300 dpi in true colours and saved in TIFF format. It contains 200 pages although only the firsts 100 have been used in this work. For each page, we used the GIDOC (Serrano et al.,) prototype for text block layout analysis and line segmentation. Concretely, a preliminary detection was performed by a fully automatic process using standard preprocessing techniques based on horizontal and vertical projection profiles. Then, the detected locations for each block and lines were verified by a human expert and corrected if necessary, resulting in a data-set of 2,926 text line images.

The main block of the whole manuscript was transcribed automatically and post-edited line by line by an expert palaeographer during the experimental process. This transcription has been carried out trying to obtain the most detailed transcription possible. That is, the words are transcribed in the same way as they appear on the text, without correcting orthographic mistakes. The book contains around 17k running words from a lexicon of around 3k different words. Table 1 summarizes the basic statistics of the corpus text transcriptions.

To carry out the study we grouped the pages of the document into 5 consecutive partitions of 20 pages each (1-20, 21-40, 41-60, 61-80, 81-100).

Table 2: Basic statistics of the different partitions for the database.

Number of:	P1	P2	P3	P4	P5
Pages	20	20	20	20	20
Lines	581	583	582	584	596
Run. words	3 560	3 523	3 560	3 533	3 615
Characters	19 539	19 234	19 544	19 644	19 818

Table 1: Basic statistics of the text transcriptions.

Number of:	Total
Pages	100
Lines	2,926
Running words	17,791
Lexicon size	2,210
Running characters	97,779
Character set size	84

All the information related with the different partitions is shown in Table 2.

4.3 Procedure

The study consisted in transcribing line by line, by a palaeographer expert, the first 100 pages of the marriage license book presented in the previous subsection. First, partition P1 was manually transcribed by the user without any help of the HTR system. Then, from partition P2 to P5, each partition was automatically transcribed by the system trained with all preceding partitions, which were previously post-edited by the user. This should help in improving the system accuracy.

The experimental process can be summarized in the following steps:

- Initially, the user manually transcribed the first 20 pages of the book (P1).
- The following block was automatically transcribed by the HTR system trained with all preceding transcriptions.
- Each automatically transcribed line was supervised and, if necessary, amended by the expert.
- After processing a block of pages, all supervised transcriptions were used to (re-)train the automatic transcription system.

- The previous 3 steps were iterated until all the blocks were perfectly transcribed.

The manual transcription process and the post-editing process were carried out by means of the GIDOC prototype (Serrano et al.,). In order to avoid possible biases due to human learnability to the user interface, the user became familiar with the engine.

5 Results and Discussion

Table 3 shows the results obtained in the transcription experiments for the different partitions. Column *Time* represents the minutes that the palaeographer needed to post-edit the HTR output of the 20 pages of each partition (except for the first block that was transcribed without any help). For each partition, the HTR system was trained with all the pages in the preceding partitions. The value in parentheses in column *Time* is the average number of minutes that the palaeographer needed to post-edit a marriage license in each partition. Column *WER* is the Word Error Rate for each partition. Note that this value can be interpreted as the percentage of corrections (deletions, insertions and substitutions) that the palaeographer needed to obtain the correct transcription. Column *% Running OOV* is the percentage of out of vocabulary words in each partition that were not observed in the training. The HTR system was not able to provide the correct transcription of these words because they were not included in the language model and therefore these errors were unavoidable for the HTR system. In other words, if we had available a lexicon, then the WER in the OOV words could be reduced at most in the amount represented in the *% Running OOV* column.

Regarding the results, it is important to remark the following issues. First, as expected, the

Table 3: Post-editing results.

	Time	WER	% Running OOV
P1 (1-20)	393 (3.4)	100	100
P2 (21-40)	305 (2.7)	57.7	15.9
P3 (41-60)	235 (2.1)	45.0	11.6
P4 (61-80)	193 (1.7)	35.7	11.6
P5 (81-100)	170 (1.5)	36.2	8.1

WER decreased as the amount of training data increased. In particular, the system achieved around 36% of WER for the last two partitions. This can be also observed in Figure 3, where the results are shown graphically.

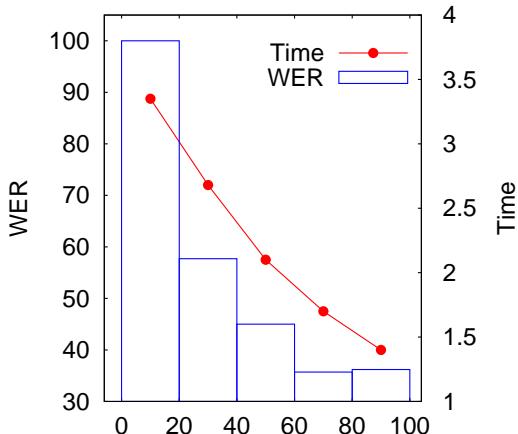


Figure 3: Transcription Word Error Rate (WER) and post-editing time (in minutes) as a function of the block of pages transcribed. For each block the HTR system is trained with all the pages in preceding blocks.

Second, regarding the time required by a human expert to post-edit the transcription proposed by the system, results showed that it became better with the number of partitions already processed. Most specifically, the relative difference between manually transcribe the first block with respect to post-edit the last block is 56%.

It is important to remark that during the transcription process the palaeographer learned to transcribe as new pages were processed. Therefore the time reduction for transcribing could be due to: i) the help of the HTR system, ii) the palaeographer’s learning process, or iii) both of them. In order to clarify this issue, we plotted the

average time that was needed to transcribe each license by page grouped by partitions, and then we fitted these times to a function. The gradient of this function may be interpreted as the “tendency” of the time needed to transcribe a page (see Figure 4). If the gradient is near to 0, then this could be interpreted as the palaeographer needed similar time to transcribe the initial pages of the partition and the final pages, and therefore the improvements in time are mainly due to the HTR system. This happened in the last two partitions.

In the first partition, the gradient was positive. This may be interpreted as the palaeographer was learning to transcribe and some pages were difficult. In partitions 2 and 3, the gradient was negative; that may be interpreted as the palaeographer was taking profit of the experience acquired in previous pages, and he was learning as new pages were post-edited.

Although we think that there is room for significant improvements, it must be noted that the results are reasonably good for effective post-editing.

6 Conclusions

In this paper we have studied how the accuracy of the HTR systems is reduced with respect to the amount of data used to train the models. In addition, we have also studied how the improvements in the system accuracy affect to the human efficiency following a post-editing approach. The experiments have been carried out with a marriage license book. These documents have interesting information that is being used by demographers that devote a lot of time to transcribe them.

Considering the results obtained in the field study, we can conclude that post-editing the output of an automatic transcription system, significant amounts of human effort can be saved.

Acknowledgements

Work supported by the Spanish Government (MICINN and “Plan E”) under the MITTRAL (TIN2009-14633-C03-01) research project and under the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), the Generalitat valenciana under grant Prometeo/2009/014 and FPU AP2007-02867 and

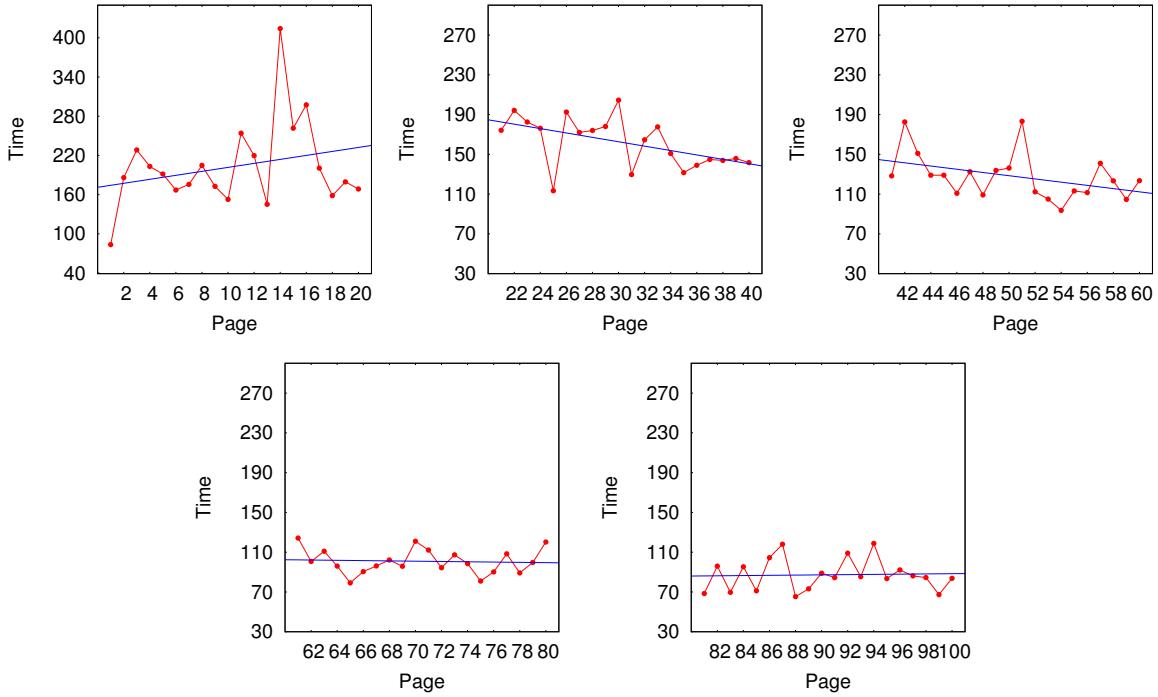


Figure 4: Average time (in seconds) that was needed to post-edit each license by page, and linear function fitted to this time.

by the Universitat Politècnica de València (PAID-05-11).

References

- I. Bazzi, R. Schwartz, and J. Makhoul. 1999. An Omnipoint Open-Vocabulary OCR System for English and Arabic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(6):495–504.
- F. Drida. 2006. Towards restoring historic documents degraded over time. In *Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pages 350–357. Lyon, France.
- S. España-Boquera, M.J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez. 2011. Improving offline handwriting text recognition with hybrid hmm/ann models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):767–779.
- A. Esteve, C. Cortina, and A. Cabré. 2009. Long term trends in marital age homogamy patterns: Spain, 1992-2006. *Population*, 64(1):173–202.
- A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.
- F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press.
- E. Kavallieratou and E. Stamatatos. 2006. Improving the quality of degraded document images. In *Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pages 340–349, Washington DC, USA.
- R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. volume 1, pages 181–184, Detroit, USA.
- U.-V. Marti and H. Bunke. 2001. Using a Statistical Language Model to improve the performance of an HMM-Based Cursive Handwriting Recognition System. *Int. Journal on Pattern Recognition and Artificial Intelligence*, 15(1):65–90.
- V. Romero, M. Pastor, A. H. Toselli, and E. Vidal. 2006. Criteria for handwritten off-line text size normalization. In *Proc. of the 5th Int. Conf. on Visualization, Imaging and Image (VIIP 2006)*, Palma de Mallorca, Spain, August.
- N. Serrano, L. Tarazon, D. Pérez, O. Ramos-Terrades, and A. Juan. The GIDOC prototype. In *Proceedings of the 10th PRIS 2010*, pages 82–89, Funchal (Portugal).
- A. H. Toselli et al. 2004. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *Int. Journal on Pattern Recognition and Artificial Intelligence*, 18(4):519–539.

bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...

Towards lexical link-up for a corpus of Old Swedish

Yvonne Adesam Malin Ahlberg Gerlof Bouma

Språkbanken

Department of Swedish

University of Gothenburg

firstname.lastname@gu.se

Abstract

We present our ongoing work on handling spelling variations in Old Swedish texts, which lack a standardized orthography. Words in the texts are matched to lexica by edit distance. We compare manually compiled substitution rules with rules automatically derived from spelling variants in a lexicon. A pilot evaluation showed that the second approach gives more correct matches, but also more false positives. We discuss several possible improvements. The work presented is a step towards natural language processing of Old Swedish texts.

1 Introduction

Corpus linguistics makes it possible to explore many interesting questions by investigating large amounts of text. These texts generally have some form of annotation, or mark-up. Språkbanken,¹ the Swedish language bank, has a large collection of modern texts with close to a billion words, which have been automatically annotated and are searchable through the corpus query interface Korp (Borin et al., 2012).² A current effort to cover different types of Swedish (Borin et al., 2010) involves older variants of Swedish texts, from the 19th century back to the 13th century. We now focus on Old Swedish, which covers two stages of Swedish: Early Old Swedish (1225-1375), and Late Old Swedish (1375-1526). Ultimately, our goal is to develop tools for various levels of annotation, such as part-of-speech, morphosyntactic information, and dependency parses.

A number of issues are problematic for annotation of Old Swedish texts. For example, sentence

splitting cannot be handled with standard tools, as sentence boundaries are often not marked by punctuation or uppercase letters. Compared to modern Swedish texts, the Old Swedish texts have a different vocabulary and richer morphology, show a divergent word order, and Latin and German influences. Finally, the lack of a standardized orthography results in a wide variety of spellings for the same word. In the following, we will discuss our first efforts to handle such spelling variations.

2 Resources

Our source material comes from Fornsvenska textbanken,³ a collection of around 160 digitized texts totaling 3 million words, mainly from the 13th to the 16th century. The data set consists of novels, poems, laws, and religious texts, ranging from 200 to 200,000 words in length. The texts thus vary in age, language, size, and type.

Språkbanken develops and makes available a range of lexical resources (Borin et al., 2010). The central resource for Contemporary Swedish is SALDO (Borin et al., 2008). Dalin's dictionary (1853/1855) covers 19th century Swedish. Additionally, we have three dictionaries for Old Swedish: Söderwall and Söderwall's supplement (1884/1953) with 44,000 entries, and Schlyter (1887) with 10,000 entries, focusing on law texts. There is overlap between these three dictionaries. Finally, a morphology for modern Swedish has been developed at Språkbanken and morphologies for 19th century Swedish and Old Swedish are under development (Borin and Forsberg, 2008). While the work presented only uses the Old Swedish dictionaries and corpora, we intend to also use the morphology in the future.

¹<http://spraakbanken.gu.se>

²<http://spraakbanken.gu.se/korp/>

³<http://project2.sol.lu.se/fornsvenska/>

3 Spelling Variation

A problem that severely hampers processing of historical text is the lack of a standardized orthography. A common solution is to normalize spelling to a modern language variety (Pettersson et al., 2012; Bollmann et al., 2011; Jurish, 2010a). In contrast, we intend to retain the original orthography in order to use the lexical and morphological resources available. Thus, we investigate mapping forms occurring in the corpora to the entries in the Söderwall and Schlyter dictionaries. This linking is motivated by the following: the dictionary entries can serve as type identifiers, which facilitate statistical processing; we gain access to the (partial) information about part-of-speech provided in the lexica; and finally, from a user-interface perspective, the link can serve as an assisting technology for studies of the Old Swedish texts.

Initial experiments suggest that only roughly one quarter of the types in unprocessed text match exactly against a lexical entry in one of the dictionaries. We therefore investigate approximate string matching to raise the coverage of the dictionaries. In the following, we give a general outline of our method and report on a pilot evaluation. As this is work in progress, we mainly focus on the qualitative findings of this evaluation and discuss directions for future work.

3.1 General Approach

A lexical entry is considered to be a match for a token in the text if it is the best candidate in terms of edit distance below a given threshold. We only allow the substitution operation, as this simplifies the implementation and the automatic collection of rules. Substitution rules can contain strings of unequal length.

Calculating the edit distance between a source word and each of the about 54,000 target dictionary entries is too expensive with million word corpora. We therefore use an *anagram hashing* filter (Reynaert, 2011) to cut down the number of exact edit distance calculations. The filter computes character based hashes for the source words and for the target entries. Character edits such as substitution can be performed numerically on these hashes. After applying a numerical edit on a source hash, we have a new hash value which we can compare against the hash table representing

the dictionary or apply further edits to. The hash function loses information about the order of characters in a string – hence the name *anagram hash* – which means that the filter underestimates the actual number of required edits between source and target is. In the experiments presented below, the filter generates, for each source word, up to 1 million hash variations with an increasing number of edits. From the target dictionary entries found in this candidate set – a much smaller number – we select the best matches using exact edit distance calculation. In its current implementation, the filter may discard valid matches, a problem we shall return to in Section 3.4.

3.2 Substitution Rules

We created sets of substitution rules in two ways. First, we composed a set of 31 correspondences by hand, based on common alternations appearing in the Old Swedish texts. An illustration of these rules is given in Fig. 1 (left). The rules have a constant weight, which makes them more costly than identity mapping but cheaper than a substitution not defined in the rule set.

Secondly, we used character aligned spelling variants to derive two further substitution rule sets. The variants are taken from the Schlyter dictionary, where 1,478 entries have an average of 2.2 variants. Note that the dictionaries do not share all orthographic conventions, and we cannot assume that all possible spellings are listed for an entry.

The spelling variants were aligned with iterated Levenshtein distance alignment (Wieling et al., 2009), using weighted substitution, insertion or deletion operations. The operations initially have unit cost, but after a complete iteration we use the alignments to calculate new weights, which are then used to realign the variant pairs. The weights converge in 3 or 4 iterations, resulting in fewer alignments, which are of higher quality. For instance, the pair *alita-ålijthe* ‘be assured’ can be aligned in the following two ways if all operations have unit cost:

1	a	l	i	t	a		
	å	l	i	j	h	e	
2	a	l	i	t	a		
	å	l	i	j	t	h	e

After convergence of the alignment weights, however, only the former alignment is produced, as

Rule	Cost	Rule	Cost	Rule	Cost
v↔u	0.1	o↔u	0.14	u→o	0.20
v↔w	0.1	d↔þ	0.16	æ→e	0.27
u↔o	0.1	k↔g	0.24	pt→ft	0.31
y↔ö	0.1	y↔ö	0.24	g#→gg#	0.42
r#↔#	0.1	æ↔e	0.30	þer→n	0.43
þ↔t	0.1	å↔a	0.32	au→ö	0.44
þ↔th	0.1	þ↔t	0.32	th→þ	0.44
e#↔a#	0.1	z↔#	0.42	mp→m	0.45
#hv↔#v	0.1	r↔l	0.47	li→eli	0.45
f↔ff	0.1	r↔#	0.60	ghi→i	0.62

Figure 1: Example rules: hand-written (left), automatically derived character substitution (mid) and n-gram substitution (right). # marks word boundary.

$a \leftrightarrow e$ is a cheaper substitution than $a \leftrightarrow h$.

From these alignments, we first extracted simple 1–1 substitution rules by collecting all character substitutions that were observed more than once in the data set. This resulted in 106 rules – a sample is found in Fig. 1 (mid). The costs are taken directly from the iterative Levenshtein distance alignment procedure. These 1–1 substitutions can only be used to match words of equal length. We thus extracted a more comprehensive set of substitutions by collecting all uni-, bi- and tri-gram alignments occurring more than once in the dataset. Because an n-gram in an alignment may contain ϵ -s, we not only create n–n rules, but also n–m rules. The cost associated with a substitution is $-\log p(\text{RHS}|\text{LHS})$, scaled down by a factor 10 to bring it in line with the costs of the other rule sets. Conditional probabilities were estimated using simple additive smoothing. Some examples of the resulting 6,045 rules are given in Fig. 1 (right).

3.3 Pilot Experiments

To get an impression of the coverage and accuracy of our approximate match approach, we applied our system to a small test collection consisting of a fragment of law text (Scania Law, Holm B 76) and a fragment of bible text (Marcus Gospel, 1526), together 249 tokens and 168 types based on string identity. The test corpus is too small to support more than a very coarse impression of the accuracy of our approach. However, the pilot evaluation is intended to give insight into the potential and challenges of our approach. We are planning a larger corpus to facilitate future development and

Method	Match		Correct		Top 3		Top 10	
	#	%	#	%	#	%	#	%
exact match	70	28	55	22				
manual rules	147	59	118	47	122	49	122	49
auto char	197	79	116	47	131	53	134	54
auto n-gram	240	96	154	62	195	78	208	84

Table 1: Results of the pilot evaluation on 249 tokens.

allow more thorough evaluation. Because of the spelling variation, the manual annotation of the link-up is non-trivial.

For each word in the text, we checked whether the algorithm finds a match in one of the dictionaries, and whether this match is correct. We count a match as correct if the intended meaning of the word in its context is among the entries selected from the dictionary. We do not distinguish between homographs at this point, i.e., words with different sense or part-of-speech tag. We also count entries that refer to a correct entry as correct. For instance, we consider **aer.X**⁴ in the Schlyter dictionary a correct match for *aer* ‘is’, as it refers to **vara.V**, which is the main entry for the verb *to be*. Such references are rare, though. Finally, we also applied a relaxed notion of correctness, using an oracle, where we consider a word as correctly linked if it is among the top three or top ten candidates. Note that we use the term *coverage* for both overall matches and correct matches. We avoid using the term *recall*, as we do not know how many words have a correct entry in the dictionary to begin with.

The results of the evaluation are in Table 1. As we can see, less than a third of the tokens in the texts had an exact match in one of the three dictionaries, and only counting correct exact matches, coverage is 22%, giving a precision of 79% (=55/70).⁵ The manual rules roughly double this coverage, both in terms of overall matches and correct matches. Almost half of the tokens in the texts are now matched against the correct lexical entry. Note that the top 3 and top 10 oracles do

⁴Boldface indicates lexical entries. The following capital indicates part-of-speech as given by the dictionary, where X means unspecified.

⁵Considering that the used lexica are partly built upon these corpora, this is a low number. However, we match inflected forms against dictionary entries that are mostly in their base form.

not really increase the number of correct matches, as the rules only allow a very restricted amount of variation. Precision of the manual rules remains at 80% (= 118/147).

The automatically extracted character mapping fails to improve upon the correct match score of the manual rules, and precision drops to 59% (= 116/197). These rules thus create many misleading matches, compared to the manual rules. Finally, the automatically extracted n-gram rules find a match for almost all words in the text, and moreover, retrieve the greatest number of correct matches at 62% of all tokens. This, however, comes at a cost of low precision: 64% (= 154/240). When using the top k oracles the precision for the automatically extracted n-gram rules is instead encouragingly high at 81% (= 195/240) and 87% (= 208/240), respectively.

3.4 Evaluation and Future Work

One of the reasons that the automatically extracted n-gram method has such high overall coverage is that the substitution rules not only capture spelling variation, but inflection, too. For instance, for the inflected *öknen* ‘desert.DEF.SG’ we find the correct match **ökn.N**, a match we would have otherwise missed even though the spelling convention happens to be the same between source and target. In future work, we intend to treat morphology more systematically by incorporating the computational morphology mentioned in Section 2. For instance, we could use the morphology to lemmatize the text, after which the lemmata are linked-up to the dictionaries using our edit distance approach.

The top k oracles for the automatically extracted n-gram rules combine high coverage with high precision. We hope to be able to capitalize on this potential in the future by using context information to model the oracle. For instance, for *stijghar* in the sentence *görer hans stijghar retta* ‘straighten his paths’, we find the verb **stigha.V** ‘ascend, walk’ as a closer match than the noun **stigher.N** ‘path’. Looking at the context, however, we may be able to guess that the nominal entry is more likely to be correct (see Wilcox-O’Hearn et al. (2008), Jurish (2010b) for the use of word level context in similar selection tasks).

Another avenue for research lies in the anagram hash filter. At the moment, we consider one mil-

lion operations on the anagram hash, submitting on average just under 900 candidate matches for exact edit distance calculation. However, since we have over 6,000 substitution rules, many of which may apply for a particular word, even one million hash variations may only represent a fraction of the search space. When an even smaller number of operations is considered – attractive from a processing effort perspective – the filter removes many valid alternatives, either resulting in suboptimal top candidates or empty candidate lists. For instance, *cristindom–kristindomber.N* ‘christianity, baptism’ is found by the top 3 oracle in the 1 million operations, but not in the 10k filter. Such incorrectly filtered source-target pairs are typically long and either allow many different operations or require repeated applications of the same cheap rules. Resolving this mismatch between the pre-filter step and the exact distance calculation is part of ongoing work.

We found that, in addition to linking the tokens to entries in the lexica, the entries need to be linked between the lexica. There is overlap between the dictionaries, but the different dictionaries give different information. For instance, *aer* ‘is’, mentioned above, is spelled *aer* in the law text, but *är* in the bible text. Only the Schlyter entry **aer.X** links this inflected form to **vara.V**. The corresponding entries in the Söderwall dictionaries, **är.X**, give only noun, conjunction and pronoun uses. Consolidating the dictionaries will increase the amount of information available.

Finally, the evaluation revealed the necessity to be able to handle multi-word tokens. For example, the text contains the expression *köpæ iorth* ‘bought land’. While the word *iorth* is correctly matched in the lexicon, the lexicon entries found for *köpæ* are incorrect, as they refer to the verb buying or the people involved in buying. The entry **köpe iorþ.N** does appear in two of the lexica but cannot presently be found as we only consider graphic words. There are also cases where we need to split a graphic word to match it against a dictionary. For instance, the parts of the compound *villhonugh* ‘wild honey’ can separately be matched against **vilder.AV** ‘wild’ and **hunagh.N** ‘honey’.

Splitting and merging compounds and matching them against the dictionary can be readily integrated by allowing substitution rules to contain

graphic word boundaries and considering multiple source tokens at one time (Reynaert, 2011). An effective way of doing separate matching as needed in the case of *vill-honnugh* remains a question for future work.

4 Conclusions

We are working towards automatic annotation of about 3 million words of historical Swedish texts. As a first step, we are developing a module to handle spelling variation. Three sets of approximate matching rules were used, one with hand-crafted substitutions, and two with substitutions automatically extracted from alternative entries in a lexicon. We presented a pilot evaluation by matching words in two short texts to our historical lexica. While the automatically created rules find more matches, the manual rules have higher precision. Future work includes improving the anagram hash filter, incorporating Old Swedish morphology, handling multi-word units, and exploiting context information to improve precision.

Acknowledgements

The research presented here is carried out in the context of the Centre for Language Technology of the University of Gothenburg and Chalmers University of Technology.

References

- Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the RANLP Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42. Hissar, Bulgaria.
- Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech, Morocco. ELRA.
- Lars Borin, Markus Forsberg, and Lennart Lönnqvist. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. In Nivre, Dahllöf, and Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, volume 7 of *Studia Linguistica Upsaliensia*, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.
- Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010. Diabase: Towards a diachronic BLARK in support of historical studies. In Calzolari, Choukri, Maegaard, Mariani, Odijk, Piperidis, Rosner, and Tapias, editors, *Proceedings of the LREC 2010 workshop on Semantic relations. Theory and Applications*, Valletta, Malta. ELRA.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In Calzolari, Choukri, Declerck, Doğan, Maegaard, Mariani, Odijk, and Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. ELRA.
- Anders Fredrik Dalin. 1853/1855. *Ordbok öfver svenska språket*, volume I-II. Stockholm, Sweden.
- Bryan Jurish. 2010a. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77, Uppsala, Sweden. Association for Computational Linguistics.
- Bryan Jurish. 2010b. More than Words: Using Token Context to Improve Canonicalization of Historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–40.
- Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Parsing the past - identification of verb constructions in historical text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.*, Avignon, France.
- Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187. 10.1007/s10032-010-0133-5.
- Carl Johan Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar*, volume 13 of *Saml. af Sweriges Gamla Lagar*. Lund, Sweden.
- Knut Fredrik Söderwall. 1884/1953. *Ordbok Öfver svenska medeltids-språket. Ordbok Öfver svenska medeltids-språket. Supplement*. Lund, Sweden.
- Martijn Wieling, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In Borin and Lendvai, editors, *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009) Workshop at the 12th Meeting of the EACL*, pages 26–34.
- Amber Wilcox-O’Hearn, Graeme Hirst, and Alexander Budanitsky. 2008. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *LNCS*, pages 605–616. Springer.

Automatic extraction of potential examples of semantic change using lexical sets

Karin Cavallin

Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Sweden
karin.cavallin@gu.se

Abstract

This paper describes ongoing work on automatically finding candidates for semantic change by comparing two corpora from different time periods. Semantic change is viewed in terms of distributional difference with a computational and linguistically motivated approach. The data is parsed, lemmatized and part of speech information is added. In distributional semantics, meaning is characterized with respect to the context. This idea is developed from Firth (1957) and is formulated according to ‘the distributional hypothesis’ of Harris (1968). A method is developed to describe distributional behaviour in order to track semantic change over time. We will explore statistically ranked lists of *verbal predicate - nominal object* constructions and examine differences at the level of word types.

1 Introduction

When I asked a lexicographer how he finds *semantic change* the answer was “I read a lot”. This method, however pleasant, might not be the most efficient way, and we here propose a way of extracting candidates for semantic change automatically, in the hope that this gives the lexicographers more time to do the proper analysis of the candidates instead of time-consuming reading. The main idea is to make a quantitative study in order to automatically find candidates for semantic change in two corpora, one of 19th century data and one of data from 1990’s. In this exploratory stage of the project, we work with transitive verb predicates and nominal objects.

A distributional standpoint is adopted, in that meaning is characterized in terms of the context in which words occur. The distributional hypothesis is attributed to Harris (1968) but the most famous quote representing this view is by Firth (1957): “You shall know a word by the company it keeps”¹. The distributional hypothesis is not uncontroversial, but for computational means this assumption has proven fruitful in several tasks.

The data used for this project consisted of 19th century Swedish literary texts from *Litteraturbanken* (the Swedish Literature Bank) (LB) and parts of the Swedish Parole corpus (PA) with text from the 1990’s.

We expect that a comparison of the ranks and other differences in distribution, both on frequency level and type level, will yield information that will pick out candidates for semantic change. We use the terms “type” and “token” as it is standard in corpus linguistics. Whereas *token* refers to number of word occurrences, *type* refers to the (orthographic) representation of a word group.

1.1 Related work

The field of semantic change seems to have received little attention in natural language processing (NLP). There has been a small amount of research in the last few years. Sagi et al. (2009) adapt latent semantic analysis and present a way of detecting semantic change based on a semantic *density* measure. The density is conveyed by the cohesiveness of the vectors. Cook and Hirst (2011) focus on finding amelioration and pejo-

¹A thorough survey of distributional semantics is found in Sahlgren (2006) and Sahlgren (2008).

ration. Gulordava and Baroni (2011), “address the task of automatic detection of the semantic change of words in quantitative way” focusing on detecting semantic change rather than specifically widening, narrowing or emotive change. Rohrdantz et al. (2011) “presents a new approach to detecting and tracking changes in word meaning by visually modelling and representing diachronic development in word contexts”[p.305].

The GoogleNgram-viewer (Michel et al., 2011), provides a quick and easy way of detecting changes in ngram frequencies, which can be interpreted for different aspects of cultural change. Lau et al. (2012) “apply topic modelling to automatically induce word senses of a target word, and demonstrate that [their] word sense induction method can be used to automatically detect words with emergent novel senses, as well as token occurrences of those senses.”[p.591].

Hilpert (2012) works with diachronic collocational analysis, which is mainly about semantic change in grammatical constructions. So far there is no consensus or standard way of approaching semantic change from a more quantitative perspective.

1.2 Lexical sets

We work with syntactically motivated collocational pairs, in this case verbal predicates and the head noun of object arguments, from here on ‘verb-object pairs’, we collect these together in *lexical sets*. Any set of lexical units that share a common feature can constitute a lexical set, be it phonological, ontological, orthographical, etcetera. Here, a lexical set can be a *verbal lexical set*, the verbs that occur as governing verb to a given nominal argument, or a *nominal lexical set*, the nouns that occur as argument to a given verb². The nominal and verbal lexical sets are arranged as ranked lists according to an association measure.

1.3 Log-likelihood

The ranking is based on a log-likelihood (Dunning, 1993) count and the verb-object pairs with stronger association are ranked higher. The log-likelihood implementation is taken from the

²The definition and the ranking of *nominal* and *verbal* lexical set follow the work of Jezek and Lenci (2007).

Ngram Statistic Package (Banerjee and Pedersen, 2003). “The log-likelihood ratio measures the deviation between the observed data and what would be expected if <word1> and <word2> were independent. The higher the score, the less evidence there is in favor of concluding that the words are independent.”³. The log-likelihood measure is applied to the extracted verb-object pairs in the corpora, thus providing us with a ranking such that the higher the ranking, the less likely it is that the words are independent within the respective pair, that is they have a stronger association with each other. This measure has primarily been used for collocation extraction and therefore seems appropriate to verb-object pairs.

The data here is summarized in ranked lists from the perspective of the predicate or argument, respectively. Senses will be analysed through *frequency* and a log-likelihood based *ranking* and also the difference in number of different *types* a word co-occurs with in its lexical set.

2 Preparing the data

2.1 The data

In order to track semantic change over time there is a need for corpora containing material from different time periods. For modern Swedish we use a selected subset of the Parole corpus (Språkbanken, 2011). The subcorpus we use contains novels, magazines, press releases and newspaper texts from the 1990s. For the older Swedish we use the Swedish Literature Bank (Litteraturbanken, 2010), a resource not adapted for NLP purposes, but intended for people with literary interests. The Swedish Literature Bank carries material from as early as the 13th century, but for the present study only the 19th century texts have been tagged, parsed and lemmatized.

The subcorpus of the Swedish Literature Bank we are building amounts to approximately 10 million word tokens. Of these approximately 2 million are tagged as nouns, and 1.5 million as verbs⁴. The subcorpus of Parole amounts to approximately 8 million tokens, whereof 1.4 million

³<http://search.cpan.org/~tpederse/Text-NSP-1.25/lib/Text/NSP/Measures/2D/MI/11.pm>, 2012-08-16

⁴Accuracy has not been estimated yet.

are words tagged as nouns, and 1.3 million tagged as verbs.

Table 1: No. of nouns, verbs and extracted verb-object pairs in the datasets.

	LB	PA
Word tokens	10M	8M
Noun tokens	2M	1.4M
Verb tokens	1.5M	1.3M
Extracted VO pairs tok	290.878	482.221
Extracted VO pairs typ	157.869	97.077

We see that we have been able to extract more VO pairs from the PA corpus, despite the fact that it is slightly smaller. However, the number of different pair-types also differs within the datasets. There is, for example, a much greater difference between the verb-object pair tokens than the verb-object pair types in the PA data set. This suggests that looking at types is a more promising starting point than raw frequency since that might be less sensitive to, for instance, over-representation.

2.2 PoS-tagger

The PoS-tagger used for the Swedish Literature Bank is Trigrams'n'Taggers, TnT (Brants, 1998), since its output is compatible with the MaltParser (see section 2.3). Good results are also attested in tagging texts containing many misspellings⁵ such as those of primary school students.⁶ Treating 19th century spellings as if they were misspellings is one heuristic way of addressing the problem of tagging 19th century Swedish. During the manual lemmatization (described in section 2.4), the PoS errors detected were omitted from the final data set.

2.3 Parser

One of the most important features in pursuing this sense tracking is to ensure that the corpus is parsed in order to identify predicates and objects. A parser freely available and widely used in the NLP community is the MaltParser (Nivre and Hall, 2005). The MaltParser is a system for data-driven dependency parsing. We have used a pre-trained Swedish model available from the

⁵Personal communication, Sofie Johansson Kokkinakis.

⁶As far as I know there is currently no account of why the TnT works well with misspellings.

MaltParser distribution. This model is of course trained on modern Swedish, which gives rise to noise in non-modern data, but we hope this is insignificant given the amount of data. In a recent paper, Pettersson et al. (2012) attempt to improve parsing by normalising their data from 1550-1880 (i.e. mostly by normalization spelling before tagging and parsing). We hope to take advantage of this result in future work.

2.4 Lemmatization

Some of the material was lemmatized automatically. The automatic lemmatization works as follows: There are two linked morphologies. One is originally a 19th century dictionary, Dalin (Dalin, 1850–1853), now enhanced with a full-form morphology⁷. The contemporary morphology is SALDO (Borin and Forsberg, 2008), an among other things full-form morphology lexicon. Both were developed at Språkbanken at Gothenburg University. Given the PoS information, which reduces ambiguity, the base form is extracted, and via the linking to SALDO, a contemporary base form (lemma) is extracted.⁸

However, in order to improve coverage, we performed a manual lemmatization on all the unlemmatized verbs and nouns in the LB corpus. The lemmatization is on the word form level and semantic ambiguities are not resolved. This is partly for practical reasons and partly to make the material as unbiased as possible with regard to sense, and also to avoid discussion of how fine-grained distinctions should be.

3 Method and Theory

The main data is basically a data set with information regarding ranked frequency occurrence from different perspectives and values computed on these frequencies by different statistical measures for the two data sets of extracted verb-object pairs of the Swedish Literature Bank and the Pa-role corpus.

The ranking can be made in different ways. By not ranking merely according to raw frequency,

⁷<http://spraakbanken.gu.se/eng/resource/dalin>

⁸This was carried out by Markus Forsberg, employing a similar lemmatization approach to that used, but not presented, in Borin et al. (2010) and Borin et al. (2011).

but adding a statistical measure to the data set (here a log likelihood count), we get a more reliable pattern where the most strongly associated pairs are ranked higher. This is a common approach for collocational extraction (Manning and Schütze, 1999) which this lexical set extraction shares similarities with. This data provides a basis for extracting viable candidates for semantic change.

It would be too ambitious to attempt to find *reasons* for semantic change automatically. What we should be able to find is the *consequences of semantic change*. We follow the distinction of consequences of semantic change of *widening*, *narrowing*, *amelioration* and *pejoration*, discussed by Bloomfield (1933). Whether a detected change is a widening, narrowing or a difference in emotive value is of course not detected by a difference in number, but needs manual lexical inspection to be determined. However, an increase or decrease in different types (as in item 3 below) should be a good indicator of a widening or narrowing, respectively. The latter two are not possible to perform computationally without a proper ontology. We want to consider the following aspects:

1. Increased or decreased raw frequency of a given *verb*, *object* or *verb-object pair*.
2. A higher or lower rank, given a statistical measure of a verb-object pair.⁹
3. Increased or decreased number of different types for a given verb or object.
4. Difference in the semantic type or class of the words in a lexical set.
5. A summarized difference of the semantic types of the respective lexical sets.

Differences in frequency between the two corpora is a first, but crude, indicator of semantic change. An ontologically motivated analysis would provide greater insights, but there is no full-coverage ontology available for Swedish. A more useful and theoretically motivated strategy is looking at the number of different types as shown in section 5.

⁹Here log-likelihood has been used, other association measures can also be applied.

4 Preliminary results

4.1 Lexical sets as means to finding semantic change

First attempts (Cavallin, 2012) showed distributional differences for what would be a widening. The noun “kontakt”, ‘contact’, must have undergone a widening, from merely a closeness between surfaces, into a connection between people. This is manifested as an increase of raw frequency (1878 occurrences in the later period versus 9 occurrences in the earlier), and also a higher rank in the lexical sets extracted from the Parole corpus (140 versus 7666). With respect to difference in type frequency we find a striking difference of 68 different verb (types) solely occurring in Parole (PA), five only occurring in the Literature Bank (LB), and two occurring in both. Thus information about raw frequency, ranking and type difference give us strong indications that there has been a change. If we look further at the lexical set of “kontakt” in Table 2, we also find that there are many words in the top fifteen that refer to “kontakt” in the sense of social connection, rather than the contact of surfaces or an electric plug. The last items on the list, which only occur in the Literature Bank seemingly mainly concern physical contact.

5 Type level differences as means to finding semantic change

In this next step we find a way of comparing lexical sets from a *type* perspective, and make this information give us a set of viable candidates for semantic change.

By extracting the elements in the datasets (mentioned in section 2) that differs the most, we get an overview of candidates for semantic change. The elements in the respective dataset that differ the most are computed by first normalizing the type counts, i.e. the number of different types each given verb/object governs or is argument to (this does not take frequency into consideration). By dividing the given count by the maximal type count of each set the values are normalized and hence comparable by taking the difference between the value of the PA data and the LB data. The greater difference on type level between a given word in the two datasets the higher the po-

Table 2: Lexical set of “kontakt”.

Trans	Verb	PaRank	LbRank
take	ta	140	-
have	ha	686	-
hold	hålla	1003	-
give	ge	1871	-
tie	knyta	2154	-
establish	etablera	2157	-
loose	förlora	2468	19175
find	finna	3684	-
come	komma	7134	-
loose	tappa	10118	-
miss	sakna	10806	-
avoid	undvika	10873	-
re-establish	återknyta	11533	-
get	få	11933	-
convey	förmedla	12332	-
...
seek	söka	13322	21363
...
connect	koppla	-	6423
screw	skruva	-	6907
Maintain	bibehålla	-	13852
see	se	-	39260

sition in the table (see Table 3). The word type displaying the maximal difference is then ranked highest under *DiffTyp*. *DiffToken* is the difference of the normalized frequency, and does not constitute the rank in the given table. *LbSum* and *PaSum* refer to the unnormalized raw frequency in the data sets. *LbTyp* and *PaTyp* refer to the unnormalized number of types in the datasets.

A high position can indicate a widening, whereas a low position indicates a narrowing.¹⁰ Words in the middle of the type ranking should then not have experienced any major changes. Where to draw the line between what is a significant difference and thus a candidate for semantic change is not straightforward and left for future discussion.

The top word in Table 3, of words displaying the words that differ most regarding the number of types they occur with, is “procent” *percent*. Looking more thoroughly at the lexical sets of “procent” we see that the 12 verbs in the LB data are

¹⁰So far only widening has been explored.

Table 3: Examples of top objects differing in number of verb types.¹²

Transl.	Object	Lb/PaTyp	Lb/PaSum	DiffTyp/Tok
percent	procent	12/155	17/796	0.31/0.03
crown	krona	115/215	238/2611	0.28/0.03
problem	problem	27/140	39/3134	0.26/0.11

fairly vague, or semantically light (such as “kalla” ‘call’, “ha” ‘have’ , “göra” ‘do’, “få” ‘get’, “ta” ‘take’). The 155 different verbs occurring with “procent” in the PA data show many verbs that are semantically rich, for example: “kontrollera” ‘control’, “äga” ‘own’, “samla” ‘collect’, “producera” ‘produce’ and “utgöra” ‘constitute’. It appears that the word “procent” is used in a wider context in PA than in LB, where it occurs primarily in connection with money, a sense attested in the beginning of the 18th century. However, the more general notion of “procent” as part of a whole is also attested in the beginning of 18th century (SAOB, 1954)[p.1932ff]. Even though the sense of PART OF A WHOLE is listed in traditional dictionaries, the widening of the *usage* of this interpretation of “procent” is not noted as gaining in the traditional Swedish dictionaries.

“Krona”, as a unit of currency, was introduced in 1873, when the Scandinavian coin-union was created.¹³ Looking at the lexical set for “krona” ranked according to the difference of the normalized log likelihood values we see a great number of senses among the top ten which refer to MONEY rather than to something WORN ON THE HEAD. This shows us that counting types can point us in the direction of semantic change. However, we must always confirm candidates by manual inspection of the lexical set. This distributional difference for “krona” is definitely due to a non-linguistic historical change which caused the widening. Whereas some semantic change evolves slowly, this newer sense of “krona” has been very actively created and thus easily dated.

“Problem” has a major type and frequency increase. In the early 20th century “problem” is attested in compounds as in “problembarn”, ‘problem child’ (SAOB, 1954)[p.p.1927ff], where the notion is more of a psycho-social problem than

¹³krona. <http://www.ne.se/lang/krona/232211>, Nationalencyklopedin. Accessed July 13, 2012.

e.g. philosophical and political problems. The more concrete or easier form of problem as ‘difficulty’ or ‘trouble’ is more widely used in PA. The widened sense of “problem” has an impact on the frequency and number of types.

We see that words can have kept their original sense into the 20th century, and that even though the subsense has been co-occurring all along, it increases in usage, which we define as a widening of the sense, especially when the increased usage seem to be prevailing.

The approach is a promising contribution towards an automatic prediction of semantic change. It can suggest candidates for widening and narrowing. It would, however, benefit from more accurately parsed and tagged data. By combining the different indicators of semantic change in item 3 we would get even more reliable predictions.

6 Conclusion and future work

Differences in frequency is a first, but crude, indicator of semantic change. A more useful and theoretically motivated approach is looking at the number of different types as shown in the present paper, and assessing the candidates by looking at the appropriate lexical sets. (An ontologically motivated analysis would provide even greater insights, but there is no full-coverage ontology available for Swedish). Semantic change appears to reveal itself distributionally in different ways. Summarizing, it can be *measurably* manifested as:

- an increase in raw frequency (*Tokens*)
- differences in the number of types (*Types*)
- differences in the ranking of the words in the corresponding lexical sets
- differences in the lexical distribution in the corresponding lexical sets

It is important to note that not all distributional differences can be assumed to be semantic change *per se*. One must resort to manual inspection (and theoretical considerations) in order to confirm cases of semantic change. “Krona” is a semantic change brought about by a change in the world. However, there can also be socio-historical changes reflected in language. In our data we

have the example of the verb-object pair “läsa-bibel” ‘read-bible’. This has a much lower frequency in the modern data in comparison with for instance “läsa-tidning” ‘read-news paper’. What is reflected in the data is not that “läsa” has changed, but rather the fact that Sweden has gone through a secularization. At this level of analysis there is no obvious way to distinguish socio-historical change from semantic change by distributional means.¹⁴

It is important to remember that there are tagging errors where non-nouns (or non-objects) and non-verbs (or non-root predicates) have been tagged as nouns and verbs, especially in the older data. This can wrongly increase the number of different types in the LB data. If we combine the different measurable aspects mentioned above, the errors can hopefully be marginalized awaiting more accurately annotated data.

By comparing (fairly¹⁵) comparable corpora from different time periods we can be made aware of changes. Given a more fine-graded time distinction of the corpora, we could even attempt tracking where the sense starts being polysemous and where the new sense possibly exceeds the older sense in frequency.

The manually made lemmatization is a valuable resource in enhancing search in the 19th century material in the Swedish Literature Bank, and can be used for building better automatic lemmatization on other older Swedish material.

We would like to compare lexical sets where the given words are within semantically similar domains, which presumably will render further input in the pursuit of semantic change. We are also planning to compare the data taking the corresponding lexical sets and compute which lexical sets are the closest and most distant from each other.

¹⁴However, an ontological analysis could at least point us in the direction of where the predicate or argument is of a different semantic type. “Krona” would be distinguished as a candidate of semantic change, whereas “läsa” would not, since “bibel” and “tidning” are both PRINTED MATTER, and for instance the verbs in “trycka-krona” ‘coin-crown’ and “pryda-krona” ‘decorate-crown’ are ontologically distant.

¹⁵We are aware of the fact that the difference attested could be a difference only between the corpora, rather than between time periods.

The outcome of the present work is a starting point for automatically detecting semantic change. The approach gives us indications, but there is still a need for manual inspection, which we hope to decrease as resources and methods are refined. This approach is not restricted to Swedish, and would benefit from an attempt in a language with more elaborate language resources.

References

- S. Banerjee and T. Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, February.
- L. Bloomfield. 1933. *Language*. Henry Holt, New York.
- L. Borin and M. Forsberg. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, Göteborg universitet. <http://spraakbanken.gu.se/eng/saldo/>.
- L. Borin, M. Forsberg, and D. Kokkinakis. 2010. Dia-base: Towards a diachronic blark in support of historical studies. *LREC2*.
- L. Borin, M. Forsberg, and C. Ahlberger. 2011. Semantic search in literature as an e-humanities research tool: Conplisit – consumption patterns and life-style in 19th century swedish literature. In *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*, volume 11, pages 58–65.
- T. Brants. 1998. TnT - Statistical Part-of-Speech Tagging. <http://www.coli.uni-sb.de/~thorsten/tnt/>.
- K. Cavallin. 2012. Exploring semantic change with lexical sets. In *Proceedings of the 15th EURALEX International Congress*, pages 1018–1022.
- P. Cook and G. Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274, Singapore, December.
- A.F. Dalin. 1850-1855. *Ordbok Öfver svenska språket*, volume I, II. Svenska Akademien, Stockholm.
- T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- J.R. Firth. 1957. A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, pages 1–32.
- K. Gulordava and M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.
- Z. Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.
- M. Hilpert. 2012. Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In K-L. Allan and J.A. Robinson, editors, *Current Methods in Historical Semantics*, pages 133–160. De Gruyter Mouton, Berlin.
- E. Ježek and A. Lenci. 2007. When GL meets the corpus: A data-driven investigation of semantic types and coercion phenomena. In *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon*, Paris.
- J.H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France, April.
- Litteraturbanken. 2010. <http://litteraturbanken.se/>.
- C. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.
- J-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 14.
- J. Nivre and J. Hall. 2005. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005)*, pages 137–148, December.
- E. Pettersson, B. Megyesi, and J. Nivre. 2012. Parsing the Past - Identification of Verb Constructions in Historical Text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities*, Avignon, France.
- C. Rohrdantz, A. Hautli, T. Mayer, M. Butt, D.A. Keim, and F. Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 305–310, Stroudsburg, PA, USA. Association for Computational Linguistics.
- E. Sagi, S. Kaufmann, and B. Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings*

- of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, March.
- M. Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.
- M. Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):1–18.
- SAOB. 1954. *Svenska Akademiens Ordbok*. Svenska Akademien, Lund.
- Språkbanken. 2011. <http://spraakbanken.gu.se/korp/>.

Automatic classification of folk narrative genres

Dong Nguyen¹, Dolf Trieschnigg¹, Theo Meder², Mariët Theune¹

¹University of Twente, Enschede, The Netherlands

²Meertens Institute, Amsterdam, The Netherlands

{d.nguyen, d.trieschnigg}@utwente.nl

theo.meder@meertens.knaw.nl, m.theune@utwente.nl

Abstract

Folk narratives are a valuable resource for humanities and social science researchers. This paper focuses on automatically recognizing folk narrative genres, such as urban legends, fairy tales, jokes and riddles. We explore the effectiveness of lexical, structural, stylistic and domain specific features. We find that it is possible to obtain a good performance using only shallow features. As dataset for our experiments we used the Dutch Folktale database, containing narratives from the 16th century until now.

1 Introduction

Folk narratives are an integral part of cultural heritage and a valuable resource for historical and contemporary comparative folk narrative studies. They reflect moral values and beliefs, and identities of groups and individuals over time (Meder, 2010). In addition, folk narratives can be studied to understand variability in transmission of narratives over time.

Recently, much interest has arisen to increase the digitalization of folk narratives (e.g. Meder (2010), La Barre and Tilley (2012), Abello et al. (2012)). In addition, natural language processing methods have been applied to folk narrative data. For example, fairy tales are an interesting resource for sentiment analysis (e.g. Mohammad (2011), Alm et al. (2005)) and methods have been explored to identify similar fairy tales (Lobo and de Matos, 2010), jokes (Friedland and Allan, 2008) and urban legends (Grundkiewicz and Gralinski, 2011).

Folk narratives span a wide range of genres and in this paper we present work on identifying these genres. We automatically classify folk narratives as *legend*, *saint's legend*, *fairy tale*, *urban legend*, *personal narrative*, *riddle*, *situation puzzle*, *joke* or *song*. Being able to automatically classify these genres will improve accessibility of narratives (e.g. filtering search results by genre) and test to what extent these genres are distinguishable from each other. Most of the genres are not well defined, and researchers currently use crude heuristics or intuition to assign the genres.

Text genre classification is a well-studied problem and good performance has been obtained using surface cues (Kessler et al., 1997). Effective features include bag of words, POS patterns, text statistics (Finn and Kushmerick, 2006), and character n-grams (Kanaris and Stamatatos (2007), Sharoff et al. (2010)).

Finn and Kushmerick (2006) argued that genre classifiers should be reusable across multiple topics. A classifier for folk narrative genres should also be reusable across multiple topics, or in particular across story types¹. For example, a narrative such as *Red Riding Hood* should not be classified as a fairy tale because it matches a story type in the training set, but because it has characteristics of a fairy tale in general. This allows us to distinguish between particular genres, instead of just recognizing variants of the same story. In addition, this is desirable, since variants of a story type such as *Red Riding Hood* can appear in other genres as well, such as jokes and riddles.

¹Stories are classified under the same type when they have similar plots.

Most of the research on genre classification focused on classification of text and web genres. To the best of our knowledge, we are the first to automatically classify folk narrative genres. Our dataset contains folk narratives ranging from the 16th century until now. We first give an overview of the dataset and the folk narrative genres. We then describe the experiments, discuss the results and suggest future work.

2 Dataset

2.1 Overview

Our dataset is a large collection of folk narratives collected in the Netherlands². Although the collection contains many narratives in dialect, we restrict our focus to the narratives in standard Dutch, resulting in a total of 14,963 narratives. The narratives span a large time frame, with most of them from the 19th, 20th and 21st century, but some even dating from the 16th century as can be seen in Table 1³. Each narrative has been manually annotated with metadata, such as genre, named entities, keywords and a summary.

2.2 Narrative genres

Folk narrative genres vary between cultures. Legend, myth and folktales are major genres that are present in many cultures. Bascom (1965) proposed a formal definition of these genres, based on belief, time, place, attitude and principal characters of the narratives. In this work, we restrict our attention to genres that are applicable to the Dutch folk narratives. The selected genres are described below and based on how annotators assign narratives to genres in the Dutch Folktale database.

Fairy tales are set in an unspecified time (e.g. the well-known *Once upon a time ...*) and place, and are believed not to be true. They often have a happy ending and contain magical elements. Most of the fairy tales in the collection are classified under the Aarne-Thompson-Uther classification system, which is widely used to classify and organize folk tales (Uther, 2004).

²Dutch Folktale database: <http://www.verhalenbank.nl/>.

³Although standard Dutch was not used before the 19th century, some narratives dating from before that time have been recorded in standard Dutch.

Time period	Frequency
- 1599	8
1600 - 1699	11
1700 - 1799	24
1800 - 1899	826
1900 - 1999	8331
2000 -	4609
Unknown	1154

Table 1: Spread of data over time periods

Legends are situated in a known place and time, and occur in the recent past. They were regarded as non-fiction by the narrator and the audience at the time they were narrated. Although the main characters are human, legends often contain supernatural elements such as witches or ghosts.

Saint's legends are narratives that are centered on a holy person or object. They are a popular genre in Catholic circles.

Urban legends are also referred to as contemporary legends, belief legends or FOAF (Friend Of A Friend) tales in literature. The narratives are legends situated in modern times and claimed by the narrator to have actually happened. They tell about hazardous or embarrassing situations. Many of the urban legends are classified in a type-index by Brunvand (1993).

Personal narratives are personal memories (not rooted in tradition) that happened to the narrator himself or were observed by him. Therefore, the stories are not necessarily told in the first person.

Riddles are usually short, consisting of a question and an answer. Many modern riddles function as a joke, while older riddles were more like puzzles to be solved.

Situation puzzles, also referred to as *kwispels* (Burger and Meder, 2006), are narrative riddle games and start with the mysterious outcome of a plot (e.g. *A man orders albatross in a restaurant, eats one bite, and kills himself*). The audience then needs to guess what led to this situation. The storyteller can only answer with ‘yes’ or ‘no’.

Jokes are short stories for laughter. The collection contains contemporary jokes, but also older jokes that are part of the ATU index (Uther, 2004). Older jokes are often longer, and do not necessarily contain a punchline at the end.

Songs These are songs that are part of oral tradition (contrary to pop songs). Some of them have a story component, for example, ballads that tell the story of *Bluebeard*.

Genre	Train	Dev	Test
Situation puzzle	53	7	12
Saint's legend	166	60	88
Song	18	13	5
Joke	1863	719	1333
Pers. narr.	197	153	91
Riddle	755	347	507
Legend	2684	1005	1364
Fairy tale	431	142	187
Urban legend	2144	134	485
Total	8311	2580	4072

Table 2: Dataset

2.3 Statistics

We divide the dataset into a training, development and test set, see Table 2. The class distribution is highly skewed, with many instances of legends and jokes, and only a small number of songs. Each story type and genre pair (for example *Red Riding Hood* as a fairy tale) only occurs in one of the sets. As a result, the splits are not always even across the multiple genres.

3 Experimental setup

3.1 Learning algorithm

We use an SVM with a linear kernel and L2 regularization, using the liblinear (Fan et al., 2008) and scikit-learn libraries (Pedregosa et al., 2011). We use the method by Crammer and Singer (2002) for multiclass classification, which we found to perform better than one versus all on the development data.

3.2 Features

The variety of feature types in use is described below. The number of features is listed in Table 3. The frequency counts of the features are normalized.

I Lexical features. We explore *unigrams*, and *character n-grams* (all n-grams from length 2-5) including punctuation and spaces.

II Stylistic and structural features.

POS unigrams and bigrams (CGN⁴ tagset) are extracted using the Frog tool (Van Den Bosch et al., 2007).

⁴Corpus Gesproken Nederlands (Spoken Dutch Corpus), <http://lands.let.kun.nl/cgn/ehome.htm>

Punctuation. The number of punctuation characters such as ? and ", normalized by total number of characters.

Whitespace. A feature counting the number of empty lines, normalized by total number of lines. Included to help detect songs.

Text statistics. Document length, average and standard deviation of sentence length, number of words per sentence and length of words.

III Domain knowledge. Legends are characterized by references to places, persons etc. We therefore consider the number of *automatically tagged named entities*. We use the Frog tool (Van Den Bosch et al., 2007) to count the number of references to persons, organizations, location, products, events and miscellaneous named entities. Each of them is represented as a separate feature.

IV Meta data. We explore the added value of the manually annotated metadata: *keywords*, *named entities* and *summary*. Features were created by using the normalized frequencies of their tokens. We also added a feature for the manually annotated *date* (year) the story was written or told. For stories of which the date is unknown, we used the average date.

Feature type	# Features
Unigrams	16,902
Char. n-grams	128,864
Punctuation	8
Text statistics	6
POS patterns	154
Whitespace	1
Named entities	6
META - Keywords	4674
META - Named entities	803
META - Summary	5154
META - Date	1

Table 3: Number of features

3.3 Evaluation

We evaluate the methods using precision, recall and F₁-measure. Since the class distribution is highly skewed, we focus mainly on the macro average that averages across the scores of the individual classes.

	Precision	Recall	F₁
Unigrams	0.551	0.482	0.500
Char. n-grams	0.646	0.578	0.594

Table 4: Baselines (Macro average)

	Precision	Recall	F₁
All	0.660	0.591	0.608
-Unigrams	0.646	0.582	0.597
-Char. n-grams	0.577	0.511	0.531
-Punctuation	0.659	0.591	0.607
-Text statistics	0.659	0.589	0.606
-POS patterns	0.659	0.590	0.607
-Whitespace	0.660	0.591	0.608
-Domain knowl.	0.660	0.591	0.607

Table 5: Ablation studies, without meta data (Macro average)

4 Results

The penalty term C of the error term for SVM was set using the development set. All results reported are on the test set. We first report two baselines in Table 4, using only unigrams and character based n-grams. Results show that the character based n-grams are highly effective. This is probably because they are able to capture punctuation and endings of words, and are more robust to spelling mistakes and (historical) variations.

Next, ablation studies were performed to analyze the effectiveness of the different feature types, by leaving that particular feature type out. We experimented with and without metadata. The results without metadata are reported in Table 5. We find that only character n-grams contribute highly to the performance. Removing the other feature types almost has no effect on the performance. However, without character n-grams, the other features do have an added value to the unigram baseline (an increase in macro average of 0.50 to 0.53 in F₁ score). One should note that some errors might be introduced due to mistakes by the automatic taggers for the POS tokens and the domain knowledge features (named entities), causing these features to be less effective.

The results with all features including the metadata are reported in Table 6. We find that when using all features the F₁ score increases from 0.61 to 0.62. The ablation studies suggest that especially the keywords, summary and date are effective. However, overall, we find that only using character n-grams already gives a good perfor-

	Precision	Recall	F₁
All	0.676	0.600	0.621
-META - Keywords	0.659	0.595	0.611
-META - Named Entities	0.682	0.599	0.623
-META - Summary	0.664	0.596	0.614
-META - Date	0.674	0.592	0.614

Table 6: Ablation studies all features (Macro average)

	Precision	Recall	F₁
Sit. puzzle	0.70	0.58	0.64
Saint's legend	0.81	0.40	0.53
Song	0.00	0.00	0.00
Joke	0.93	0.71	0.81
Pers. narr.	0.69	0.52	0.59
Riddle	0.86	0.83	0.84
Legend	0.82	0.92	0.86
Fairy tale	0.69	0.53	0.60
Urban legend	0.59	0.91	0.71

Table 7: Results per genre

mance. We therefore believe they are a valuable alternative against more sophisticated features.

We also find that including the date of a narrative as a feature leads to an increase from 0.614 to 0.621. This feature is effective since some genres (such as urban legends) only occur in certain time periods. Noise due to the many documents (1154 of 14963) for which the date is not known, could have affected the effectiveness of the feature.

In Table 7 the results per genre are listed for the run including all features (with metadata). The best performing genres are jokes, riddles and legends. We find that songs are always incorrectly classified, probably due to the small number of training examples. Personal narratives are also a difficult category. These narratives can be about any topic, and they do not have a standard structure. Fairy tales are often misclassified as legends. Some of the fairy tales do not contain many magical elements, and therefore look very similar to legends. In addition, the texts in our dataset are sometimes interleaved with comments that can include geographical locations, confusing the model even more.

Initially, annotators of the Dutch Folktale database were allowed to assign multiple genres. From this, we observe that many narratives were classified under multiple genres (these narratives were excluded from the dataset). This is evidence that for some narratives it is hard to assign a single genre, making it unclear what optimal performance can be achieved.

5 Conclusion

Folk narratives are a valuable resource for historical comparative folk narrative studies. In this paper we presented experiments on classifying folk narrative genres. The goal was to automatically classify narratives, dating from the 16th century until now, as *legend*, *saint's legend*, *fairy tale*, *urban legend*, *personal narrative*, *riddle*, *situation puzzle*, *joke* or *song*. Character n-grams were found to be the most effective of all features. We therefore plan to explore historical texts in non standard Dutch, since character n-grams can be easily extracted from them. We also intend to explore features that help detect difficult genres and generalize across specific stories. For example, features that can detect humorous components.

6 Acknowledgements

This research was supported by the Folktales as Classifiable Texts (FACT) project, part of the CATCH programme funded by the Netherlands Organisation for Scientific Research (NWO).

References

- J. Abello, P. Broadwell, and T. R. Tangherlini. 2012. Computational folkloristics. *Communications of the ACM*, 55(7):60–70, July.
- C. O. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, pages 579–586.
- W. Bascom. 1965. The Forms of Folklore: Prose Narratives. *The Journal of American Folklore*, 78(307).
- J. H. Brunvand. 1993. *The Baby Train and Other Lusty Urban Legends*. W. W. Norton & Company.
- P. Burger and T. Meder. 2006. -A rope breaks. A bell chimes. A man dies.- The kwispel: a neglected international narrative riddle genre. In *Toplore. Stories and Songs*, pages 28–38.
- K. Crammer and Y. Singer. 2002. On the learnability and design of output codes for multiclass problems. *Mach. Learn.*, 47(2-3):201–233, May.
- R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9(6/1/2008):1871–1874.
- A. Finn and N. Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.
- L. Friedland and J. Allan. 2008. Joke retrieval: recognizing the same joke told differently. In *Proceedings of CIKM*, pages 883–892.
- R. Grundkiewicz and F. Gralinski. 2011. How to Distinguish a Kidney Theft from a Death Car? Experiments in Clustering Urban-Legend Texts. In *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition*.
- I. Kanaris and E. Stamatatos. 2007. Webpage Genre Identification Using Variable-Length Character n-Grams. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI '07, pages 3–10, Washington, DC, USA. IEEE Computer Society.
- B. Kessler, G. Nunberg, and H. Schuetze. 1997. Automatic Detection of Text Genre. In *Proceedings of the 35th ACL/8th EACL*.
- K. A. La Barre and C. L. Tilley. 2012. The elusive tale: leveraging the study of information seeking and knowledge organization to improve access to and discovery of folktales. *Journal of the American Society for Information Science and Technology*, 63(4):687–701.
- P. V. Lobo and D. M. de Matos. 2010. Fairy Tale Corpus Organization Using Latent Semantic Mapping and an Item-to-item Top-n Recommendation Algorithm. In *Proceedings of LREC*.
- T. Meder. 2010. From a Dutch Folktale Database towards an International Folktale Database. *Fabula*, 51(1-2):6–22.
- S. Mohammad. 2011. From once upon a time to happily ever after: tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 105–114.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.
- S. Sharoff, Z. Wu, and K. Markert. 2010. The Web Library of Babel: evaluating genre collections. In *Proceedings of LREC*.
- H.-J. Uther. 2004. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*. Vols 1-3. Suomalainen Tiedeakatemia, Helsinki.
- A. Van Den Bosch, B. Busser, S. Canisius, and W. Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational Linguistics in the Netherlands Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114. OTS.

The DTA ‘base format’: A TEI-Subset for the Compilation of Interoperable Corpora

Alexander Geyken
BBAW
Jägerstr. 22/23
D-10117 Berlin
geyken@bbaw.de

Susanne Haaf
BBAW
Jägerstr. 22/23
D-10117 Berlin
haaf@bbaw.de

Frank Wiegand
BBAW
Jägerstr. 22/23
D-10117 Berlin
wiegand@bbaw.de

Abstract

This article describes a strict subset of TEI P5, the DTA ‘base format’, which combines the richness of encoding non-controversial structural aspects of texts while allowing only minimal semantic interpretation. The proposed format is discussed with regard to other commonly used XML/TEI schemas. Furthermore, the article presents examples of good practices showing how external corpora can either be converted into the DTA ‘base format’ directly or after cautiously extending it. Thus, the proposed encoding schema contributes to the paradigm shift recently observed in corpus compilation, namely from private encoding to interoperable encoding.

1 Introduction

Up to the end of the 1990s corpus compilation on the basis of the *Guidelines of the Text Encoding Initiative* (TEI; most recent release: P5, cf. Burnard and Baumann, 2012) was mainly a project specific activity. Corpus documents were validated against a project specific document grammar, (possibly) private character encodings were used, and the documents were transformed into proprietary formats in order to get indexed for full text retrieval. In that era of project specific encoding, exchange of documents across projects was not a primary goal, and, in general, character encoding problems as well as differences in the document type grammar (DTD) were obstacles to a broader exchange of data. With the advent of XML and Unicode, documents encoded according to the recommendations of the

TEI became interchangeable, or, more precisely, documents encoded in TEI P5 could be safely exchanged on different platforms without worrying about incompatibilities of character encoding. However, differences in structural encodings still remained. The large flexibility of using the TEI Guidelines to encode similar semantic phenomena with different XML elements is one major reason for this problem: for example, there are several ways to encode the hierarchy of sections in documents, either with numbered division elements (`<div1>...<div7>`) elements or by enumerating the hierarchy with numeric @n-values: `<div n="1">`, `<div n="2">`, etc. Likewise, there are different ways to encode information about person names (`<persName>`, `<name type="person">`, `<rs type="person">`), several ways to link text passages (`<ref>`, `<ptr>`, `@corresp`, `@next/@prev`, ...) etc. However, the main reason for differences in structural encoding resides in the fact that different projects use different subsets of the TEI according to their needs. Problems like these become apparent when the attempt is made to carry out specific tasks with the exchanged data on another platform together with another document collection.

Problems occur on several levels, the first one being the difficulty to create a common style sheet across different document collections encoded in different TEI P5 schemas in order to present all document collections uniformly on the web. Another problem concerns the exchange of TEI metadata: Due to the flexibility of the TEI tag set, the structure of TEI Headers may dif-

fer considerably, which forces harvesting mechanisms exploiting this information in a uniform way to deal with a lot of different cases. This is obviously not the idea of a standard. Examples 1 and 2 illustrate this fact: the information about an author of a work can be either underspecified (Ex. 1) or very detailed (Ex. 2). However, both are valid according to the TEI Header specification.

Example 1

```
<author>Ernst, Ferdinand</author>
```

Example 2

```
<author>
  <persName>
    <forename>Ferdinand</forename>
    <surname>Ernst</surname>
  </persName>
</author>
```

Furthermore, machine-exploitable extraction of document components such as ‘retrieve all letters of the document collection’ or ‘display all quotations in a chapter’ pose an enormous problem since division types or entity encoding for quotes do not have to be realized in an ubiquitous way across document collections. Clearly the problem is even worse for complex XPath queries or for data mining tasks where ubiquitous encoding is a necessary prerequisite. To sum up: at present, document collections encoded in TEI can be exchanged only by accepting the loss of interoperability on one or several of the above-mentioned levels. These problems are widely acknowledged (cf. e.g. Ramsay et al., 2011: p. 1-4; Pytlík Zillig, 2009: p. 187 seq.; Unsworth, 2011: p. 1 seq.; Stührenberg, 2012: p. 141 seq.).

More recently, several attempts were made to increase the interoperability among different document collections by creating common formats. Therefore, subsets of TEI P5 were created reducing the tei_all tag set to a considerably smaller number of elements and attributes (cf. Day, 2010: p. 1). The TEI consortium recommends such customizations of the TEI inventory according to the individual needs of projects instead of taking the whole TEI tagset as a basis for the annotation of a corpus (Burnard and Baumann, 2012: ch. 15.5, 23.2). TEI formats like TEI Lite (Burnard and Sperberg-McQueen,

2006), TEI Tite (Trolard, 2011) or the *Best Practices for TEI in Libraries* (TEI SIG on Libraries, 2011; henceforth: TEI-Lib) are promoted.¹ In addition, several corpus and data curation projects have developed other TEI- or TEI-related formats according to their particular purposes, e.g. TEI Analytics developed by the MONK project (cf. Unsworth, 2011; Pytlík Zillig, 2009; Pytlík Zillig et al., 2012; henceforth: TEI-A), IDS-XCES by the Institute for the German Language in Mannheim and *TextGrid’s Baseline Encoding for Text Data in TEI P5* (TextGrid, 2007–2009; henceforth: TextGrid’s BE). These formats have been designed to allow for the basic structuring of all written texts and therefore serve as a starting point from which more detailed, possibly project specific text structuring could start.

The remainder of this paper starts with a short presentation of the above-mentioned subsets of the TEI (section 2). In section 3, we motivate the creation of a TEI format for the “Deutsches Textarchiv” (DTA), the DTA ‘base format’² (henceforth: DTA-BF). Section 4 presents examples of good practice illustrating how different external corpora can be converted into the DTA-BF, thus being interoperable in a wider context, e.g. as part of the text corpora provided by the large European infrastructure project CLARIN.³ We conclude with a short summary and some ideas about future prospects.

2 Comparison between existing TEI Encoding Formats

In this section, we compare some (well-known) existing XML annotation formats, which are fully or partially based on the TEI Guidelines, namely the above-mentioned formats TEI Tite, TEI Lite, TEI-Lib, TEI-A, IDS-XCES, and TextGrid’s BE. The formats are evaluated with respect to their applicability for the annotation of historical corpora such as the “Deutsches Textarchiv”.

All of the mentioned encoding formats have in common their attempt to unify large amounts of – possibly different – texts. However, considerable differences persist. TEI Tite and TEI-

¹Cf. www.tei-c.org/Guidelines/Customization.

²www.deutsches-textarchiv.de/doku/basisformat.

³Common Language Resources and Technology Infrastructure; www.clarin.eu.

Lib are complementary in the sense that they provide annotation guidelines for text digitization undertaken by libraries. While TEI Tite was created to allow for basic text structuring undertaken by external vendors, therefore intending “to prescribe *exactly* one way of encoding a particular feature of a document in as many cases as possible” (Trolard, 2011: ch. 1; Day, 2010: p. 16), TEI-Lib is intended “to support in-house encoding that adheres as closely as possible to common TEI practice and library standards yet still leaves room for variation in local practice” (TEI SIG on Libraries, 2011: ch. 2; cf. Dalmau/Schlosser, 2010: p. 355 seq.). Both formats are therefore especially suited to the task of annotating large amounts of heterogeneous text material in a library context. TEI Lite pursues a similar goal, being meant to “meet 90 % of the needs of 90 % of the TEI user community” (Burnard and Sperberg-McQueen, 2006: Prefatory note), but without being restricted to library usage. TEI-A results in a customization which is supposed to be suitable for the annotation of diverse texts from variable sources, as well, but has a different starting point than TEI-Lib, TEI Tite and TEI Lite, since it was created as a format to bring together texts which were already annotated individually (Pytlík Zillich, 2009: p. 188 seq.). Similarly, TextGrid’s BE is intended as basic encoding format enabling the intertextual search within TextGrid (TextGrid, 2007-2009: p. 6.). Finally, IDS-XCES serves as an encoding scheme for the IDS corpus texts. It is originally based on XCES, the XML adaption of CES, which was extended, partially with respect to the TEI Guidelines, according to the requirements of the IDS corpora (Institute for the German Language Mannheim, 2012; Stührenberg, 2012: p. 175-180).

Despite their individual genesis and purpose there is a set of structuring elements common to all of the named formats. E.g. the text of a document is divided into a <front>, a <body>, and a <back> area, paragraphs are structured as such using the element <p>, verse (at some point) should be encoded using <lg> and <l>, speech acts in a drama are encoded with the <sp> element etc. Such analogies show that there is a commonplace structuring level, which might be classified as level-1-encoding, or as, what the TEI

P5 *Recommendations for the Encoding of Large Corpora* subsume under “required” elements, demanding that “texts included within the corpus will always encode textual features in this category, should they exist in the text” (Burnard and Baumann, 2012: ch. 15.5). Still, in some cases the selections of TEI P5 elements differ. E.g. only TEI-A offers tagging solutions for screenplays, such as <view> and <camera>. Furthermore, the flexibility of the TEI specification allows that semantically similar phenomena are addressed differently by the encoding formats. E.g. TEI Lite, TEI Tite, and TEI-Lib allow for the encoding of additions and deletions which were performed on the source document by providing the elements <add> and , whereas TextGrid’s BE offers the elements <addSpan> and <delSpan> for this purpose.

The appropriate selection of elements is just one factor for the evaluation of annotation formats. Almost equally important is the appropriate choice of attributes and their corresponding values, ideally expressed as a fixed value set. In addition, there are more general factors to be taken into account with regard to the practical applicability in specific project contexts, namely the determination of annotation levels, solutions to the provision of metadata, comprehensive guidelines on text transcription and editorial interventions as well as the documentation of the format itself. Last but not least annotation formats differ in their degree of conformance to the TEI Guidelines - is the format a strict subset of TEI-P5 or does it make use of extensions.

Tables 1 and 2 summarize the commonalities and differences between the annotation formats considered here with respect to the above-mentioned factors. These factors serve as a guideline for the discussion of the DTA base format that is presented in the next section.

3 The DTA ‘base format’

This section discusses the DTA-BF, a customization of the TEI P5 tag set, created for the encoding of (historical) German text in large text corpora. The format emerged from previous work on a TEI P5 corpus project, the DWDS corpus (Geyken, 2007). The DWDS corpus is a cor-

	TEI P5 subset	documentation	element-wise attribute selection	fixed/recommended attribute values	levels
TEI Tite	no	yes, mainly element-wise	class-wise	no	no
TEI Lite	yes	yes	class-wise; some element-wise recommendations for attributes	no	no
TEI for Libraries	yes	yes	selection of generally recommended attributes	no	yes ^a
TEI Analytics	no	yes, element-wise; but examples include undocumented elements	yes	in some cases (e.g. recommended values for @unit and @part; fixed values for @scope)	no
TextGrid's Baseline Encoding	yes	yes, but examples include undocumented elements	in some cases (e.g. for inline elements)	in some cases	no
IDS-XCES	no	only changes to XCES are communicated; no documentation of the usage of elements	yes	in some cases	no
DTA ‘base format’	yes	yes	yes	yes	yes

^alevels are not strictly cumulative

Table 1: Comparison of annotation formats – part 1

pus of the 20th century German language of written text. It is roughly equally distributed over time and over five genres: journalism (approx. 27 % of the corpus), literary texts (26 %), scientific texts (approx. 22 %) and functional texts (approx. 20 %), as well as a smaller number of transcripts of spoken language (5 %). The focus of encoding was put on the non-controversial structural aspects of the documents with the goal to facilitate cross-document full text retrieval for linguistic purposes.⁴ With the start of the project *Deutsches Textarchiv*⁵ (DTA) in 2007, the TEI P5 compliant schema had to be extended considerably for two main reasons: faithful page per page presentation of the entire works, and the necessity to deal with older prints, thus having to cope with additional structural variation. The DTA project works on building a text corpus for the historical New High German. Within seven years of work, a selection of 1,300 texts of different text types, originating from the 17th to 19th century, are being digitized and annotated according to the TEI P5 Guidelines. Linguistic analyses are added to

the digitized text sources in a stand-off format for further corpus research.

The goal of the DTA-BF is to provide a homogeneous text annotation for a collection of historical texts being heterogeneous with respect to the date of their origin (1650–1900) and text types (literary texts, functional texts, scientific texts). To achieve this, the DTA-BF follows some overall restrictions, this way combining the different benefits of the named formats.

In the remainder of this section we show how the DTA-BF deals with the factors mentioned in section 2 ensuring the quality of corpora encoded according to the DTA-BF as well as the applicability of the DTA-BF for other projects.

3.1 Selection of Elements, Attributes, and Values

The selection of DTA-BF elements corresponds to a large extent to the tagset of TEI Lite. However, unlike TEI Lite, the DTA-BF also provides a restricted set of attribute values in order to minimize the possibility of using different tagging solutions for similar structural phenom-

⁴Cf. www.dwds.de.

⁵Cf. www.deutsches-textarchiv.de. The DTA is funded by the German Research Foundation between 2007 and 2014.

	inline metadata	solutions for editorial interventions ^a	transcription guidelines	text type specific encoding guidelines ^b
TEI Tite	no	no	no	newspapers
TEI Lite	yes	CN; AD-ST; AD-Ed (except <supplied>); AE	no	no
TEI for Libraries	yes	CN (except <reg>, <orig>); AD-ST; AD-Ed (except <supplied>)	instructions for quotation marks and hyphens	interviews
TEI Analytics	yes	CN; AD-ST (except); AD-Ed (no <gap>, <unclear>; <supplied> is not documented, but used in examples)	no	screenplays
TextGrid's Baseline Encoding	yes	CN; AD-ST (<addSpan>, <delSpan> instead of <add>,); AD-Ed (except <supplied>); AE	no	dictionary entries
IDS-XCES	no ^c	CN (except <choice>, <sic>; <corr> with @sic); AD-Ed (except <unclear>, <supplied>); AE (except <expan>; <abbr> with @expan)	no	spoken language (e.g. dialogues, speeches, debates, interviews)
DTA 'base format'	yes	CN; AD-Ed (except <unclear>; usage of <supplied> instead); AE	yes	funeral sermons, newspapers

^a I.e. correction and normalization (CN; includes <choice>, <sic>, <corr>, <reg>, <orig>); deletions, and additions: in the source text (AD-ST; includes <add>,), editorial (AD-Ed; includes <gap>, <unclear>, <supplied>); abbreviations and expansions (AE; <choice>, <abbr>, <expan>)

^b Other than prose, verse, drama, letter

^c A metadata format is provided, which contains TEI Header elements as well as a considerable amount of other elements

Table 2: Comparison of annotation formats – part 2

ena.⁶ This goal is explicitly expressed by the TEI Tite guidelines, as well.⁷ However, only some recommendations for attribute values are given, whereas no firm value lists are integrated in the TEI Tite schema. Other formats, such as TEI-Lib, explicitly decided against the restriction of attribute values.⁸

In our opinion, it is crucial to provide a detailed specification not only of elements but of corresponding attributes and values as well to mini-

⁶The necessity of minimal semantic ambiguity of the tagset has been pointed out by Unsworth (2011: § 7).

⁷“Tite is meant to prescribe exactly one way of encoding a particular feature of a document in as many cases as possible, ensuring that any two encoders would produce the same XML document for a source document” (Trolard, 2011: ch. 1).

⁸Cf. e.g. the statement of TEI-Lib about possible @type-values: “Constructing a list of acceptable attribute values for the @type attribute for each element, on which everyone could agree, is impossible. Instead, it is recommended that projects describe the @type attribute values used in their texts in the projects ODD file and that this list be made available to people using the texts” (TEI SIG on Libraries, 2011: ch. 3.8.1).

mize ambiguities of the tag set. Therefore, each of the 105 TEI P5 elements currently contained by the DTA-BF tagset⁹ is provided with a fixed list of possible attributes and values. The selection of attributes specified for each element is restricted not only class-wise but element-wise. Attribute values may occur within the DTA-BF in three different ways:

1. In general, the DTA-BF prescribes a fixed set of possible values for each attribute, thus being even more restrictive than TEI Tite. E.g. possible values for the @unit attribute of the element <gap> are: "chars", "words", "lines", or "pages". The selection of values in the DTA-BF can either apply for an attribute in every possible context or depend on the surrounding element.

2. In rare cases, where attribute values cannot

⁹I.e. about 80 elements used for the annotation of the texts themselves plus 25 additional elements needed specifically for the representation of metadata within the TEI Header.

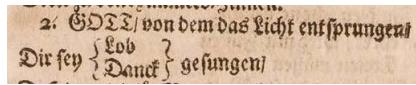
be reduced to a fixed set, restrictions are made with respect to the data type. E.g. the value of the attribute @quantity within the element <gap> has to be a non-negative integer (data.count).

3. Finally, there are cases, in which attribute values cannot be restricted by the schema at all. E.g. the value of @n in <pb> may consist of alphanumeric characters (e.g. <pb n="16">, <pb n="XVI">) as well as strings (e.g. <pb n=" [16] ">). In such (rare) cases value restrictions are given in the DTA-BF guidelines.

The DTA-BF has been designed to cover all annotation requirements for a basic structuring of the large variety of historical texts that are dealt with in the DTA. On this common structural level each typographically marked segment in the source text (centered, printed in bold or italics, printed with an individual typeface, smaller or larger letters) is labelled preferably with one basic semantic category (citation, poem, title, note etc.) or, if a semantic function cannot easily be assigned, with the formal category describing the typographical characteristics of the respective text segment. Fig. 1 shows the combination of semantic (<lg>, <l>) and formal (<list rendition="#leftBraced #rightBraced">, <item>) tagging.

3.2 Annotation Levels

As stated above, the DTA-BF is supposed to serve as a guideline for the homogeneous annotation of heterogeneous historical corpus texts. However, the necessity of a homogeneous format for the annotation of historical text seems to be opposed to the fact, that different projects usually have different needs as for how detailed text structuring should be. This problem is aggravated by the fact that text structuring becomes more labor intensive the more detailed it is. To address this problem, the TEI *Recommendations for the Encoding of Large Corpora* advise users who wish to create language corpora to define four levels of text annotation (required, recommended, optional, and proscribed) when determining a subset of TEI elements appropriate to the anticipated needs of the project (Burnard and Bauman, 2012: ch. 15.5).



```
<lg n="2">
  <head>2.</head>
  <l>GOTT/ von dem das Licht
    ent&#x17f; prungen/<l><lb/>
  <l>Dir &#x17f; ey
    <list rendition="#leftBraced
      #rightBraced">
      <item>Lob</item><lb/>
      <item>Danck</item>
    </list>
    ge&#x17f; ungen/<l><lb/>
    [...]
</lg>
```

Figure 1: Friedrich Rudolph Ludwig von Canitz: Gedichte. Berlin, 1700. Image 16.
www.deutsches-textarchiv.de/canitz_gedichte_1700/16

Like TEI-Lib, the DTA-BF defines different encoding levels according to the depth of text structuring, thus categorizing all available DTA-BF elements due to different text structuring necessities and depths. In accordance with the TEI Guidelines but in contrast to TEI-Lib (TEI SIG on Libraries, 2011: ch. 1), the levels 1 to 3 of the DTA-BF are strictly cumulative.¹⁰ The first level represents the least common denominator for text structuring, therefore containing elements that are mandatory for basic semantic text annotation. The elements in level 2 are strongly recommended but not mandatory. Level 3 contains optional elements, which can be used for in-depth semantic text structuring, but are not applied extensively throughout the DTA core corpus.¹¹ Thus, the obligation to use the provided elements decreases with the increase of levels and, in connection with that, the depth of text structur-

¹⁰For an overview on the DTA-BF elements and their corresponding levels cf. www.deutsches-textarchiv.de/doku/basisformat_table.

¹¹However, the DTA-BF remains an annotation format for the structuring of historical corpus texts, esp. serving linguistic purposes. Projects, which aim at providing historical critical editions of texts, will need further annotation possibilities (e.g. an inventory for a critical apparatus as specified in chapter 12 of the TEI guidelines; cf. Burnard and Baumann, 2012: ch. 12). Such projects might want to start off with the DTA-BF for annotation and extend it according to their requirements.

ing. The fourth level contains an exception list of elements which should be avoided in favor of a different solution provided by the DTA-BF.

3.3 Metadata

Like TEI-Lib (TEI SIG on Libraries, 2011: ch. 4.1), the DTA-BF provides a TEI Header customization which allows to express rich bibliographic metadata for each corpus text. The DTA-BF metadata specification focuses on the bibliographic description of written corpora. We provide conversion routines to other standards such as the Component Metadata Infrastructure CMDI¹², which is the recommended metadata format in CLARIN.¹³

3.4 Text Transcription and Editorial Interventions

In addition to the DTA-BF, extensive transcription guidelines are provided in order to support common transcription practices for each text in the DTA corpus.¹⁴ To this end, furthermore, the DTA-BF contains regulations for possible editorial interventions. From the TEI formats mentioned above, only the TEI-Lib guidelines point out the necessity of transcription guidelines, but limit their advice to the handling of punctuation and hyphenation problems (TEI SIG on Libraries, 2011: ch. 3.2).

3.5 Documentation

The DTA-BF comes with a detailed documentation¹⁵ explaining the usage of each element, attribute, and value according to the possible annotation needs in text structuring. The documentation contains examples taken from the DTA corpus and illustrates typical encoding scenarios as well as exception cases.

There is also an ODD¹⁶ specification for the DTA-BF together with a corresponding RNG schema¹⁷ generated with TEI-ROMA.¹⁸

¹²Cf. www.clarin.eu/cmdi.

¹³Cf. footnote 3.

¹⁴Cf. www.deutsches-textarchiv.de/doku/richtlinien.

¹⁵Cf. www.deutsches-textarchiv.de/doku/basisformat.

¹⁶One document does it all; cf.

www.tei-c.org/Guidelines/Customization/odds.xml.

¹⁷Cf. www.deutsches-textarchiv.de/basisformat.odd; www.deutsches-textarchiv.de/basisformat.rng.

¹⁸Cf. www.tei-c.org/Roma.

3.6 Relation to the TEI Guidelines

Like TEI Lite, TEI-Lib, and TextGrid's BE, the DTA-BF is a strict subset of the TEI P5 tag set. It is therefore entirely compatible with the TEI P5 Guidelines in that they are only customized by selection, but not extended in any way.

4 Lifecycle of the DTA ‘base format’ within DTAE

DTA Extensions (DTAE) is a module of the DTA project with the goal to integrate digitized historical German texts drawn from external sources into the DTA core corpus. There are two prerequisites for those texts: they need to be considered as influential with respect to the goal of the DTA to compile a historical reference corpus, and they have to dispose of a high transcription quality.

External resources may be transcribed either in a word processing or HTML environment – a case we do not discuss here since it has no effect on the DTA-BF – or more often (as more and more philological projects adopt the TEI) be encoded in a customized TEI P5 format. In this case, a transformation of the customized TEI schema into the DTA-BF has to be specified. In general, all texts are subject to a quality assurance phase before being published in the DTA environment (Geyken et al., 2012; Haaf et al., 2012). For this task, a web-based distributed quality assurance platform has been implemented (Wiegand and Geyken, 2011), where users can proofread texts page by page and report different kinds of errors. As a result of the conversion and correction process, material from heterogeneous corpus formats is made accessible in the context of one homogeneous, high-quality text corpus.

So far, corpus texts from 10 external projects with a total of 200,000 pages were integrated into the DTA corpus after being converted into the DTA-BF, including *Blumenbach online*, *AEDit*, and *Dinglers Polytechnisches Journal*.¹⁹

We distinguish three cases for the integration of external TEI-encoded corpora into the DTA environment: 1. The conversion of the customized TEI schema into the DTA-BF can be done automatically, since the DTA-BF provides a semanti-

¹⁹Cf. www.deutsches-textarchiv.de/doku/dtae for the full list of DTAE projects.

cally equivalent solution. 2. The solution adopted in the customized TEI schema corresponds to a text phenomenon which has not been considered by the DTA-BF so far and which in turn leads to a modification of the DTA-BF. 3. The external text corpus cannot be automatically converted, either because the underlying TEI schema is too flexible thus leading to structuring ambiguities (cf. section 1), or because the schema is applied inconsistently over the text collection. Since this last case requires manual intervention, it is only considered for external texts which are stable, either because the project is finished, or because the quality of the transcription and the structural encoding is sufficient, which means no additional annotation work is likely to be carried out on the source text.

The customized TEI schema of the project *Dinglers Polytechnisches Journal* may serve as an example for the first case. This schema defines missing transcriptions due to illegibility of the text source as follows:²⁰

```
<unclear reason="problem">
  [Fehlender Text
   (engl.: missing text)]
</unclear>
```

Even though the DTA-BF does not include the TEI element `<unclear>`, this expression can easily be converted into the equivalent of the DTA-BF annotation:

```
<gap reason="illegible"/>
```

The following two examples illustrate the second case, i.e. modifications of the DTA-BF according to the requirements of external corpus projects:

The *Blumenbach online* project provides editorial figure descriptions (`<figDesc>`), a kind of additional information about the source text given by the editor. Such additional information was not foreseen by the DTA-BF. Since the `<figDesc>` is only a special case of an editorial comment, the DTA-BF element `<note>` was extended by the attribute-value combination `@resp="editorial"`. With this extension, we were able to preserve the figure descriptions

²⁰Cf. dingler.culture.hu-berlin.de/article/pj003/ar003042 for an example.

of the edited Blumenbach texts and to generally allow for editorial comments elsewhere.

Furthermore, modifications of the DTA-BF may become necessary due to the integration of new (special) text types in the DTA corpora. E.g. in the context of the DFG funded project *AEdit Frühe Neuzeit (Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit)* the *Forschungsstelle für Personalschriften (Academy of Sciences and Literature in Mainz)* is currently digitizing funeral sermons of the former municipal library in Wrocław. The digitized texts are being annotated according to the DTA-BF. The addition of new specific @type-values for `<div>` elements became necessary in order to allow for the naming of different text types within a funeral sermon. The new values added to the existing value selection were prefixed `fs` in order to limit their usage to the document type “funeral sermon” (e.g. `fsSermon`, `fsConsolationLetter`, `fsCurriculumVitae`, `fsEpitaph` etc.).

However, possible modifications of the DTA-BF are considered carefully in order to avoid negative effects on the annotation consistency of the DTA corpus.

5 Conclusion and further prospects

In this article, we have presented the DTA ‘base format’, a strict subset of TEI P5. The DTA-BF has been designed and developed with the goal to cope with a large variety of text types of written German corpora. It is a reasonable common denominator for a large reference corpus of the historical New High German ranging from 1650 to 1900. It goes without saying that the success of the DTA-BF is largely dependent on its adoption by other projects, namely the number of documents encoded in the format. Establishing the usage of the DTA-BF in a broader context may be supported considerably within a large infrastructure such as provided by CLARIN and, for the German context, CLARIN-D, where major text corpus providers are gathered pursuing the goal to define policies which guarantee the interoperability of resources which are integrated into the infrastructure. The *Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)* as a partner of the CLARIN-D project, as a future CLARIN Center, and as the coordinator of

the work package ‘Resources and Tools’ plays an important role in the discussion process. In addition, a CLARIN-D corpus project has recently been started with the goal to curate already existing corpus texts of the 15th to 19th century and to integrate them into the CLARIN-D infrastructure by using the DTA-BF as a starting point, thus enabling the DTA-BF to evolve in an environment of even more heterogeneous text resources. The project partners of this CLARIN-D curation project are the BBAW (coordination), the *Herzog August Library of Wolfenbüttel*, the *Institute for the German Language Mannheim*, and the *University of Gießen*.

References

- Lou Burnard, and Syd Bauman. 2012. *P5: Guidelines for Electronic Text Encoding and Interchange*, Version 2.1.0, June 17th, 2012. www.tei-c.org/release/doc/tei-p5-doc/en/html.
- Lou Burnard, and C. M. Sperberg-McQueen. 2006. *TEI Lite: Encoding for Interchange: an introduction to the TEI – Revised for TEI P5 release* (February 2006). www.tei-c.org/Guidelines/Customization/Lite/.
- Michelle Dalmau, and Melanie Schlosser. 2011. *Challenges of serials text encoding in the spirit of scholarly communication*. In: Library Hi Tech 28,3 (2011), pp. 345–359. <http://dx.doi.org/10.1108/07378831011076611>.
- Michael Day. 2010. *IMPACT Best Practice Guide: Metadata for Text Digitisation & OCR*. www.impact-project.eu/uploads/media/IMPACT-metadata-bpg-pilot-1.pdf/.
- Alexander Geyken. 2007. *The DWDS corpus: A reference corpus for the German language of the 20th century*. In: Christiane Fellbaum (Hg.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London, 23–41.
- Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, and Frank Wiegand. 2012. *TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv*. In: *Jahrbuch für Computerphilologie*, Jg. 9, 2012.
- Susanne Haaf, Frank Wiegand, and Alexander Geyken. 2012. *Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text*. Accepted to be published in the *Journal of the Text Encoding Initiative* 3, 2012.
- Institute for the German Language (IDS) Mannheim. 2012. *IDS-Textmodell: Unterschiede gegenüber XCES*. www.ids-mannheim.de/kl/projekte/korpora/idsxces.html (accessed June 24th, 2012).
- TEI SIG on Libraries. 2011. *Best Practices for TEI in Libraries. A TEI Project*. Ed. by Kevin Hawkins, Michelle Dalmau, and Syd Bauman, Version 3.0 (October 2011). www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html.
- Stephen Ramsay, and Brian Pytlak Zillig. 2011. *Code Generation Techniques for Document Collection Interoperability*. Chicago Colloquium on Digital Humanities and Computer Science 2011. http://chicagocolloquium.org/wp-content/uploads/2011/11/dhcs2011_submission_6.pdf.
- Maik Stührenberg. 2012. *Auszeichnungssprachen für linguistische Korpora: theoretische Grundlagen, De-facto-Standards, Normen*. Bielefeld: Universität Bielefeld. urn:nbn:de:hbz:361-24927723.
- TextGrid. 2007–2009. *TextGrid’s Baseline Encoding for Text Data in TEI P5*. www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf (accessed June 24th, 2012).
- Perry Trolard. 2011. *TEI Lite – A recommendation for off-site text encoding*. Version 1.1, September 2011. www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite.doc.html.
- John Unsworth. 2011. *Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI*. *Journal of the Text Encoding Initiative* 1 (June 2011). <http://jtei.revues.org/215> (access June 24th, 2012).
- Frank Wiegand, and Alexander Geyken. *Quality assurance of large TEI corpora*. Poster, presented at the 2011 Annual Conference and Members’ Meeting of the TEI Consortium: *Philology in the Digital Age*. Würzburg, 2011. www.deutschestextarchiv.de/doku/tei-mm-2011_poster.pdf.
- Brian Pytlak Zillig. 2009. *TEI Analytics: converting documents into a TEI format for cross-collection text analysis*. *Literary and Linguistic Computing*, 24(2), 187–192. doi:10.1093/linc/fqp005.
- Brian Pytlak Zillig, Syd Bauman, Julia Flanders, and Steve Ramsay. 2012. *MONK TEIAnalytics. Target schema for MONK ingestion transformation*. segonku.unl.edu/teianalytics/TEIAnalytics.html (accessed June 24th, 2012).

Building An Old Occitan Corpus via Cross-Language Transfer

Olga Scrivner

Indiana University

Bloomington, IN, USA

obscrivn@indiana.edu

Sandra Kübler

Indiana University

Bloomington, IN, USA

skuebler@indiana.edu

Abstract

This paper describes the implementation of a resource-light approach, cross-language transfer, to build and annotate a historical corpus for Old Occitan. Our approach transfers morpho-syntactic and syntactic annotation from resource-rich source languages, Old French and Catalan, to a genetically related target language, Old Occitan. The present corpus consists of three sub-corpora in XML format: 1) raw text; 2) part-of-speech tagged text; and 3) syntactically annotated text.

1 Introduction

In the past decade, a number of annotated corpora have been developed for Medieval Romance languages, namely corpora of Old Spanish (Davies, 2002), Old Portuguese (Davies and Ferreira, 2006), and Old French (Stein, 2008; Martineau, 2010). However, annotated data are still sparse for less-common languages, such as Old Occitan. For example, the only available electronic database “The Concordance of Medieval Occitan”¹, published in 2001, is not free and is limited to lexical search.

While the majority of historical corpora are built by means of specific tools developed for each language and project, such as TWIC for NCA (Nouveau Corpus d’Amsterdam) (Stein, 2008), or a probabilistic parser for the GTRC project (MCVF Corpus) (Martineau et al., 2007),

the goal of this project is to implement a resource-light approach, exploiting existing resources and common characteristics shared by Romance languages.

It is well known that Romance languages share many lexical and syntactic properties. The following example illustrates the similarity in word order and lexicon of Old Occitan, Old Catalan, and Old French²:

- | | | | | | |
|-----|------|------------|---------|---------|---------|
| (1) | Oc: | Dedins | la | cambra | son |
| | Cat: | Dins | la | cambra | són |
| | Fr: | Dans | la | chambre | elles |
| | | vengudas, | dejosta | lui | son |
| | | vingudas, | davant | ell | són |
| | | entrées, | devant | lui | elles |
| | | assegudas. | | | se sont |
| | | assegudas. | | | |
| | | assises. | | | |
- ‘They came into the room, they sat down next to him.’

Several recent experiments have demonstrated that genetically related languages can share their knowledge through cross-language transfer (Hana et al., 2006; Feldman and Hana, 2010) between closely related languages. That is, a resource-rich language can be used to process unannotated data in other genetically related languages. For example, Hana et al. (2006) have used Spanish morphological and lexical data for automatic tagging of Brazilian Portuguese. While the idea of cross-language transfer is not new and

¹<http://www.digento.de/titel/100553.html>

²Catalan and French translation is ours. We approximated Old Catalan and Old French word order as far as possible.

is mainly used with parallel corpora and large bilingual lexicons (Yarowsky and Ngai, 2001; Hwa et al., 2005), experiments by Hana et al. (2006) have demonstrated the usability of this method in situations where there are no parallel corpora but instead resources for a closely related language.

Our goal is to build a corpus of Old Occitan in a resource-light manner, by using cross-language transfer. This approach will be used not only for part-of-speech tagging, but also for syntactic annotation.

The organization of the remainder of the paper is as follows: Section 2 provides a brief description of Old Occitan. Section 3 reviews the concept of a resource-light approach in corpus linguistics. Section 4 provides details on corpus pre-processing. The methods for cross-linguistic part-of-speech (POS) tagging and cross-linguistic parsing are described in Sections 5 and 6. Finally, the conclusions and directions for further work are presented in Section 7.

2 Old Occitan (Provençal)

Occitan, often referred to as Provençal, constitutes an important element of the literary, linguistic, and cultural heritage in the history of Romance languages. Provençal (Occitan) poetry was a predecessor of French lyrics. Moreover, Occitan was the only administrative language in Medieval France, besides Latin (Belasco, 1990). While the historical importance of this language is indisputable, Occitan, as a language, remains linguistically understudied. Compared to Old French, Provençal is still lacking digitized copies of scanned manuscripts, as well as annotated corpora for morpho-syntactic or syntactic research.

Typologically, Old Occitan is classified as one of the Gallo-Roman languages, together with French and Catalan (Bec, 1973). If one examines Old Occitan, Old French, and Old Catalan, on the one hand, it is striking how many lexical and morphological characteristics these languages share. For example, French and Occitan have rich verbal inflection and a two-case nominal system (nominative and accusative), illustrated with an example of the word ‘wall’ in (2):



Figure 1: Linguistic map of France, from (Bec, 1973)

	Case	Old Occitan	Old French
(2)	Nominative	lo murs	li murs
	Accusative	lo mur	lo mur

On the other hand, Occitan has syntactic traits similar to Catalan, such as a relatively free word order and null subjects, illustrated in (3) and (4).

- (3) *Gran honor nos fai*
great honor us_{Dat} does
'He grants us a great honor' (Old Occitan - Flamenca)
- (4) *molt he gran desig*
much have big desire
'I have a lot of great desire...' (Old Catalan - Ramon Llull)³

The close relationship between these three languages is also marked geographically. The northern border of the Occitan-speaking area is adjacent to the French linguistic domain, whereas in the south, Occitan borders on the Catalan-speaking area, as shown in Figure 1.

This project focuses on the 13th century Old Occitan romance “Flamenca”, from the edition by Meyer (1901). Apart from a very intriguing fable of beautiful Flamenca imprisoned in a tower by her jealous husband, this story presents a

³<http://orbita.bib.ub.edu/ramon/velec.asp>

very interesting linguistic document consisting of 8097 lines of the “universally acknowledged masterpiece of Old Occitan narrative” (Fleischmann, 1995). Multiple styles, such as internal monologues, dialogues and narratives, provide a rich lexical, morphological and syntactic database of a language spoken in southern France.

3 Linguistic Annotation via Cross-Language Transfer

Corpus-based approaches often require a great amount of parallel data or manual labor. In contrast, the cross-language transfer, as proposed by Hana et al. (2006), is a resource-light approach. That is, this method does not involve any resources in the target language, neither training data, a large lexicon, nor time-consuming manual annotation.

While cross-language transfer has been previously applied to languages with parallel corpora and bilingual lexica (Yarowsky and Ngai, 2001; Hwa et al., 2005), Hana et al. (2006) introduced a method in the area where these additional resources are not available. Feldman and Hana (2010) performed several experiments with Romance and Slavic languages. The only resources they used were i) POS tagged data of the source language, ii) raw data in the target language, and iii) a resource-light morphological analyzer for the target language. For POS tagging, the Markov model tagger TnT (Brants, 2000) was trained on a source language, namely Spanish and Czech, in order to obtain transition probabilities. The reasoning is that the word order patterns of source and target languages are very similar so that given the same tagset, the transition probabilities should be similar, too. Since the languages differ in their morphological characteristics, a direct transfer of the lexical probabilities was not possible. Instead, a shallow morphological analyzer was developed for the target languages, using cognate information, among other similarities. The trained models were then applied to the target languages, Portuguese, Catalan, and Russian. Tagging accuracies for Catalan, Portuguese, and Russian yielded 70.7%, 77.2%, and 78.6% respectively.

In contrast, syntactic transfer is mainly used in machine translation. This approach requires a bilingual corpus aligned on a sentence level.

That is, words in a source language are mapped to words in a target language. Dien et al. (2004) used this method on English-Vietnamese corpus. They extracted a syntactic tree set from English and transferred it into the target language. Obtained two sets of parsed trees, English and Vietnamese, were further used as training data to extract transfer rules. In contrast, Hanneman et al. (2009) extracted unique grammar rules from English-French parallel parsed corpus and selected high-frequency rules to reorder position of constituents. Hwa et al. (2005) describe an approach that focuses on syntax projection per se, but their approach also relies on word alignment in a parallel corpus. They show that the approach works better for closely related languages (English to Spanish) than for languages as different as English and Chinese. It is widely agreed that word alignment and thus, syntactic transfer, is best applied in similar languages due to their word order pattern (Watanabe and Sumita, 2003). Therefore, genetically related Romance languages should be well suited for syntactic cross language transfer. McDonald et al. (2011) and Naseem et al. (2012) describe novel approaches that use more than one source language, reaching results similar to those of a supervised parser for the source language.

While cross-language transfer has been applied successfully to modern languages, we decided to use it to transfer linguistic annotation to a historical corpus. The choice of source languages was based on the availability of annotated resources and the similarity of language characteristics. Thus, Old French corpus (Martineau et al., 2007) was selected as a source for the morpho-syntactic annotation of Occitan. However, to transfer syntactic information, we used the Catalan dependency treebank (Civit et al., 2006) since modern Catalan displays a pro-drop feature and a relatively free word order, similarly to Old Occitan.

4 Corpus Pre-Processing

The romance ‘Flamenca’ is available in scanned images format, therefore, the initial step included conversion to an electronic version via OCR and manual correction. Figure 2 shows a sample of the manuscript.

Corrections by the editor were omitted. As can be seen in the first line of the document (see Figure 2), the editor (Meyer, 1901) enclosed a silent letter ‘s’ in brackets which we have excluded from our pre-processed text. In addition, we detached the clitic pronouns that are joined to the verbs, as in (5), as shown in (6).

- (5) *Per son anel dominim manda Que*
 for his ring coat of arms sends that
Flamenca penra sim voi.
Flamenca takes if me want
 ‘He is sending his family ring as a guarantee
 that, if I want, he will marry Flamenca’
- (6) *Per son anel domini m manda Que*
Flamenca penra si m voi.
- « Poissas lur di[s] tot en apert : (fol. 2)
 « Vostre cor nom tengas cubert,
 « Mais digas mi : si Dieus mi dona
 4 « Un'aventura que m'es'bona,
 « Non sabra bon a totz ensems?
 « Ieu ai desirat mout lonc temps
 « C'ap N'Archimbaut agues paria,
 8 « Ar son vengutz d'en lai al dia
 « Ques el la quer e la demanda :
 « Per son anel dominim manda
 « Que Flamenca penra sim voi.

Figure 2: Sample of page from ‘Flamenca’

Currently, we have pre-processed and formatted 3 095 lines, which corresponds to 12 573 tokens. The text file was then converted to XML format using EXMARaLDA⁴. While EXMARaLDA is mostly used for transcriptions, it also imports files from several formats, such as plain text or tab format, and exports them as EXMARaLDA XML files. This XML is timeline-based and supports the annotation of different linguistic levels in different tiers.

5 Cross-Linguistic POS Tagging

Since Old French and Old Occitan share many morphological features, we have adopted the POS tagset from the MCVF corpus of Old French (Martineau et al., 2007). The MCVF tagset is based on the annotation scheme of the Penn-Helsinki Parsed Corpus of Middle

⁴www.exmaralda.org

Tag	Description
ADJ	adjective
ADJR	comparative form of adjective
ADV	adverb
ADVR	comparative form of adverb
AG	gerundive of auxiliary ‘to have’
AJ	present of auxiliary ‘to have’
APP	past participle of auxiliary ‘to have’
AX	infinitive of auxiliary ‘to have’
CONJO	coordinative conjunction
CONJS	subordinate conjunction
COMP	comparative adverb
D	determiner (indefinite, definite, demonstrative)
DAT	dative
DZ	possessive determiner
EG	gerundive of auxiliary ‘to be’
EJ	present of auxiliary ‘to be’
EPP	past participle of auxiliary ‘to be’
EX	infinitive of auxiliary ‘to be’
ITJ	interjection
MDG	gerundive of modal verb
MDJ	present of modal verb
MDPP	past participle of modal verb
MDX	infinitive of modal verb
NCPL	noun common plural
NCS	noun common singular
NEG	negation
NPRPL	noun proper plural
NPRS	noun proper singular
NUM	numeral
P	preposition
PON	punctuation inside the clause
PONFP	the end of the sentence
PRO	pronoun
Q	quantifier
VG	gerundive of the main verb
VJ	present of the main verb
VPP	past participle of the main verb
VX	infinitive of the main verb
WADV	interrogative, relative or exclamative adverb
WD	interrogative, relative or exclamative determiner
WPRO	interrogative, relative or exclamative pronoun

Table 1: Occitan tagset

English (PPCME) (Kroch and Taylor, 2000), which was modified to represent French morphosyntactically, as illustrated in (7).

- (7) *Les/D petites/ADJ filles/NCPL*
 the/Det. little/Adj. girls/Noun-Cmn-pl.
 ‘the little girls’

While the MCVF tagset consists of 55 tags, we have decreased the tagset to 39 tags for our corpus. The Occitan tagset is shown in Table 1. The simplification included joining certain subclasses into one class. The reason for this modification lies in the particularities of Occitan. First, as a pro-drop language, Occitan omits an impersonal pronoun (8), in contrast to Old French (9).

- (8) *No m' o cal dir*
 Not me it_{Acc} must say
 ‘it is not necessary for me to tell this’
- (9) *Que te faut il en ce*
 what you_{Acc} must it_{Impersonal} in this
 pais?
 country
 ‘What do you need in this country?’
 (MCVF corpus)

Furthermore, the grammar of Old Occitan (Anglade, 1921) does not use “near future” tense, which is common in Old French and is formed by the verb *aller* ‘to go’ and the infinitive of a main verb (10). Therefore, the specific labels LJ, LX, LPP for the auxiliary *aller* from the corpus of Old French are mapped to the corresponding tags of the main verb, such as VJ, VX, VPP, VG (see Table 1).

- (10) *Et que iroi ge faire?*
 and what will go I do
 ‘What am I going to do?’

Finally, the French tagset contains a label FP for focus particle, such as *seullement/seulement* and *ne...que* ‘only’ (11). The following comparison shows that the latter construction *ne..que* does not convey focus in Old Occitan (12):

- (11) *le jeune homme ne le fait que*
 the young man not it does only
pour l' avarice
 for the greed
 ‘Young man does it **only** for greed’ (MCVF
 corpus)
- (12) *Don non cug que ja mais reveinha*
 of it not think that ever returns
 ‘I do **not** think **that** he ever recovers from it’

We trained TnT on 28 265 sentences from the Medieval French texts (MCVF). This trained model was used to POS tag Old Occitan without any modification to the lexicon. For the performance evaluation we extracted 50 sentences (1000 tokens) from the corpus and annotated them manually. Then, the tagger output was compared to the gold standard. The POS tagger

Accuracy	N	%
All Words	640/1000	64.00
Known Words	539/742	72.64
Unknown Words	101/258	39.15

Table 2: POS evaluation using the unmodified Medieval French lexicon

reached an accuracy of 64.00% for all words and 72.64% for known words, cf. Table 2.

A manual analysis of randomly selected 20 sentences revealed that a number of errors are caused by lexical duplicates that have different meanings in each language. For example, in (13) *no* is a possessive determiner ‘our’ in Old French, while in Occitan *no* is a negation. The second type of errors is the result of TnT’s algorithm for handling unknown words by a suffix trie. That is, unknown words are assigned to an ambiguity class depending on their suffix. For example, the unknown Occitan word *ancar* ‘yet’ is recognized as an infinitive (VX), based on the ending -ar which is common for French infinitives; whereas *entremes* ‘involves’ receives higher probability as an adjective (ADJ) because of its ending -es (see (13)).

- (13) *Ancar d' amor no s'*
 TnT: VX P NCS DZ ADV
 gold: ADV P NCS NEG PRO
entremes
 ADJ
 VJ
 ‘he is not yet involved in the love affair’

Therefore, to improve the fit of the lexicon extracted from Medieval French with words specific to Occitan, we added 171 manually annotated sentences from ‘Flamenca’ to the training data. The validation on the test set yielded 78.10% accuracy for all words, 81.10% for known, and 57.48% for unknown words, see Table 3. The results prove that adding even a small set of high quality, manually annotated sentences in the target languages improves POS tagging quality considerably, bringing the tagger’s performance close to results reached for modern languages (given

Accuracy	N	%
All Words	781/1000	78.10
Known Words	708/873	81.10
Unknown Words	73/127	57.48

Table 3: Evaluation with the Occitan-enriched lexicon

a morphologically richer language and a small training set), thus validating our approach.

Furthermore, in the course of our project we have compiled an Occitan dictionary, consisting of 2 800 entries. The glossary to Flamenca (Meyer, 1901) was used as a reference guide. Thus, our tagged corpus was augmented with lemmas from the dictionary. We have further manually checked and corrected 8 284 tags in the corpus. Finally, the POS tagged version was converted to XML, again using EXMARaLDA.

6 Syntactic Cross-Linguistic Parsing

While it has been widely accepted that syntactic annotation in terms of constituent trees provides a rich internal tree structure, recent years have shown an increased interest in dependency graphs (Civit et al., 2006). Dependency graphs provide an immediate access to lexical information for words, word pairs, and their grammatical relations. For example, each word in the Catalan sentence (14) has exactly one head, as demonstrated in Figure 3. The arcs show dependencies from heads to dependents.

- (14) *Li agrada l'actualizació que*
 to him likes the modernisation that
Barea ha fet del text.
 Barea has made of the text.
 ‘He likes how Barea modernized the text’.

In addition, it is argued that dependency grammars “deal especially well with languages involving relatively free word order” (Bamman and Crane, 2011). Since Old Occitan has relatively free word order, we aim for a syntactic annotation in form of dependency graphs. At present, dependency treebanks are available only for modern Romance languages, namely French (Abeillé et al., 2003), Catalan, and Spanish (AnCora), (Civit

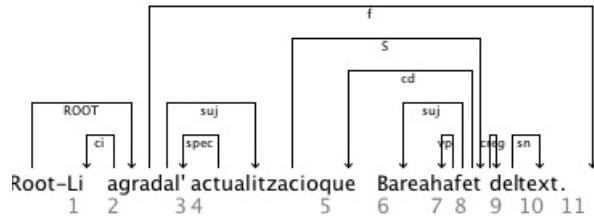


Figure 3: Example of Dependency Relation from the Catalan Treebank (Ancora)

ID	Word	Lemma	Pos	Head	Dep
1	Li	ell	PP3CS00	2	ci
2	agrada	agradar	VMIP3S0	0	sent
3	l'	el	DA0CS0	4	spec
4	actualització	actualització	NCFS000	2	suj
5	que	que	PR0CC000	8	cd
6	Barea	Barea	NP00000	8	suj
7	ha	haver	VAIP3S0	8	v
8	fet	fer	VMP00SM	4	S
9	del	del	SPCMS	8	creg
10	text	text	NCMS000	9	sn
11	.	.	Fp	2	f

Table 4: Annotation in AnCora

et al., 2006). For training a dependency parser, we use Catalan rather than modern French since the syntactic characteristics of Catalan are more similar to Occitan than French. For example, Old Occitan is a pro-drop language while French is not. We used the Catalan treebank from AnCora⁵.

The Catalan treebank consists of 16 591 sentences extracted from newspapers and annotated syntactically and semantically. The dependency treebank has been converted automatically from a constituency format with the help of a table of head finding rules (Civit et al., 2006). The sentence in (14), for example, is annotated in the treebank format as shown in Table 4.

As shown in Table 4, the Catalan tagset describes information about the major POS classes, represented as letters, and morphological features, such as gender, number, case, person, time and mode, represented as digits. The tagset has a total of 280 different labels (Taulé et al., 2008). The rich morphological information allowed us to map the Catalan tags to our Occitan tagset with high accuracy. For example, in the sentence from

⁵<http://clic.ub.edu/corpus/en/ancora-descarregues>

Table 4, verbal features such as VMIP (main verb indicative present) and VMP (main verb past participle) were mapped to Occitan tags VJ and VPP, respectively. The example of the mapped Catalan tags is shown in (15).

- (15) *Li agrada l'actualizació que
PRO VJ NCS WPRO
Barea ha fet del text .
NPC AJ VPP P NCS PONFP*
'He likes how Barea modernized the text'.

The Catalan dependency representation contains a large set of grammatical relations. We found 48 different labels in the AnCora Corpus. We decided to map these dependencies to the Penn Treebank core dependencies, such as subject (SBJ), direct object (OBJ), predicate (PRED), nominal modifier (NMOD), verbal modifier (VMOD). In addition, we added a language specific relation - CL (clitic) (16). The complete list of Occitan dependency labels and their corresponding labels in Catalan is shown in Table 5.

- (16) *Pero vostre sen m' en digas
but your opinion me about it tell
'But tell me you opinion about it.'*

It is necessary to note that verbal head selection in the AnCora Corpus differs from the constituency head assignment (Civit et al., 2006). In the dependency annotation, the head is assigned to the rightmost element in the verbal phrase. For example, past participles, and gerunds are heads, whereas auxiliaries are dependents. In addition, the treebank is automatically augmented by empty elements to represent null subjects.

In order to annotate our Old Occitan texts, we trained a transition-based dependency parser, MaltParser (Nivre et al., 2007) on the Catalan treebank with the reduced tagset. We then used the trained model to parse our corpus.

We manually annotated 30 sentences to evaluate the accuracy of the parser. For the evaluation we used MaltEval⁶, an evaluation tool for dependency trees. The results yielded 63.1% of label accuracy and 55.8% of labeled attachment. The highest score of precision and recall was for

Occitan	Relation type	Catalan
ROOT	Main clause	Sentence
S(bar)	Infinitival sentence	infinitiu
	Subordinate complement	ao
SBJ	Subject	suj
OBJ	Nominal Direct Object	cd
CL	Pronominal direct object	cd
	Pronominal morpheme	morfema.pron
	Impersonal	impers
	Passive 'se'	pass
PRED	Predicative	cpred
	Atribute	atr
NMOD	Determiner	d
	Numbers	z
	Prepositional phrase	sp
	Adjectival phrase	s.a
	Determiner	spec
	Second adj inside sn	grup.a
VMOD	Prepositional complement	creg
	Adverbial phrase	sadv
	Verb adjunct	cc
	Indirect object	ci
PMOD	Noun phrase	sn
	Adverbial modifier	r
	Second noun phrase inside sp	grup.nom
	Direct object	ci
V	Auxilliary	v
NEG	Verb modifier 'no'	mod
CONJ	Conjunction que	conj
COORD	Coordination	c
REL	Relative pronoun	relatiu
INC	Inserted phrase	inc
P	Punctuation	f

Table 5: Dependency Labels

Deprel	precision	recall
P	91.2	86.7
CL	89.3	78.1
NEG	77.8	77.8
NMOD	86.1	87.3
PMOD	76.9	78.9
PRED	42.9	16.7
ROOT	44.8	78.9
OBJ	42.3	34.4
S	45.8	16.2
SBJ	37.9	55.0
V	53.3	57.1
VMOD	46.2	56.3

Table 6: Parsing results

nominal modifiers, prepositional modifiers, clitic, negation and punctuation, cf. Table 6.

As can be seen from Table 6, subject, object and predicate relations are the least accurate. This is due to a relatively free word order in Old Occi-

⁶<http://w3.msi.vxu.se/users/jni/malteval/>

```

$<$sentence id="19" user="" date=""$>$
  $<$word id="1" form="Le" postag="D" head="2" deprel="NMOD"/$>$
  $<$word id="2" form="coms" postag="NCS" head="3" deprel="SBJ"/$>$
  $<$word id="3" form="fes" postag="VJ" head="0" deprel="ROOT"/$>$
  $<$word id="4" form="sa" postag="DZ" head="5" deprel="NMOD"/$>$
  $<$word id="5" form="mollier" postag="NCS" head="3" deprel="OBJ"/$>$
  $<$word id="6" form="venir" postag="VX" head="5" deprel="S"/$>$
  $<$word id="7" form"." postag="PONFP" head="3" deprel="P"/$>$
$<$/sentence$>$

```

Figure 4: An example of the syntactic annotation

tan and its pro-drop feature. The example in (17) illustrates that in some cases, the noun, here *Flamenca*, can be ambiguous between subject or object of the sentence.

- (17) *Que Flamenca penra*
 that Flamenca will take
 'that Flamenca will take' / 'that he will take
 Flamenca'

In contrast, the low precision of ROOT is due to MaltParser's design which may lead to incomplete syntactic annotations. To repair this type of errors, we performed post-editing, which readjusts ROOT labels, increasing its accuracy to 71%, instead of 44.8%. In addition, it yielded better results for label accuracy - 76.4% and label attachment score - 63.4%.

Finally, the parsed output was converted to XML format using Malt Converter⁷. An example of the resulting XML file is presented in Figure 4.

7 Conclusion and Future Work

While the annotation of historical corpora requires precision, the first steps in building a syntactically annotated corpus can be resource-light. If we use a cross-language transfer, we can profit from existing resources for historical or modern languages sharing similar morphological and syntactic features. We have shown that Old French presents a good source for POS tagging of Old Occitan while modern Catalan is a good source for the syntactic annotation of Occitan.

The POS tagger model trained on the Old French Corpus MCVF (Martineau et al., 2007) yielded 78% accuracy when we added a small Occitan lexicon into training, whereas dependency

parsing results yielded 63.4% of labeled attachment.

For the future, we plan to annotate the whole corpus of 8 000 lines, manually correct them, and make it available on the web. At present, 3 095 lines of tagged and parsed XML files are available upon request.

References

- Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*. Kluwer.
- Joseph Anglade. 1921. *Grammaire de l'ancien provençal ou ancienne langue d'oc*. Librairie C. Klincksieck.
- David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 79–98. Springer.
- Pierre Bec. 1973. *La langue occitane*. Number 1059 in Que sais-je? Paris: Presses universitaires de France.
- Simon Belasco. 1990. France's rich relation: The Oc connection. *The French Review*, 63(6):996–1013.
- Thorsten Brants. 2000. TnT-a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, pages 224–231, Seattle, WA.
- Monserat Civit, Antònia Martí, and Nuria Bufí. 2006. Cat3LB and Cast3LB: From constituents to dependencies. In *Advances in Natural Language Processing*, pages 141–153. Springer.
- Mark Davies and Michael Ferreira. 2006. Corpus do Portugues: 45 million words, 1300s-1900s. Available online at <http://www.corpusdoportugues.org>.
- Mark Davies. 2002. Corpus del Español: 100 million words, 1200s-1900s. Available online at <http://www.corpusdelespanol.org>.

⁷<http://w3.msi.vxu.se/~nivre/research/MaltXML.html>

- Dinh Dien, Thuy Ngan, Xuan Quang, and Chi Nam. 2004. The parallel corpus approach to building the syntactic tree transfer set in the English-to-Vietnamese machine translation. In *2004 International Conference on Electronics, Information, and Communications (ICEIC2004)*, pages 382–386, Ha Noi, Vietnam.
- Anna Feldman and Jirka Hana. 2010. *A Resource-Light Approach to Morpho-Syntactic Tagging*. Rodopi.
- Suzanne Fleischmann. 1995. The non-lyric texts. In F.R.P. Akehurst and Judith M. Davis, editors, *A Handbook of the Troubadours*, pages 176–184. University of California Press.
- Jirka Hana, Anna Feldman, Chris Brew, and Luiz Amaral. 2006. Tagging Portuguese with a Spanish tagger using cognates. In *Proceedings of the EACL Workshop on Cross-Language Knowledge Induction*, pages 33–40, Trento, Italy.
- Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An improved statistical transfer system for French-English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 140–144.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.
- Anthony Kroch and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania.
- France Martineau, Constanta Diaconescu, and Paul Hirschbühler. 2007. Le corpus ‘voies du français’: De l’élaboration à l’annotation. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d’Amsterdam*, pages 121–142. Steiner.
- France Martineau. 2010. Corpus MCVF, modéliser le changement: les voies du français. http://www.arts.uottawa.ca/voies/voies_fr.html.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, UK.
- Paul Meyer. 1901. *Le Roman de Flamenca*. Librarie Emile Bouillon.
- Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea.
- Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.
- Achim Stein. 2008. Syntactic annotation of Old French text corpora. *Corpus*, 7:157–161.
- Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, pages 3405–3408, Marrakech, Morocco.
- Taro Watanabe and Eiichiro Sumita. 2003. Example-based decoding for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 9–12, Boston, MA.
- David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, PA.

Digitised Historical Text: Does it have to be mediOCRe?

Bea Alex, Claire Grover, Ewan Klein and Richard Tobin

Institute for Language, Cognition and Computation

School of Informatics, University of Edinburgh

10 Crichton Street, EH8 9AB, UK

{balex|grover|ewan|richard}@inf.ed.ac.uk

Abstract

This paper reports on experiments to improve the Optical Character Recognition (OCR) quality of historical text as a preliminary step in text mining. We analyse the quality of OCRed text compared to a gold standard and show how it can be improved by performing two automatic correction steps. We also demonstrate the impact this can have on named entity recognition in a preliminary extrinsic evaluation. This work was performed as part of the TRADING CONSEQUENCES project which is focussed on text mining of historical documents for the study of nineteenth century trade in the British Empire.

1 Introduction

The task of applying text mining techniques to digitised historical text faces numerous hurdles. One of the most troublesome of these is the ‘garbled’ nature of the plain text which often results when the scanned original undergoes Optical Character Recognition (OCR). In this paper we discuss two areas which cause problems: soft-hyphen splitting of word tokens and “long s”-to-f confusion. We evaluate the extent to which both issues degrade the accuracy of OCRed text compared to all OCR errors and describe methods for automatically correcting them.

A representative example of scanned text is shown in Figure 1, followed by the plain text output from OCR and the manually corrected gold standard text.

(4)

BEING sensible therefore, that the committee had been amused by partial representations ; that a much more extensive trade may be established in Hudson's-Bay, both for pelts and furs ; that there are great appearances of valuable mines along the coast; and that a profitable fishery for whales, seals, &c. might be

Example 1: Fragment of scanned page from Robinson (1752)¹

1 (4) BEING sensible therefore, that the
2 committee had been amused by partial
3 representations ; that a much more
4 extensive trade may be established in
5 Hudson's-Bay, both for pelts and furs;
6 that there are great appearances of
7 valuable mines along the coast; and
8 that a profitable fishery for whales,
9 seals, &c. might be

_____ OCR output _____

1 (4) BEING sensible therefore, that the
2 committee had been amused by partial
3 representations ; that a much more
4 extensive trade may be established in
5 Hudson's-Bay, both for pelts and furs;
6 that there are great appearances of
7 valuable mines along the coast; and
8 that a profitable fishery for whales,
9 seals, &c. might be

_____ manually corrected output _____

As can be seen in Example 1, non-final lowercase letter s appears as “long s” (f or f). Not surprisingly, OCR tends to confuse this with

¹<http://eco.canadiana.ca/view/oocihm.20155/18?r=0&s=1>

the lowercase letter *f*, since the only distinction between the two letters is that the long *s* has a nub on the left side of the letter whereas the real lowercase *f* has a nub on both sides. However, we need to correct this conflation of *s* with *f* in order to achieve reasonable accuracy in text mining.

A second issue can be seen in line 8 of the OCR output, where a line-break hyphen from the input text persists as a within-word hyphen, e.g., as **pro-fitab**le (abstracting away from the additional insertion of a period and a space). Although the OCR has correctly recognised this ‘soft’ hyphen, it is still desirable to remove it in order to increase text mining accuracy. This removal can be regarded as a normalisation rather than correction *per se*.

The background and related work to our research are described in Sections 2 and 3, respectively. We have developed two tools which tackle the two issues mentioned above as accurately as possible; note that not every hyphen can be deleted and not every *f* should be turned into *s*. In both cases, we use a lexicon-based approach which is explained in more detail in Section 5. We evaluate the tools against a human corrected and normalised gold standard described in Section 4 in an attempt to quantify OCR accuracy and improvement. We describe the evaluation metric in Section 6 and report all the experiments we have conducted in Section 7. We include a preliminary extrinsic evaluation to determine the effect of text accuracy on named entity recognition.

2 Background

The TRADING CONSEQUENCES project aims to assist environmental historians in understanding the economic and environmental consequences of commodity trading during the nineteenth century. We are applying text mining to large quantities of historical text, converting unstructured textual information into structured data that will in turn populate a relational database. Prior historical research into commodity flows has focused on a small number of widely traded natural resources. By contrast, this project will pro-

vide historians with data from large corpora of digitised documents, thereby enabling them to analyse a broader range of commodities.

We analyse textual data from major British and Canadian datasets, most importantly the House of Commons Parliamentary Papers (HCPP),² the Canadiana.org data archive,³ the Foreign and Commonwealth Office Collection from JSTOR⁴ and a number of relevant books. Together these sources amount to millions of pages of text. The datasets include a wide range of official records from the British and Canadian governments, making them ideal for historical text mining. However, there are significant challenges in the initial step of transforming these document collections into a format that is suitable for subsequent text mining. Poor OCR quality is a major factor, together with artefacts introduced by the scanning process and nineteenth century language. For much of the corpus, the OCR was carried out several years ago, and is far inferior to what can be achieved nowadays with contemporary scanning hardware and OCR technology. The problems of OCR are aggravated for our corpus by the use of old fonts, poor print and paper quality, and nineteenth century language.

The project’s underlying text mining tools are built on the LT-XML⁵ and LT-TTT⁶ tools. While they are robust and achieve state-of-the-art results for modern digital newspaper text, their output for historical text will necessarily involve errors. Apart from OCR imperfections, the data is not continuous running text but passages interspersed with page breaks, page numbers and headers and occasionally hand-written notations in page margins. In order for our text mining tools to pull out the maximum amount of information, we are carrying out automatic correction of the text as a preliminary processing step in our text mining pipeline.

²<http://parlipapers.chadwyck.co.uk/home.do>

³<http://www.canadiana.ca>

⁴<http://www.jstor.org/>

⁵<http://www.ltg.ed.ac.uk/software/ltxml2>

⁶<http://www.ltg.ed.ac.uk/software/lt-ttt2>

3 Related Work

Previous research on OCR post-correction includes use of a noisy channel model, where the true sequence of characters is generated given the noisy OCR output (Kolak and Resnik, 2002). Other studies have focussed on combining the output of multiple OCR systems to improve the text through voting (Klein and Kopel, 2002), efficient text alignments combined with dictionary lookup (Lund and Ringger, 2009) and merging outputs by means of a language model (Volk et al., 2011) or by active learning of human post-editing (Abdulkader and Casey, 2009).

Recent work has suggested improving OCR by making use of online search engine spelling corrections (Bassil and Alwani, 2012). While this has shown substantial reductions in the OCR error rates for English and Arabic text, the evaluation data sets are small with only 126 and 64 words, respectively. Combining dictionary lookup and querying candidates as part of trigrams in a search engine was also proposed by Ringlstetter et al. (2005), specifically to correct alphabet confusion errors in mixed-alphabet documents.

With specific focus on processing historical documents, there have been recent initiatives to improve OCR quality through manual correction via user collaboration. For example, the Australian Newspaper Digitisation Program set up an experiment to let the public correct the OCR output of historical documents, but found it difficult to measure the quality of the corrected text (Holley, 2009a; Holley, 2009b). The IMPACT project (Neudecker and Tzadok, 2010) is aimed at developing tools to improve OCR results via crowd sourcing to improve the digitisation of historical printed text. Related to this is the ongoing TEXTUS project⁷ which plans to develop an open source platform for users to read and collaborate around publicly available texts. One of its planned functionalities is a mechanism that enables scholars to transcribe plain text versions of scanned documents.

The more specific issue of “long s” to *f* con-

version has been addressed in an interesting but uncredited blog post.⁸ This shows that a simple rule-based algorithm that maps *f*-containing strings not in the dictionary to words that are listed in the dictionary, improves OCRed text sufficiently to in turn improve recognition of taxon entities using TaxonFinder.⁹ There is no detailed information on the dictionaries used, but we assume that they include general English dictionaries as well as specialised taxon dictionaries given that such terms are in Latin.

The second specific issue, namely fixing end-of-line hyphenation, has been addressed by Torget et al. (2011), who delete hyphens automatically in contexts s_1-s_2 whenever s_2 is not in the dictionary while the string with the hyphen omitted, namely s_1s_2 , is in the dictionary. They do not provide information on how well their method performs.

There has also been work on determining the effect of OCR accuracy on text processing, be it information retrieval (IR) or text mining. In terms of IR, direct access to historical documents is hindered through language change and historical words have to be associated with their modern variants in order to improve recall. Hauser et al. (2007), for example, designed special fuzzy matching strategies to relate modern language keywords with old variants in German documents from the Early New High German period. Gotscharek et al. (2011) argue that such matching procedures need to be used in combination with specially constructed historical lexica in order to improve recall in IR. Reynaert (2008) proposes a language-independent method to clean high-frequency words in historical text by gathering typographical variants within a given Levenshtein distance combined with text-induced filtering.

OCR errors have been shown to have a negative effect on natural language processing in general. Lopresti (2005; 2008a; 2008b), for example, examines the effect that varying degrees of OCR accuracy have on sentence boundary detection, tokenisation and part-of-

⁸<http://inthefaith.net/rdp/botanicus/>

⁹<http://taxonfinder.sourceforge.net/>

speech tagging, all steps which are typically carried out as early stages of text mining. Ko-lak and Resnik (2005) carried out an extrinsic evaluation of OCR post-processing for machine translation from Spanish into English and show that translation quality increases after post-correcting the OCRed text.

4 Data Preparation

We limited our analysis to the Early Canadiana Online collections since they contain a reasonably large number ($\sim 83k$) of different documents and since we wanted to get an idea of their OCR quality. We randomly selected a set of records from Canadiana, where a *record* is the OCRed counterpart of a page in the source document. More specifically, we shuffled the list of all Canadiana documents and randomly selected a record from the first 1,000 documents. These records were drawn from the first 20 records in the document; this limitation was due to the fact that Canadiana only provides free access to the first few scanned pages of a document, and our annotator needed to be able to access these in order to correct records and create a gold standard. When selecting a random record per document, we imposed some further restrictions, namely that the record had to be marked as English (`lang="eng"`) and it had to be more than 150 characters long. The latter length limitation was simply applied to avoid lots of short titles since we were mostly interested in running text. We also programmatically excluded records resembling tables of content or technical notes and disclaimers. This resulted in a set of passages which we gave to the human annotator to correct. For Experiment 1, described in Section 7.1, we used subsets of this random set: the original OCRed records (from now on referred to as ORIGINAL), the same records corrected and annotated as a gold standard (GOLD) and finally the records containing automatically corrected OCRed text (SYSTEM).

4.1 Preparing a Gold Standard

We asked the annotator to spend one day correcting and normalising records in ORIG-

INAL to create readable English text, leaving historical variants as they are but removing soft hyphens at the end of lines and changing wrongly recognised *f* letters into *s* as well as correcting other OCR errors in the text to the best of their ability. The annotator commented that while some records were relatively easy to correct, most required looking at the scanned image to determine what was intended. We asked the annotator to ignore any records which contained text in other languages, i.e., where the `lang` attribute in the original data was assigned wrongly. The annotator also ignored records which were so garbled that it would be quicker to write the record from scratch rather than correct the OCR output.

In total, the annotator corrected 25 records from ORIGINAL. These contained 8,322 word tokens, reduced to 7,801 in GOLD (including punctuation), when tokenised with our in-house English tokeniser. An illustration of the annotator’s input and output was shown in Section 1.

While “long *s*” confusion was found to be an issue in only four of the 25 records, the soft hyphen splitting issue was pervasive throughout the ORIGINAL dataset.

4.2 Preparing System Output

Finally, we processed the uncorrected versions of the records that comprised the GOLD set, using our tools to automatically remove soft hyphens and to convert incorrect *f* letters to *s* before tokenising the text to create the SYSTEM set. The methods of both these steps are described in detail in the next section, while their performance is evaluated in Section 7.

5 Automatic OCR Post-Correction

Our automatic OCR post-correction and normalisation involves two steps:

1. Removing end-of-line hyphens if they are soft hyphens; i.e., they hyphenate words that would normally not be hyphenated in other contexts.
2. Correcting *f* letters to *s* if they were a “long *s*” in the original document.

5.1 End-of-line Soft Hyphen Deletion

The program for removing end-of-line soft hyphens is called `lxdehyphen`. It expects an XML file containing elements corresponding to lines of text. If all lines of text in a page are concatenated into one in the OCR output, then the text is first split at tokens ending in hyphen + whitespace + new token by inserting an artificial newline character. The program then tokenises the text simply using whitespace and newline as delimiters to identify hyphenated words; a more sophisticated tokenisation is not needed for this step. If the last token on a line ends with a hyphen it is considered as a candidate for joining with the first token of the next line. The tokens are joined if, after removing the hyphen and concatenating them, the result is either a word that appears in the Unix/Linux system dictionary `dict`,¹⁰ or is a word that appears elsewhere in the document. The latter heuristic, which can be considered as using the document itself as a dictionary, is very effective for documents with technical terms and names that do not appear in `dict` as well as for historical documents which contain historical variants of modern terms. Provided that a word appears somewhere else in the document unhyphenated, it will be recognised and the soft hyphen will be removed.

The tokenisation markup is then deleted and a more sophisticated tokenisation can be carried out. If soft hyphen splits have been removed, this will typically result in a reduction in the number of tokens.

5.2 $f \rightarrow s$ Character Conversion

The crucial component of the $f \rightarrow s$ conversion tool is `fix-spelling`, an `lxtransduce` grammar that replaces words based on a lexicon that maps misspelled words to the correct version. Lexicons can be constructed for various purposes; in this case, we use the lexicon `f-to-s.lex` for correcting poorly OCRed historical documents where the “long s” has been

¹⁰The dictionary (`/usr/share/dict/words`) can vary between operating systems. We ran our experiments using the Scientific Linux release 6.2 with a dictionary containing 479,829 entries.

conflated with f . The `f-to-s.lex` lexicon is created from a corpus of correct text. For each word in that corpus, a word frequency distribution is collected and all the possible misspellings caused by the long- s -to- f confusion are generated. It is possible that some of these generated words will also be real words (e.g., *fat* < *sat*).¹¹

The unigram frequency counts of each word in the corpus is therefore used to determine its likelihood. For example, *difclose* will be corrected to *disclose* because *difclose* does not occur in the corpus. *fat* will be corrected to *sat* because *sat* occurs more often. But *feed* will not be changed to *seed* because *feed* occurs more often. The corpus can be chosen to be similar to the target texts so that the results are more reliable; in particular, using old texts will prevent words that were not common then from being incorrectly used. In our experiment, we used the text from a number of books in the Gutenberg Project as the corpus to create the lexicon.¹²

6 Text Alignment and Evaluation

In the set of experiments described below, we compare different versions of a given text to determine a measure of text accuracy. In order to carry out the evaluation, we first need to align two files of text and then calculate their differences. OCR software suppliers tend to define the quality of their tools in terms of character accuracy. Unfortunately, such figures can be misleading if the digitised text is used in an information retrieval system, where it is searched, or if it is processed by a text mining tool. An OCRed word token that contains only a single character error (i.e., insertion, deletion or substitution) will score more highly on this measure than one which contains multiple such errors. However, the word will still be incorrect when it

¹¹We use $w_1 < w_2$ to indicate that string w_1 has been derived from w_2 , either by one of our tools, as in this case, or by OCR.

¹²The books were *Adventures of Sherlock Holmes*, *Christmas Carol*, *Dracula*, *Great Expectations*, *Hound of the Baskervilles*, *Paradise Lost*, *Tale of Two Cities*, *The Adventures of Sherlock Holmes*, *The History of England*, vol. 1 and *Works of Edgar Allan Poe*, vol. 1.

comes to analysing and processing it automatically, especially as many text processing tools involve dictionary, gazetteer or other exact-match word-lookup methods.

In this paper, therefore, we evaluate the quality of the converted text in terms of word error rate (WER):

$$\text{WER} = \frac{I + D + S}{N}$$

where I is the number of insertions, D the number of deletions and S the number of substitutions between the hypothesis and the reference string, while N is the number of word tokens in the reference (e.g., the GOLD dataset). We calculate WER by means of the GNU **wdiff** program,¹³ a front end to **diff**¹⁴ for comparing files on a word-per-word basis, where a word is anything between whitespace or newline characters. **diff** aligns a sequence of text by determining the longest common subsequence (Miller and Myers, 1985; Myers, 1986; Ukkonen, 1985). **wdiff** then determines the counts for the words in common between both files as well as word insertions, deletions and substitutions. Note that **wdiff** counts a word as a substitution if it is replaced or is part of a larger replacement. This is not an issue for our evaluation as we are not interested in the distribution of insertions, deletions and substitutions but merely their sum.

7 Experiments

The following first three experiments report on the quality of Canadiana’s OCRed text, and the extent to which our processing tools can improve it. The forth experiment provides an initial examination of how the OCR quality affects the recognition of commodity entity mentions, the latter being one of the aims of the TRADING CONSEQUENCES project.

7.1 Experiment 1

In Experiment 1, we first compare the difference between the ORIGINAL and the GOLD data set. The difference between these two

versions of the text shows how much the original OCRed text needs to change to be transformed to running English text without errors. We then run both the **lxdehyphen** and the $f \rightarrow s$ conversion step over the original text and create the **SYSTEM_{all}** data set. By comparing the **SYSTEM_{all}** against GOLD and determining their differences, we can then calculate the improvement we have made to the ORIGINAL data in light of all the changes that could be made.

Test Set	I	D	S	WER
ORIGINAL	71	24	1,650	0.224
SYSTEM _{all}	73	24	1,434	0.196
SYSTEM _{dehyph}	74	24	1,519	0.207
SYSTEM _{f2s}	70	24	1,567	0.213

Table 1: Number of insertions (I), deletions (D) and substitutions (S) and word error rate (WER) when comparing the different test sets against the GOLD dataset of 7,801 tokens.

Table 1 shows that the ORIGINAL OCRed text has a word error rate of 0.224 compared to the corrected and normalised GOLD version. After running the two automatic conversion steps over the ORIGINAL data set creating **SYSTEM_{all}**, word error rate was reduced by 12.5% to 0.196. Given that the GOLD set was created by correcting all the errors in the OCR, this is a substantial improvement.

We also ran each correction step separately to see how they each improve the OCR individually. The scores for **SYSTEM_{dehyph}** and **SYSTEM_{f2s}** show that the soft hyphen removal step contributed to reducing the error rate by 7.6% (0.017) to 0.207 and $f \rightarrow s$ conversion by 4.9% (0.011) to 0.213. Soft hyphens are an issue throughout the text but long- s -to- f confusion only occurred in four out of the 25 records, which explains why the effect of the latter step is smaller. The results show that after fixing the OCR for two specific problems automatically a large percentage of errors (87.5%) still remain in the text.

7.2 Experiment 2

Since we evaluate each conversion tool in light of all other corrections of the OCRed text, it is

¹³<http://www.gnu.org/software/wdiff/>

¹⁴<http://www.gnu.org/software/difftools/>

quite difficult to get an idea of how accurate they are. We therefore evaluated each tool separately on a data set which was only corrected for their particular issue. For Experiment 2, we therefore hand-corrected the 25 Canadiana records again, but this time only for soft hyphen deletion, ignoring all other issues with the OCR. We call this data set GOLD_{dehyp} and compare it to ORIGINAL to determine the difference between both texts. We also compare GOLD_{dehyp} to SYSTEM_{dehyp} created in the previous experiment.

Test Set	I	D	S	WER
ORIGINAL	1	0	177	0.022
SYSTEM_{dehyp}	1	0	47	0.006

Table 2: Number of insertions (I), deletions (D) and substitutions (S) and word error rate (WER) when comparing against GOLD_{dehyp} (8,203 tokens).

The results in Table 2 show that soft hyphens in the OCRed text reduce the word token accuracy by 0.022 compared to the normalised gold standard. The automatic fixing reduces this error by 72.7% to 0.006. Error analysis showed that the remaining differences between SYSTEM_{dehyp} and GOLD_{dehyp} are caused by missing soft hyphen deletion because the split tokens in question contained other OCR problems and were either not present in the dictionary or not repeated in the text itself, e.g., *3ndow- ment* (< *endowment*) or *patron- aye* (< *patronage*). The $1xdehyphen$ tool correctly did not remove hyphens which were meant to stay in the text.

7.3 Experiment 3

As with Experiment 2, we wanted to separately evaluate the performance of the $f \rightarrow s$ conversion step. Since the GOLD data set only contained four page records with the “long s ” confusion, we created a bigger gold standard of ten Canadiana page records which all have the issue in their original OCR and which the annotator hand-corrected only for $f \rightarrow s$ ignoring all other errors in the text. We first compare the original text of this data set (f2s_ORIGINAL) to the hand-corrected one

(f2s_GOLD) to quantify the effect of the problem on the text. We then ran the automatic $f \rightarrow s$ conversion over the f2s_ORIGINAL set and compared the output (f2s_SYSTEM f_{2s}) to the gold standard (f2s_GOLD).

Test Set	I	D	S	WER
f2s_ORIGINAL	0	0	720	0.095
f2s_SYSTEM f_{2s}	0	0	206	0.027

Table 3: Number of insertions (I), deletions (D) and substitutions (S) and word error rate (WER) when comparing against f2s_GOLD (7,618 tokens).

The results in Table 3 show that the character confusion problem increases the word error rate dramatically by 0.0945. This means that OCRed documents in which this issue occurs will benefit significantly from an automatic post-correction step that fixes this problem. Our own system output reduces the error by 71.6%, yielding a word error rate of 0.027. An error analysis of the remaining differences shows that, as with the previous experiment, the tool does not successfully convert false f letters in tokens containing other OCR issues, e.g., *adminiftzr* (< *administer*) or *fireet* (< *street*). In a few cases, it also causes some real-word errors, i.e., converting $f \rightarrow s$ when it should not have done so. This happens either for tokens where a letter other than s was confused with f during the OCR process, e.g., *fhe* (< *the*) is converted to *she*, or for tokens whose unigram frequency is smaller than that of the converted token, e.g., *fees* was incorrectly converted to *sees*.

7.4 Experiment 4

After the annotator finished correcting all of the 25 GOLD page records, she was asked to mark up commodity entities in that data set. Our definition of a commodity entity is something that is sold or traded and can be either natural or man-made. The annotator marked up 167 entities including general commodities such as *apples* and *copper* but also more specific items such as *Double Ended Taper Saw Files* or *Iron Girders*. We then transferred the annotations manually to the ORIGINAL and the SYSTEM all data as accurately as possible.

Our next task in the TRADING CONSEQUENCES project is to develop a named entity recogniser which identifies commodities that were important in nineteenth century trade. This work is being carried out in collaboration with historians at The University of York, Toronto, and involves creating a commodities thesaurus/ontology using the SKOS framework.¹⁵ This thesaurus will be the basis for our further system development. In the meantime, we are using WordNet¹⁶ to approximate commodity terms; that is, we use a chunker to recognise noun chunks in the text and label them as commodity mentions if they are a hyponym of the WordNet classes **substance**, **physical matter**, **plant** or **animal**.

Test Set	Found	Correct
ORIGINAL	79	28
GOLD	86	34
SYSTEM _{all}	80	29

Table 4: Number of found and correct commodity entities in the various test sets.

Even though this is a very crude method with a low performance, the effects of the OCR are apparent in the results shown in Table 4. The smallest number of entities and of correct entities are found in the uncorrected ORIGINAL set. 7 more entities (of which 6 are correct) are recognised in the completely corrected and normalised GOLD set. The two automatic correction steps which reduce the distance from ORIGINAL to GOLD by 12.2% lead to one additional correct entity being recognised compared to the original OCR. While this means that more OCR post-corrections would be desirable to improve the named entity recognition, we also believe that the effect of both existing post-correction tools will become more apparent as we improve our commodities recognition system.

8 Discussion and Conclusion

It is widely recognised that much of the OCRed text currently available for historical docu-

ments falls far short of what is required for accurate text processing or information retrieval. We have focussed on automatically fixing two issues in such text, namely soft hyphen deletion and $f \rightarrow s$ conversion. We have evaluated both methods and shown that together they deal with just over 12% of all word error problems in our sample. In addition, each of them successfully deals with around 72% of relevant cases. We have carried out an error analysis of where the tools fail to yield the correct results. Finally, we have described a very preliminary study which indicates that fixing and normalising the OCRed text is beneficial to named entity recognition.

As we have seen, even after these steps, a large number of OCR errors remain in the text. We will need to address at least some of these to achieve our desired level of text mining performance. As part of this task, we intend to explore whether some of the techniques we briefly reviewed in Section 3 can be combined with our current approach.

When creating the gold standard data, we found that the quality of the OCRed text from different sources can vary depending on factors such as the quality of the scan, the quality of original script and printing, the contents of a page, and so one. In a few cases the text we provided to the annotator turned out to be too garbled to understand. If humans are unable to correct such records then automatic systems will have little chance of doing so, and text mining will produce no or very useless information. We are therefore planning to integrate a further pre-processing step into our system which tries to estimate text accuracy and rejects documents from processing whose accuracy falls below a certain threshold.

Acknowledgments

We would like to thank our TRADING CONSEQUENCES project partners at The University of York, Toronto, The University of St.Andrews and EDINA for their feedback (Digging Into Data CIINN01). We would also like to thank Clare Llewellyn for her valuable input and help with the annotation.

¹⁵<http://www.w3.org/2004/02/skos/>

¹⁶<http://wordnet.princeton.edu/>

References

- Ahmad Abdulkader and Mathew R. Casey. 2009. Low cost correction of OCR errors using learning in a multi-engine environment. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 576–580, Washington, DC. IEEE Computer Society.
- Youssef Bassil and Mohammad Alwani. 2012. OCR post-processing error correction algorithm using Google's online spelling suggestion. *Journal of Emerging Trends in Computing and Information Sciences*, 3(1):90–99, January.
- Annette Gotscharek, Ulrich Reffle, Christoph Ringlstetter, Klaus U. Schulz, and Andreas Neumann. 2011. Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *IJDAR*, 14(2):159–171.
- Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, and Christiane Wanzeck. 2007. Information access to historical documents from the Early New High German period. In L. Burnard, M. Dobreva, N. Fuhr, and A. Lüdeling, editors, *Digital Historical Corpora- Architecture, Annotation, and Retrieval*, Dagstuhl, Germany.
- Rose Holley. 2009a. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4).
- Rose Holley. 2009b. Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers. National Library of Australia, Technical Report.
- Shmuel T. Klein and Miri Kopel. 2002. A voting system for automatic OCR correction. In *Proceedings of the Workshop On Information Retrieval and OCR at SIGIR*, pages 1–21.
- Okan Kolak and Philip Resnik. 2002. OCR error correction using a noisy channel model. In *Proceedings of the second international conference on Human Language Technology Research*, pages 257–262.
- Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 867–874.
- Daniel Lopresti. 2005. Performance evaluation for text processing of noisy inputs. In *Proceedings of the Symposium on Applied Computing*, pages 759–763.
- Daniel Lopresti. 2008a. Measuring the impact of character recognition errors on downstream text analysis. In B. A. Yanikoglu and K. Berkner, editors, *Document Recognition and Retrieval*, volume 6815. SPIE.
- Daniel Lopresti. 2008b. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data*, pages 9–16.
- William B. Lund and Eric K. Ringger. 2009. Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL'09, pages 231–240.
- Webb Miller and Eugene W. Myers. 1985. A file comparison program. *Software - Practice and Experience*, 15(11):1025–1040.
- Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1:251–266.
- Clemens Neudecker and Asaf Tzadok. 2010. User collaboration for improving access to historical texts. *LIBER Quarterly*, 20(1).
- Martin Reynaert. 2008. Non-interactive OCR post-correction for giga-scale digitization projects. In *Proceedings of the 9th international conference on Computational Linguistics and Intelligent Text Processing*, pages 617–630.
- Christoph Ringlstetter, Klaus U. Schulz, Stoyan Mihov, and Katerina Louka. 2005. The same is not the same—postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. In *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 406–410.
- Joseph Robson. 1752. *An Account of Six Years Residence in Hudson's-Bay from 1733 to 1736, and 1744 To 1747*. Printed for J. Payne and J. Bouquet, London.
- Andrew J. Torget, Rada Mihalcea, Jon Christensen, and Geoff McGhee. 2011. Mapping texts: Combining text-mining and geovisualization to unlock the research potential of historical newspapers. University of North Texas Digital Library, White Paper.
- Esko Ukkonen. 1985. Algorithms for approximate string matching. *Information and Control*, 64(1-3):100–118.
- Martin Volk, Lenz Furrer, and Rico Sennrich. 2011. Strategies for reducing and correcting OCR errors. In C. Sporleder, A. Bosch, and K. Zervanou, editors, *Language Technology for Cultural Heritage*, chapter 1, pages 3–22. Springer-Verlag, Berlin/Heidelberg.

Comparison of Named Entity Recognition tools for raw OCR text

Kepa Joseba Rodriguez

Göttingen State and
University Library
rodriguez@
sub.uni-goettingen.de

Mike Bryant

Centre for e-Research
King's College London
michael.bryant@
kcl.ac.uk

Tobias Blanke

Centre for e-Research
King's College London
tobias.blanke@
kcl.ac.uk

Magdalena Luszczynska

University College
London
zclhd0@
ucl.ac.uk

Abstract

This short paper analyses an experiment comparing the efficacy of several Named Entity Recognition (NER) tools at extracting entities directly from the output of an optical character recognition (OCR) workflow. The authors present how they first created a set of test data, consisting of raw and corrected OCR output manually annotated with people, locations, and organizations. They then ran each of the NER tools against both raw and corrected OCR output, comparing the precision, recall, and F1 score against the manually annotated data.

1 Introduction¹

While the amount of material being digitised and appearing online is steadily increasing, users' ability to find relevant material is constrained by the search and retrieval mechanisms available to them. Free text search engines, while undoubtedly of central importance, are often more effective and friendly to users when combined with structured classification schemes that allow faceted search and iterative narrowing-down of large result sets. Such structured classification, however, is costly in terms of resources. Named Entity Recognition (NER) holds the potential to lower this cost by using natural language processing tools to automatically tag items with people, places, and organizations that may be used as access points.

The European Holocaust Research Infrastructure (EHRI)² project aims to create a sustainable research infrastructure bringing together resources from dispersed historical archives across different countries. As part of this effort, we are investigating if and how we can enhance the research experience by offering tools to EHRI project partners that enable them to semantically enrich sources that they contribute to the wider infrastructure. A component of these services is currently planned to be a flexible OCR service based on the workflow tools developed at King's College London for the JISC-funded Ocropodium project (Blanke et al., 2011). While OCR itself is a complex process that produces hugely variable results based on the input material, modern tools are capable of producing reasonable transcripts from mid 20th-century typewritten material that is quite common among the EHRI project's primary sources. The quality of these transcripts prior to correction would be considered quite low for human readers, but even when uncorrected can offer value for search indexing and other machine processing purposes (Tanner et al., 2009). This experiment was undertaken to evaluate the efficacy of some available tools for accurately extracting semantic entities that could be used for automatically tagging and classifying documents based on this uncorrected OCR output.

Section 2 briefly summarises the existing literature covering the use of OCR and named entity extraction in the field of historical research. Section 3 describes our methodology for scoring each of the trialled NER tools in terms of pre-

¹A version of this paper was presented in the proceedings of DH2012.

²<http://www.ehri-project.eu>

cision, recall, and F1. Section 4 describes the materials on which we conducted the experiment. Section 5 presents the results of our experiment and some analysis of those results, while section 6 concludes and discusses the research we intend to conduct in the future.

2 Literature Review

Grover et al. (2008) describe their evaluation of a custom rule-based NER system for person and place names on two sets of British Parliamentary records from the 17th and 19th centuries. They describe many of the problems encountered by NER systems that result from both OCR artefacts and the archaic nature of the sources themselves, such as conflation of marginal notes with body text, multi-line quoting rules, and capitalization of common nouns.

Packer et al. (2010) tested three different methods for the extraction of person names from noisy OCR output, scoring the results from each method against a hand-annotated reference. They noted a correlation between OCR word error rate (WER) and NER quality did exist, but was small, and hypothesised that errors in the OCR deriving from misunderstanding of page-level features (i.e. those that affect the ordering and context of words) have a greater impact on NER for person names than character-level accuracy.

3 Methodology

The goal of our experiment is the evaluation of the performance of existing NER tools on the output of an open source OCR system. We have selected four named entity extractors:

- a) OpenNLP³
- b) Stanford NER (Finkel et al., 2005)⁴
- c) AlchemyAPI⁵
- d) OpenCalais⁶

The entity types we chose to focus on were person (PER), location (LOC) and organization (ORG). These entity types are the most relevant

for the extraction of metadata from documents in the domain of the EHRI project, and have a good coverage by the selected tools.

The named entity extractors use different tagsets for the annotation of named entities. We have selected the different categories and normalized the annotation as follows:

- a) Stanford NER and OpenNLP: Person, location and organization categories have been annotated for all used models.
- b) OpenCalais: we have merged the categories "Country", "City" and "NaturalFeature" into the category LOC, and the categories "Organization" and "Facility" into the category ORG. This second merging was possible because in our domain almost all facilities that appear were official buildings and public facility buildings referenced in an organizational capacity.
- c) Alchemy: We have merged the categories "Organization", "facility" and "Company" into the category ORG and "Country" , "City" and "Continent" into LOC.

To our knowledge there is at the moment no freely available corpus that can be used as gold standard for our purposes. For this reason we have produced a manually annotated resource with text extracted from images of documents related to the domain of the EHRI project as described in section 4. We have compared the output of the tools with this standard and computed the precision, recall and F1.

4 Experimental Setup

The material used for our initial experiments was obtained from the Wiener Library, London, and King's College London's Serving Soldier archive.

The Wiener Library material consisted of four individual Holocaust survivor testimonies, totalling seventeen pages of type-written monospaced text. Since the resolution of the scans was low and the images quite "noisy" (containing hand-written annotations and numerous scanning artifacts) we did not expect the quality of the OCR to be high, but felt they were suitably representative of material with which we expected to be dealing with in the EHRI project.

³<http://opennlp.apache.org>

⁴<http://nlp.stanford.edu/software/CRF-NER.shtml>

⁵<http://www.alchemyapi.com>

⁶<http://www.opencalais.com>

The OCR workflow used only open-source tools, employing several preprocessing tools from the Ocropolis toolset (Breuel, 2008) for document cleanup, deskew, and binarisation, and Tesseract 3 (Smith, 2007) for character recognition. We also found that scaling the images to four times their original size using an anti-aliasing filter made a big positive difference to the quality of the OCR output. Ultimately, the word accuracy of the OCR averaged 88.6%, although the title pages of each survivor testimony, which resembles an archival finding aid with brief summaries of the contents in tabular form, were considerably worse, with only one exceeding 80% word accuracy. Character accuracy was somewhat higher, averaging 93.0% for the whole corpus (95.5% with the title pages excluded.)

The Serving Soldier material consisted of 33 newsletters written for the crew of H.M.S. Kelly in 1939. Like the Wiener Library documents, they also comprise type-written monospaced text, but have a strong yellow hue and often-significant warping and skew artifacts. The OCR workflow also used Ocropolis for preprocessing and Tesseract 3 for character recognition, with most attention being paid to the thresholding. Word accuracy averaged 92.5% over the whole set, albeit with fairly high variance due to factors such as line-skew.

The corpus was constructed by manually correcting a copy of the OCR output text, in order to evaluate the impact of the noise produced by the OCR in the overall results. Both copies of the corpus, corrected and uncorrected, were then tokenized and POS tagged using the TreeTagger⁷ (Schmid, 1994) and imported in the MMAX2 annotation tool (Müller and Strube, 2006), before being delivered to human annotators. In order to support the human annotation process we have used the POS tags to produce a pre-selection of annotation markable candidates, selecting all words tagged as PPN⁸ and building markables with chains of these words that are not separated by punctuation marks.

We have tested the reliability of our annotation carrying an agreement study. For the study we

⁷<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>

⁸Proper Name

	Corpus WL		Corpus KCL	
	Raw	Corr	Raw	Corr
Files	17	17	33	33
Words	4415	4398	16982	15639
PER	75	83	82	80
LOC	60	63	170	178
ORG	13	13	52	60
TOTAL	148	159	305	319

Table 1: Composition of test data

	AL	OC	not corrected			corrected			P
			P	R	F1	P	R	F1	
AL	PER	OC	.51	.36	.42	.46	.32	.38	-
	LOC		.92	.43	.59	.96	.51	.67	-
	ORG		.36	.31	.33	.50	.23	.32	-
	TOTAL		.61	.38	.47	.63	.38	.48	-
ON	PER	ON	.69	.30	.42	.62	.18	.28	-
	LOC		.78	.20	.32	.76	.49	.60	-
	ORG		.40	.15	.22	.50	.15	.24	-
	TOTAL		.75	.29	.41	.69	.30	.42	-
ST	PER	ST	.33	.04	.08	.50	.06	.10	-
	LOC		.56	.20	.29	.62	.25	.35	-
	ORG		.27	.23	.25	.20	.08	.11	-
	TOTAL		.42	.12	.19	.53	.13	.21	-

Table 2: Output of NERs the Wiener Library dataset

have used the Kappa coefficient (Carletta, 1996) between two native English speaker annotators, and we get a value of $K = .93^9$.

The composition of the corpus has been summarized in table 1.

5 Results

The results of the different NER tools have been summarized in table 2 for the Wiener Library's dataset and in table 3 for the King's College London's dataset¹⁰. The last row shows the statistical significance¹¹ of the differences between the performance of the NER tools on corrected and uncorrected text calculated globally and for each entity types. The stars indicate degrees of significance: one star signifies $0.025 \leq p \geq 0.05$; two stars $0.01 \leq p < 0.025$; three stars $p < 0.01$. We use a dash when differences are not statistically significant.

A first observation of the results for corrected and non-corrected text shows that the correction

⁹Values of $K > .8$ show that the annotation is reliable to draw definitive conclusions.

¹⁰The meaning of the abbreviations is: AL Alchemy, OC for OpenCalais, ON for OpenNLP, ST for Stanford NER.

¹¹We used the paired t-test to compute the statistical significance.

		not corrected			corrected			<i>p</i>
		P	R	F1	P	R	F1	
AL	PER	.15	.22	.18	.24	.32	.27	***
	LOC	.63	.46	.53	.74	.59	.65	***
	ORG	.12	.16	.13	.24	.27	.25	**
	TOTAL	.31	.33	.32	.43	.44	.44	***
OC	PER	.25	.11	.15	.35	.16	.22	-
	LOC	.65	.50	.57	.66	.57	.61	-
	ORG	.14	.20	.16	.22	.27	.18	-
	TOTAL	.42	.33	.37	.46	.39	.42	-
ON	PER	.25	.13	.17	.29	.11	.16	-
	LOC	.65	.29	.40	.69	.33	.44	-
	ORG	.10	.24	.14	.14	.27	.18	-
	TOTAL	.28	.23	.25	.33	.25	.28	-
ST	PER	.18	.25	.21	.29	.37	.33	***
	LOC	.52	.71	.60	.53	.77	.62	-
	ORG	.08	.29	.13	.17	.48	.26	*
	TOTAL	.28	.50	.36	.35	.60	.44	***

Table 3: Output of NERs on the King College London dataset

of the text by hand does not increase the performance of the tools by a significant amount¹². On the other side the improvement is not statistically significant in a reliable way: there is not an entity type for which the performance of the NER increases in a significant way for both datasets for the same NER tool.

In both cases the performance of systems is in the best case modest and some way below the performance reported in the evaluation of the tools. One reason for this is that the kind of text that we use is quite different to the texts usually used to train and evaluate NER systems¹³.

For instance, a significant number of the non-extracted entities of type PER can be explained by their being represented in a variety of different ways in the source text, as for instance [Last name, first name], use of parentheses together with initials in the name - as for instance “Captain (D)” - sometimes written all in capital, or in other non standard ways, and these variants have often been missed by the NER tools. In some cases we feel that these omissions can potentially be resolved with a fairly straightforward heuristic pre-processing of the text.

Another difficult case is when the names of per-

¹²Although we find a small improvement for some of the tools and entity types, it is not enough if we take in account the amount of ours necessary for the correction task. That is a highly resources consuming task that goes beyond the objectives of our project and can not be founded by mass digitization projects

¹³In the case of the (open source) Stanford NER and OpenNLP we know that a mixture of the MUC-6, MUC-7, and CoNLL were used as gold standard corpora. This information is not available for closed source webservices like OpenCalais and AlchemyAPI

sons or locations are used to annotate other cases of entities. For instance in one of our data sets the warship with name “Kelly”, which appears very often in all the files, has consistently been annotated incorrectly as PER, and the same occurs with the name of other warships.

Perhaps unsurprisingly, entities of type ORG proved the most difficult to extract. Organizations referenced in our test datasets (particularly the H.M.S. Kelly newsletters) tended to be once highly relevant organizations that no longer exist, and organizations with abbreviated names. A way to improve the results here could be to employ more external sources of knowledge that can enrich the output of the NE extractors.

An additional, more general, factor that makes NE extractors difficult is the high quantity of spelling errors and irregular spelling in the original files. OpenCalais tries to solve this problem using its knowledge to reconstruct the non-interpretable output. In several cases this reconstruction seems to be successful, but with the risk of introducing entities that don’t appear in the text¹⁴. That suggests that automatically extracted entities should be validated using controlled vocabularies or other kinds of domain knowledge.

In terms of relative results of the various NER tools, a few tentative conclusions can be drawn. Stanford NER gave overall the best performance across both datasets, and was most effective on PER and LOC types. Alchemy API achieved the best results for the ORG type, particularly on manually corrected text, but was not markedly better than OpenCalais or Stanford NER with this data. OpenNLP performed the least accurately of the tools we tested.

6 Conclusions and Future Work

The experiment we describe here is very much linked to practical needs. At this stage we have plans for the implementation of a system that provides OCR and NER workflow facilities for EHRI project partners, and which we hope will provide real-world usage data to augment the results of our trials.

The results indicate that manual correction of OCR output does not significantly improve the

¹⁴For instance the entity “Klan, Walter” was extracted as “Ku Klux Klan”.

performance of named-entity extraction (though correction may of course be valuable for other IR purposes), and that there is scope for improving accuracy via some heuristic preprocessing and the use of domain knowledge.

In the near term our work will likely focus on employing simple heuristics and pattern-matching tools to attempt to improve the quality of the NER for our specific domain. Longer term, we intend to exploit the domain-specific knowledge generated by the EHRI project (in the form of authority files describing people, places, and organizations) to validate and enhance the output of automatic entity recognition tools.

References

- T. Blanke, M. Bryant, and M. Hedges. Ocropodium: open source OCR for small-scale historical archives. *Journal of Information Science*, 38(1):76–86, November 2011. ISSN 0165-5515, 1741-6485. doi: 10.1177/0165551511429418. URL <http://jis.sagepub.com/cgi/doi/10.1177/0165551511429418>.
- Thomas Breuel. The OCropus open source OCR system. In *Proceedings IS&T/SPIE 20th Annual Symposium*, 2008. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.99.8505>.
- Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996. ISSN 0891-2017. URL <http://dl.acm.org/citation.cfm?id=230386.230390>.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL ’05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL <http://dx.doi.org/10.3115/1219840.1219885>.
- Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. Named entity recognition for digitised historical texts. 2008. URL <http://citeseerx.ist.psu.edu/> viewdoc/summary?doi=10.1.1.166.68.
- Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.
- Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi, and Lee S. Jensen. Extracting person names from diverse and noisy OCR text. page 19. ACM Press, 2010. ISBN 9781450303767. doi: 10.1145/1871840.1871845. URL <http://portal.acm.org/citation.cfm?doid=1871840.1871845>.
- Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- Ray Smith. An overview of the tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 2007. URL <http://dl.acm.org/citation.cfm?id=1304846>.
- Simon Tanner, Trevor Muoz, and Pich Hemy Ros. Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, 15(7/8), July 2009. ISSN 1082-9873. doi: 10.1045/july2009-munoz. URL <http://www.dlib.org/dlib/july09/munoz/07munoz.html>.

From Semi-Automatic to Automatic Affix Extraction in Middle English Corpora: Building a Sustainable Database for Analyzing Derivational Morphology over Time

Hagen Peukert

University of Hamburg

hagen.peukert@uni-hamburg.de

Abstract

The annotation of large corpora is usually restricted to syntactic structure and word class. Pure lexical information and information on the structure of words are stored in specialized dictionaries (Baayen et al., 1995). Both data structures – dictionary and text corpus – can be matched to get e.g. a distribution of certain (restricted) lexical information from a text. This procedure works fine for synchronic corpora. What is missing, however, is either a special mark-up in texts linking each of the items to a certain time or a diachronic lexical database that allows for the matching of the items over time. In what follows, we take the latter approach and present a tool set (MoreXtractor, Morphilizer, MorQuery), a database (Morphilo-DB) and the architecture of a platform (Morphorm) for a sustainable use of diachronic linguistic data for Middle English, Early Modern English and Modern English.

1 Introduction

The sustainability of linguistic resources has gained considerable attention in the last years or so (Dipper et al., 2006; Rehm et al., 2010; Schmidt et al., 2006; Stührenberg et al., 2008). This development was probably initiated and was certainly fostered by federally funded research projects on information structure (SFB 632), linguistic data structures (SFB 441), or multilingualism (SFB 538). Work on sustainable language resources culminated in a row of frameworks, data models and structures, formats and

tools. Also, it continues to prosper in related work (e.g. CLARIN). SPLICR, for example, addresses the issue of normalization of XML-annotated language records, or meta data (Rehm et al., 2010). In fact, the authors discuss the issues of a steadily growing proprietary tag set, the availability, the accessibility and the findability of linguistic resources. More precisely, search engines locate commercially produced data collections, but miss deep structured resources of small research projects. Privacy and property rights restrict accessibility. Proprietary tag sets not obeying to established standards pose a problem for automatic analysis. In this vein, PAULA concentrates on stand-off annotations and the TUS-NELDA repository states an example of integrated annotations. Both frameworks specify methods for handling and storing linguistic data. Finally, EXMERALDA is a tool for annotating spoken data in the first place. In sum, the focus in this field comprise work on annotation, format, tools, data integration (Dipper et al., 2006; Witt et al., 2009) and documentation (Simons and Bird, 2008). In a more general sense, these dimensions reflect Simons' and Bird's (2008) first three key players in a sustainable framework of language resources: creators, archives, aggregators, and users.

Although some of the repositories include historical language resources, the data structures and tools do not take into account the diachronic dimension, that is, language change over large spans of time is not represented in any of the models. Indeed, one finds tools for tagging morphological information, annotation schemas or tran-

scription (Dipper, 2011; Dipper, 2010; Dipper and Schnurrenberger, 2009), but they are not integrated in the very architecture of the present frameworks. We like to initiate a kick-start to close this gap by providing a first sketch of a platform, a tool set and a database that is specifically designed for diachronic data, i.e. adding the time dimension. We will not elaborate on the issues of annotations and formats here. For reasons of ease, the annotation is kept as simple and as minimal as possible so that they can be transferred to an appropriate XML tag set, if available or necessary.

The Morphilo tool set aims at building a representative diachronic database of English. The software consists of three components: MoreXtractor, Morphilizer and MorQuery. MoreXtractor uses a quite simple algorithm that – dependent on the given word class and a rule set – identifies the structure of the word and assigns lexical tags to it (e.g. */root* or */pref*). The identification process is based on enumerated lists comprising all prefix and suffix allomorphs listed in the OED. After inputting a tagged corpus from a specific time, MoreXtractor produces a text file, in which the structure of all words is annotated.

Since the algorithm “overgeneralizes”, the file has to be checked for wrong annotations. This tedious task is carried out by the Morphilizer component. It takes each of the text files and its time specification as an input, displays the word structure in a template and allows the user to make adjustments in a comfortable way by click and drop. Each word item, its structure and its token frequency that were checked manually are written into the Morphilo-Database. For the given time frame of the text, each word type has to be processed only once.

MorQuery provides a comfortable search of the database. Each combination of morphemes, allomorphs, compounds, word types, time frames or corpora can be chosen from drop-down menus. It is also possible to make selections of the most frequent queries or directly type SQL commands to the prompt.

Last, Morphorm is a platform incorporating the tool set and the database. Morphorm will be available to the linguistic community on a website. All researchers are encouraged to query the data, but also to contribute to the project by hav-

```
public enum SuffixEnum {
    ...
    ship("ship"), skiepe("ship"), scipe("ship"),
    scype("ship"), scip("ship"), sciop("ship"),
    scep("ship"), sip("ship"), sipe("ship"),
    schipe("ship"), schupe("ship"), schippe("ship"),
    shipe("ship"), schyp("ship"), schep("ship"),
    shep("ship"), shepe("ship"), chipe("ship"),
    chepe("ship"), schip("ship"), shyp("ship"),
    shippe("ship"), schuppe("ship"), chyp("ship"),
    chep("ship"), shyppe("ship"), shipp("ship")
}
}
```

Figure 1: representation of *ship*-suffix

ing their own diachronic corpora read in and analyzed. Since the database will have a large stock of entries by its inception, the workload for post-processing using Morphilizer for each additional new corpus will be evanescently little.

2 Morphilo Architecture: Toolset – Database – Platform

2.1 Data Structures

Prefix morphemes and suffix morphemes are stored in enumerated lists. Each entry in the list represents one morpheme referring to differing numbers of allomorphs. These allomorphs were extracted from the OED (3rd edition, online version). The OED enlists 179 entries for prefix morphemes and 390 entries for suffix morphemes. The various forms of each suffix – e.g. *mentt*, *mente*, *ment* especially present in Middle English – are referenced in the data structure as allomorphs. In some extreme cases, such as the prefix *over-*, the OED lists over 100 written variants. Other entries, such as the *trans*-prefix, have only one form listed (see figure 1).

There are some cases in which one form represents several morphemes, e.g. there are three entries for the *ant*-suffix. Since these cases are either due to assimilation, misinterpretation (peasan(t), for example) or meaning shift – all of which occur over time – these cases are captured on the time scale in the database (see section 2.2). The enumerated lists represent exactly one form of each affix (morpheme) and all its allographs. Even though there are some cases, in which an affix form corresponds to several meanings at a time (e.g. out Booij 2010: pp 19), this is clearly not the rule, most likely a transitional stadium and subject of an ongoing debate whether the same

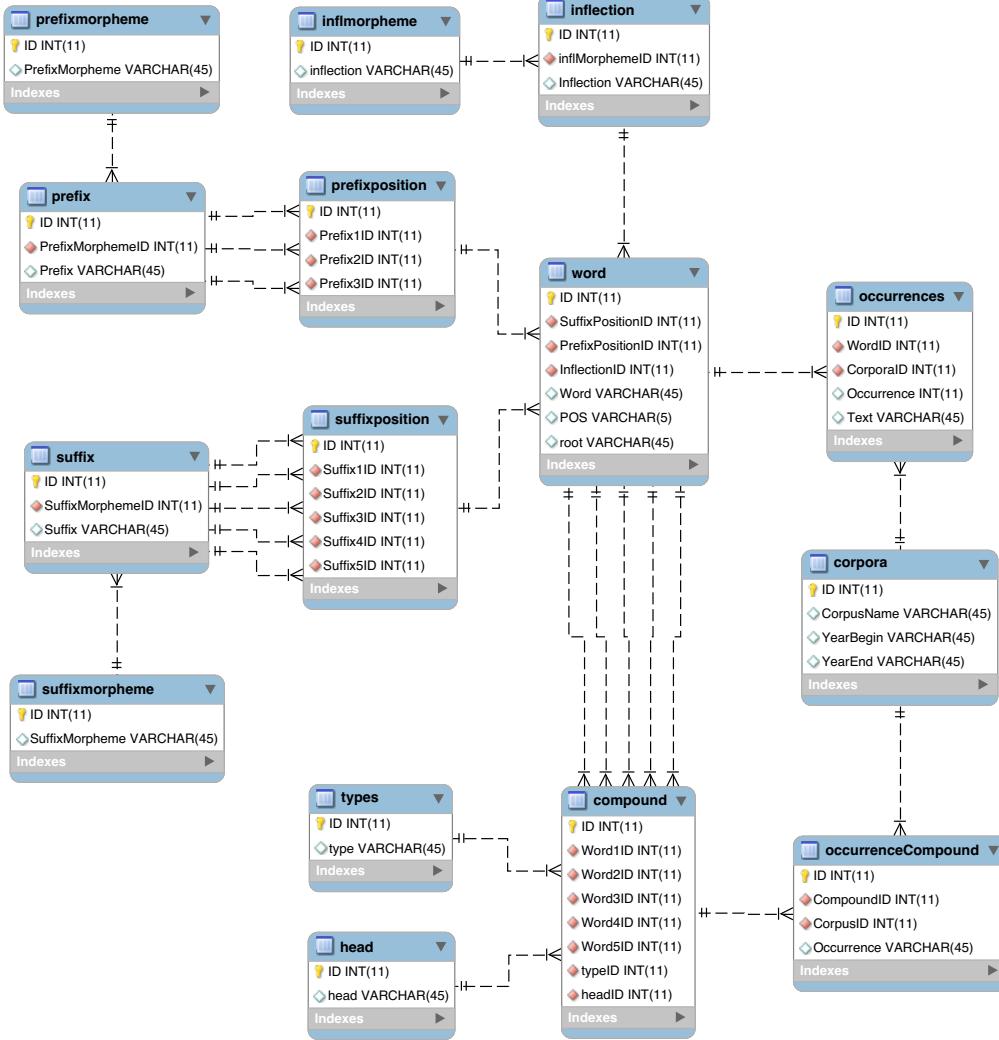


Figure 2: ER-Diagramm of the Morphilo database

meaning is involved. In addition, we find slight semantic differences in the majority of derivational affixes depending on the environment they are attached to. To illustrate, throughout the history of English, *lordship* has incorporated several meanings ranging from “a percentage on sales of books” (mainly used as such in the 19th century) to “arbitrariness” (documented in the 17th century) and “government, province, district” (OED, 2012). From the given example, it is clear that both the root *lord* and the *ship*-suffix embrace different meanings. So affix polysemy is as much a matter of degree as are slight semantic differences provoked by the semantic content of the “carrier word”. In sum, the enumerated lists alone do not include all necessary information, but need

reference to the time information stored in the database. Equal forms referring to different semantic contents are represented at different time periods.

2.2 Morphilo Database

The Morphilo database is a MySQL-database and plays a pivotal role in the design of the application (see figure 2). It holds data on the position and order of derivational and inflectional affixes per predefined time slice (here 70 years). Moreover, compounds are included. They possess information on the position of the head and its type (e.g. exocentric, dvandva).

The basic unit of analysis is the word. In the corresponding table each analyzed word is listed

once per time period. Along with the information of the word form, its root and part of speech are also given. If a word occurs more than once per specified period, the occurrence is incremented. The table *occurrences* is linked to the table *corpora*, which encodes the time information along with the name of the corpus to be analyzed. Time is specified by a beginning date and an end date. These dates are checked before the information of a new corpus can be added.

The *compounds* (figure 2) link to the *corpora*-table as well. However, *compounds* consist of words and hence *compounds* can be derived from the *words*-table. In the *compounds*-table itself, the order of its components (words) is encoded. All words, on the other hand, can be analyzed in terms of their components also, that is, affixes. The order of the affixes can be gained from the respective “position-tables”. For inflectional affixes, no position is specified. We assume for English that inflections occur at the end of the word only once. The tables *prefix* and *suffix* define all allomorphs whereas *prefixmorpheme* and *suffixmorpheme* harbor their morphemic representations.

Thus, simple as well as complex words are represented in a structured format. Each entry (compound or word) has to be analyzed only once per time period. Inconsistencies may arise if two different structures of the same word show up. In this case, one has to agree on one interpretation for the given time period.

2.3 Morphilo Toolset

The Morphilo tool set consists of three components: MoreXtractor, Morphilizer, and MorQuery. MoreXtractor commands a reductionistic logic matching a set of affix strings to the given word input by using a simple rule set of the English Morphology. Since this algorithm is highly overgeneralizing, the Morphilizer assists in correcting the overgeneralizations and storing the correct entries in a database. Last, MorQuery is a tool to conveniently query the database for all common features encountered in English derivational morphology. In short, the Morphilo tools assist in filling and querying the database.

```
...  
judg/root      ment   /N     1/suf  
im pute/root   /VB    1/pref  
de light/root   ful    /ADJ   1/pref 1/suf  
en vir/root    on     ment  s   /N     1/pref 2/suf 1/infl  
fash/root      ion    /N     1/suf  
pro p/root     er     /ADJ   1/pref 1/infl  
...
```

Figure 3: sample extract from a morphilo-tagged file

2.3.1 MoreXtractor

MoreXtractor is a morphological tagger. For the present implementation, the program reads in Penn-Treebank-tagged text corpora and stores them in a vector. The graphical user interface (GUI) offers the option of processing word classes (N, V, A, or Adv). The POS-information is there to allow the user to filter the word classes of interest. Its effect on avoiding affixal ambiguity for internal processing is insignificant.

The software will then run a simple stemmer for the inflectional system of Middle English. The stemmer follows the logic of a 2-subsequential finite state transducer (Mohri, 1997) that aligns the known inflectional endings to the word. The archaic inflectional prefix *y-* is omitted. Likewise, the remnants of the Old English stem-based morphology as well as exceptions (*ox-oxen*, *mouse-mice*, *sheep-sheep*) remain unconsidered. All inflections are marked with */infl* without any further encodings of the English inflectional morphology.

In a second step, each derivational prefix and suffix of the corresponding enumerated lists dependent on the word class is mapped to the stemmed item. Whenever several affixes can be fully mapped (e.g. *-ion* versus *-ation*), the longer item is selected because the probability that the longer affix corresponds to its lexical counterpart is higher (Best, 2003). Prefixes are mapped from left to right; suffixes from right to left. The remnant of the string alignments is tagged as */root*. Last, the updated vector is stored in a text file (see figure 3).

One can clearly see that the transducer overgeneralizes. To be precise, the last entry in figure 3 – *proper* – the inflectional suffix *-er*, which usually specifies the comparative in adjectives, as well as the prefix *pro-*, which is eligible for nouns and adjectives, are indeed marked as affixes although they belong to the root of the monomorphemic word *proper*. In fact, this behavior of the algorithm is intentional because first it prevents us

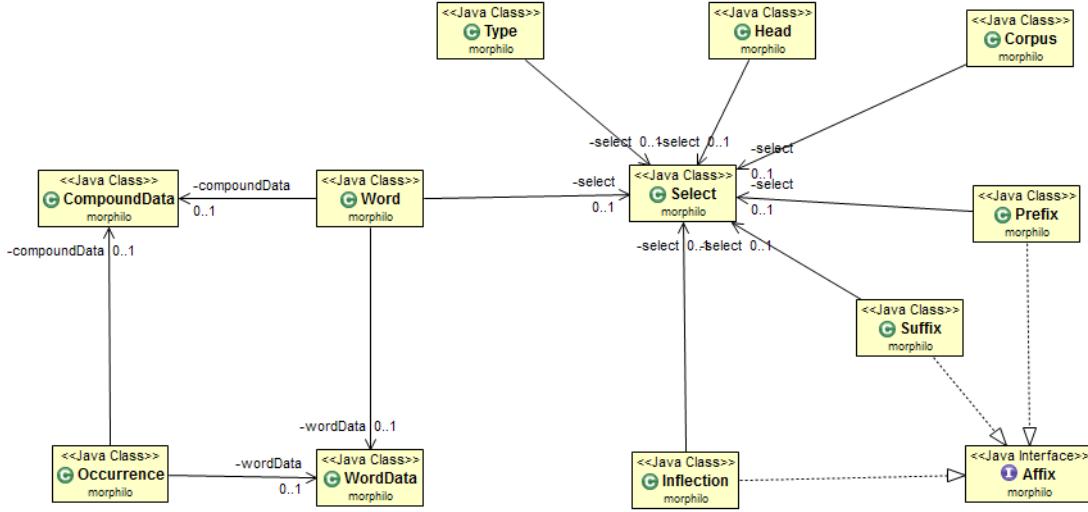


Figure 4: Morphilizer’s implementation as observer pattern

from missing any potential candidates by a manual follow-up analysis and second the algorithm is applicable to other languages more easily for its generality.

2.3.2 Morphilizer

Morphilizer organizes the final analysis by hand. MoreXtractor’s automated tagging procedure outputs a morphemically tagged output file. It is these annotations that will help the user to efficiently correct false affix annotations by click and drop and thus quickly build up a data stock that is then also used in subsequent matching procedures. Morphilizer’s design is based on the observer pattern (see figure 4). The affix interface is implemented by the *Prefix*, *Suffix* and *Inflection* classes, which register at the *Select*-class. The difference to the standard observer pattern is that the registered classes cannot resign from their “Observable” class once they are declared for a certain time period, that is, defined affixes stay the way they are (for more specific information please see the documentation section on www.morphilo.uni-hamburg.de).

Morphilizer takes three input variables: the tagged file, the time range and the corpus name. The algorithm starts by checking against the time range in the *corpora*-table of the Morphilo database (see figure 2). Once the specified dates fall within an existent time range and the corpus name is not yet included, all entries of the tag file

are matched to the *word*-table referring to both its word class (*POS* in table *word*) and its word form (*word* in table *word*). If present, the occurrences of the item in the tagged file are counted, then deleted and the table *occurrences* is updated by incrementing the respective number in the field *Occurrence*. All entries that are not available in the Morphilo database are left unchanged in the file. They will be processed in the same manner as those corpora that fall outside any represented time range. For the latter case, the table *corpora* is updated first by the new corpus information (time range and name). Eventually the manual analysis begins.

Morphilizer presents each entry that is to be analyzed manually in text fields such as “prefix 1”, “prefix 2”, “root”, “suffix 5”, etc. corresponding to the automated analysis done by MoreXtractor. At this point, the user will interfere and either confirm or correct and rearrange the suggestions. Most of the commands in Morphilizer are carried out in this manner. Compound words undergo a slightly different procedure. Some of the Penn-Treebank-tagged corpora do not indicate compounds. Whenever real compounds occur in the word section of the Morphilizer GUI, they can be shifted to the compound section by a mouse click. At the end of the analysis, all instances of the corresponding item are counted and deleted in the original file. Finally, the new word is written into the database and all relevant tables are up-

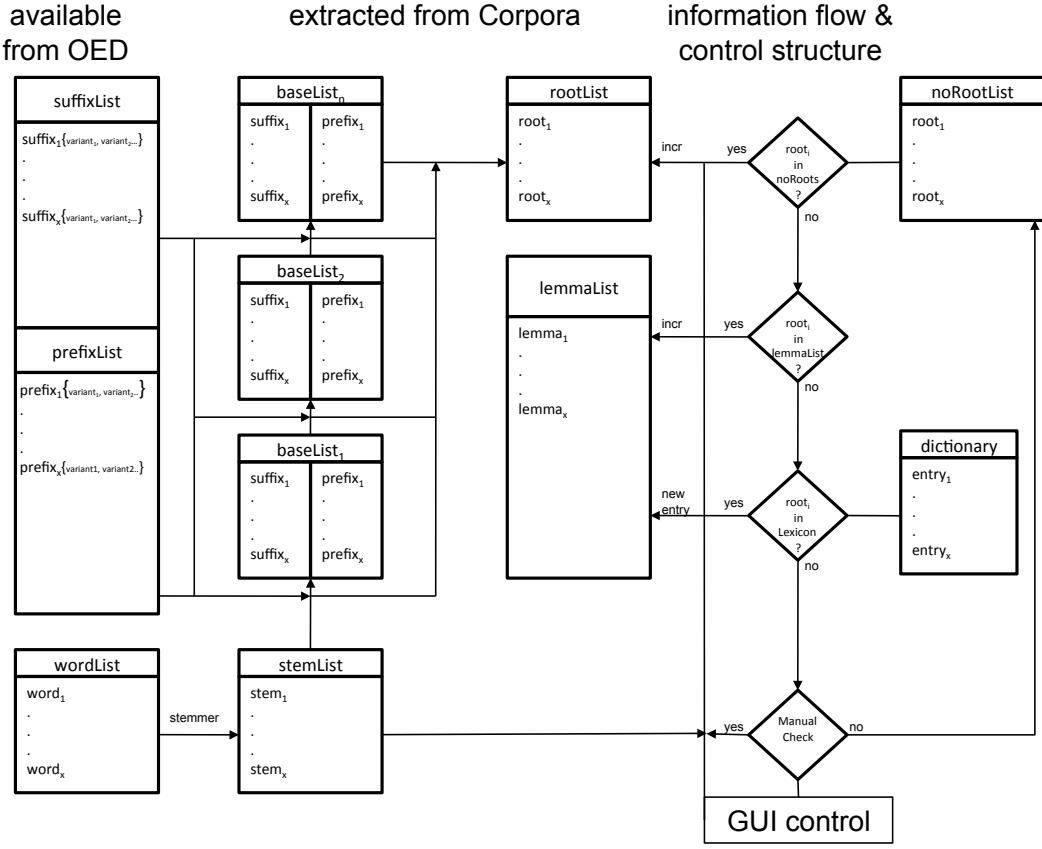


Figure 5: Architecture of Moreextractor and Morphilizer

dated. Deleting the entry from the original file enables the user to interrupt her or his work and go on at a later point in time. As a summary, the main sequences of the algorithm (MoreXtractor and Morphilizer) is visualized in figure 5.

2.3.3 MorQuery

MorQuery is the third component in the tool set. It is an independent program to query the Morphilo database more easily. A web-based interface is also available. In essence, the user makes a selection of the features of interests (corpus, types/tokens, word class, morpheme/allomorph, affix position, prefixes/suffixes/compound/words, derivation/inflection). The software combines these choices to valid SQL commands, queries the database and returns the results as textual output. The results can be saved for further statistical analyses in a tab-delimited format. While for very specific information requests, SQL queries can also be entered directly, a selection of the most

common queries can be chosen from a drop down menu.

2.4 Morphilo Platform: Morphorm

Morphorm is a platform attempting to contribute to a sustainable framework of reusability of diachronic linguistic data. The framework incorporates the Morphilo tool set and the Morphilo database. In addition, it extends the prevalent structure to meet the requirements of a multi-user design. The main idea behind Morphorm is similar to web wikis: share work - receive full profit. Users contribute to the data stock and profit themselves from a more representative set of data and less annotation work. With each additional unit of annotated text, future annotation work will be substantially less for all users since each item (word or compound) has to be analyzed only once.

Figure 6 depicts the architecture. Note that MoreXtractor receives direct input on the time range and words from the database here. This

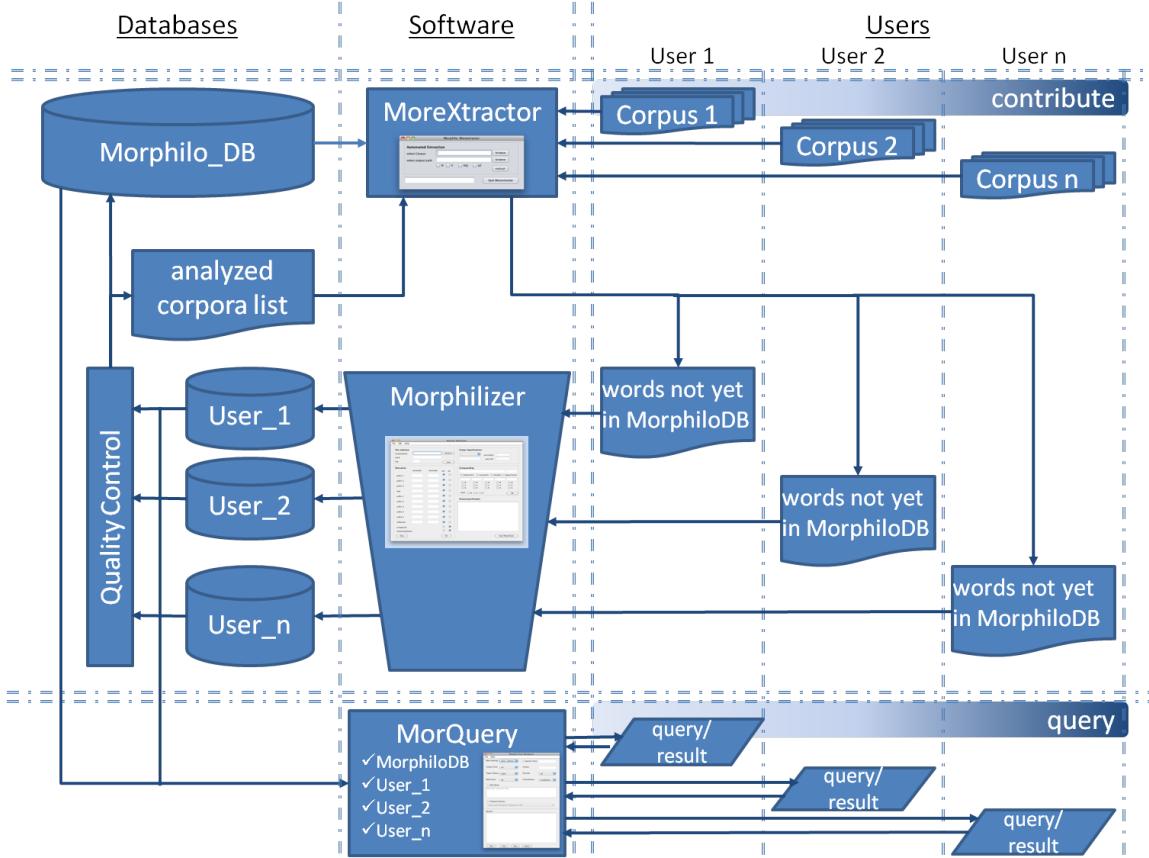


Figure 6: Architecture of Morphorm

feature is part of Morphilizer in the standalone application. Also, a list of analyzed corpora ensures that no data is processed twice. Each new corpus is written to this file. The second difference in Morphorm is that new data is not written into the original database, but to separate datasets that are structurally identical. The third adjustment made in Morphorm is quality control. Decentralizing quality control is a sensitive issue and cannot *per se* be fully automated. There is no full-fledged solution available, but we will use indicators and reported feedback by users. A first indicator is the frequency of usage of a certain dataset by the user community and in publications. A high frequency indicates a certain trust in the analyzed data. A second indicator is, if available, data of the registered user, e.g. his or her project, background, or department. Third, unexpected differences in the result sets of the Morphilo and the user dataset hint at possible erratic annotations. However, from the suspicious datasets a sample will be drawn and will be checked manually. Last,

reported errors from other users will contribute to revising or excluding datasets from accommodating it to the master file. If a “user dataset” meets all quality standards, it is incorporated into the Morphilo database.

The integration of MorQuery made an additional selection field necessary. The user makes the choice on a selection of datasets most suitable to her or him. The quality of the Morphilo database is assured; for all others that have not been checked for quality no guarantees can be made. So, it is up to the user to make a decision on the trade-off between representativeness and risk of wrong annotations. A possible way of dealing with this situation is to make several queries (similar to the procedure described above for quality control): one with the Morphilo dataset, one with all datasets and one with the personal selection of datasets. If the results deviate substantially from the Morphilo results, the selection should be treated with caution. The data should be checked individually and reported to the quality control.

3 Discussion

A first criticism could be addressed to ignoring the XML standard for making morphological annotations and a respective XML-based repository. There are two lines of argumentation to support the present configuration. First, MoreXtractor produces output for Morphilizer. The output is not meant as a tagged text for further external processing. Really, the annotation is added for reasons of user convenience. It is indeed possible to use an XML schema instead, but it does not justify the effort because the database, at least not now, is not represented as an XML repository. This leads to the second line of argumentation, it is still unclear whether XML in its present implementation will be established as a standard for linguistic annotation in general. At present, the “Morphilo data” is available in a structured format. It is unproblematic to transfer MySQL data or object data to XML subsequently if agreement on a standard is reached. Until then, it has advantageous for programming and available design patterns to use the present structure.

In the light of the recent developments of word taggers, a second criticism could be directed towards the simplicity of the algorithm of MoreXtractor. Again, the idea behind MoreXtractor is not to give a reasonable text output for further external processing. More importantly, the software is not tailored for one particular language. Even if the present implementation is for the English Language, the Morphilo framework as such could be implemented in any other language, in which derivational morphology is an important part of the grammar. A simple matching procedure that depends on word class affixation as its only constraint can be implemented for any language. In contrast, from a typological perspective, the idiosyncrasies of language-specific morphology is the most complex. Hence an architecture heavily dependent on language-specific morphology results in a large effort of adjustments.

Finally, the success of the Morphilo crucially depends on the participation of other scientists in the field of the historical derivational morphology of English. Supposedly, the number of these scientists cannot be exceedingly large and so shared annotation work will only pay off over a larger

time frame. In this case, success requires great persistence and obviously it implies data sustainability. In addition, a larger time horizon could pose an issue to quality assurance as well because it entails maintenance and as such man power. We can only speculate on the future acceptance of Morphilo, but once the initial database comprises the bulk of the known vocabulary of Middle English and Early Middle English, only very few new words will continue to be incorporated so that maintenance is then to be restricted to a minimum. At this stage, we will have arrived at a nearly fully automatic affix extraction device for derivations and inflections.

4 Summary

We have presented a tool set that helps to analyze lexical units and organize the work on historical text corpora. These tools can also be used in a web-based platform encouraging a culture of sharing and participation, but also saving time and work. The idea grew out of the need to cooperate more intensively in the field of historical linguistics on the basis of digital texts and media. From some publications in the field (Hiltunen, 1983; Dalton-Puffer, 1996; Haselow, 2011; Ciszek, 2008; Bauer, 2009; Nevalainen, 2008) and personal communication we can see that annotation work of the same corpus material is often carried out several times. In fact, often conflicting evidence is produced because of deviant procedures in the analysis of data.

By initiating a platform and making it known to the research community, not only the workload can indeed be diminished, but also a common standard for analyzing diachronic derivational affixes can be established. At the same time, large and more representative sets of diachronic linguistic data allows us to apply a larger spectrum of quantitative methods. As a consequence, the successful implementation and acceptance contributes without much ado to a sustainable use of historical linguistic data. It is in this spirit that we like to recommend the Morphilo framework to other scientists in the field.

References

- Harald R. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The celex lexical database (cd-rom).
- Laurie Bauer, 2009. *Competition in English Word Formation*, pages 177–198. Wiley-Blackwell, Malden.
- Karl-Heinz Best. 2003. *Quantitative Linguistik*. Peust und Gutschmidt, Göttingen.
- Ewa Ciszek. 2008. *Word derivation in Early Middle English*, volume 23 of *Studies in English medieval language and literature*. Peter Lang, Frankfurt/Main.
- Christiane Dalton-Puffer. 1996. *The French Influence on Middle English Morphology: A Corpus-Based Study of Derivation*. Mouton De Gruyter, New York.
- Stefanie Dipper and Martin Schnurrenberger. 2009. Otto: A tool for diplomatic transcription of historical texts. In *4th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 516–520.
- Stefanie Dipper, Erhard Hinrichs, Thomas Schmidt, Andreas Wagner, and Andreas Witt. 2006. Sustainability of linguistic ressources. In Erhard Hinrichs, Nancy Ide, Martha Palmer, and James Pustejovsky, editors, *LREC 2006 Workshop on Merging and Layering Linguistic Information*, pages 48–54.
- Stefanie Dipper. 2010. Pos-tagging of historical language data: First experiments. In *10th Conference on Natural Language Processing (KONVENS-10)*, Semantic Approaches in Natural Language Processing, pages 117–121.
- Stefanie Dipper. 2011. Morphological and part-of-speech tagging of historical language data: Comparison. *Journal for Language Technology and Computational Linguistics*, 26(2):25–37.
- Alexander Haselow. 2011. *Typological Changes in the Lexicon: Analytic Tendencies in English Noun Formation*, volume 72 of *Topics in English Linguistics*. De Gruyter Mouton, Berlin.
- Risto Hiltunen. 1983. *The decline of the prefixes and the beginnings of the English phrasal verb: the evidence from some Old and Early Middle English texts*, volume 160 of *Turun Yliopiston julkaisuja*. Turun Yliopisto, Turku.
- Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–312.
- Terttu Nevalainen, 2008. *Early Modern English (1485-1660)*, pages 209–215. Wiley Blackwell, Malden.
- Oxford English Dictionary OED. 2012. "lordship, n.", oxford university press, 3rd edition, online version: <http://www.oed.com/view/entry/110327>.
- Georg Rehm, Oliver Schonefeld, Thorsten Trippel, and Andreas Witt. 2010. Sustainability of linguistic ressources revisited. In *International Symposium on XML for the long Haul: Issues in the Long-term Preservation of XML*, volume 6 of *Balisse Series on Markup Technologies*.
- Thomas Schmidt, Christian Chiarcos, Timm Lehmburg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic ressources.
- Gary F. Simons and Steven Bird. 2008. Toward a global infrastructure for the sustainability of language ressources. In *22nd Pacific Asia Conference on Language, Information and Computation*.
- Maik Stührenberg, Michael Beißwenger, Kai-Uwe Kühnberger, Harald Lünen, Alexander Mehler, Dieter Metzing, and Uwe Mönnich. 2008. Sustainability of text-technological ressources. In *Post-LREC-2008 Workshop on Sustainability of Language Ressources and Tools for Natural Language Processing*.
- Andreas Witt, Georg Rehm, Erhard Hinrichs, Timm Lehmburg, and Jens Stegmann. 2009. Sustainability of linguistic resources through feature structures. *Literary and Linguistic Computing*, 24(3):363–372.

The Reference Corpus of Late Middle English Scientific Prose

Javier Calle-Martín

Universidad de Málaga, Dpto. Filología Inglesa, Francesa y Alemana, Campus de Teatinos s/n, 29071 Málaga, Spain
jcalle@uma.es

Teresa Marqués-Aguado

Universidad de Murcia, Dpto. Filología Inglesa, Campus de la Merced, 30071 Murcia, Spain
tmarques@um.es

Laura Esteban-Segura

Universidad de Murcia, Dpto. Filología Inglesa, Campus de la Merced, 30071 Murcia, Spain
lesteban@um.es

Antonio Miranda-García

Universidad de Málaga, Dpto. Filología Inglesa, Francesa y Alemana, Campus de Teatinos s/n, 29071 Málaga, Spain
amiranda@uma.es

Abstract

This paper presents the current status of the project *Reference Corpus of Late Middle English Scientific Prose*, which pursues the digital editing of hitherto unedited scientific, particularly medical, manuscripts in late Middle English, as well as the compilation of an annotated corpus. The principles followed for the digital editions and the compilation of the corpus will be explained; the development and application of several specific tools to retrieve linguistic information within the framework of the project will also be discussed. Our work joins in with worldwide initiatives from other research teams devoted to the study of medical and scientific writings in the history of English (see Taavitsainen, 2009).

1 Introduction

Digital editing has been much debated for more than a decade since the advent of the first projects in English like *The Canterbury Tales* (started in 1993) and *The Electronic Beowulf* (in 1994). The active scholarly thinking is corroborated not only by the publication of a plethora of *ad hoc* monographs discussing the nature of digital editions from different perspectives (Sutherland, 1997; Burnard, O'Brien O'Keefe and Unsworth, 2006; Deegan and Sutherland, 2009), but also by the special-themed issue published by *Literary and Linguistic Computing* (2009), approaching the topic from theoretical and empirical domains.

It is now a common practice among many manuscript holders to digitise and to publicise previously unpublished texts and/or manuscripts, offering not only an edition in itself but also the foundations for further work. The benefits of this activity are manifold to such extent that “digital editions of manuscripts [...] have opened new possibilities to scholarship as they normally include fully searchable and browsable transcriptions and, in many cases, some kind of digital facsimile of the original source documents, variously connected to the edited text” (Pierazzo, 2009: 169; also Ore, 2009: 114).

In the light of this trend, the present paper describes the model of electronic editions followed in the *Reference Corpus of Late Middle English Scientific Prose*, an on-going research project developed at the universities of Málaga and Murcia with the following objectives: (a) the implementation of on-line electronic editions of hitherto unedited *Fachprosa* written in the vernacular in which the manuscript high-resolution images are accompanied by their diplomatic transcriptions; and (b) the compilation of an annotated corpus from this material facilitating Boolean and non-Boolean searches, both word- and lemma-based.

The justification for our work lies in the need for faithful transcriptions for research purposes, avoiding the use of published editions. Lavagnino's definition of electronic/digital

edition has been adopted. This is described as an archive offering “diplomatic transcriptions of documents, and facsimiles of those documents. And it should avoid [...] the creation of critically edited texts by means of editorial emendation [...] What readers need is access to original sources” (1998: 149). The rationale behind the project is then to offer faithful reproductions of the originals which can be eventually used as primary sources for research in Linguistics and other side areas like Palaeography, Codicology, Ecdotics or History of Science. This material also serves as the input for the compilation of a corpus of late Middle English scientific prose, thus allowing for the synchronic study of the language from different perspectives, such as phono-orthographic and morpho-syntactic. The internal coherence of the project derives from the contribution of two variables, one qualitative and the other quantitative. On the one hand, the project exclusively focuses on 14th- and 15th-century scientific treatises and, on the other, it displays complete texts, not samples.

The present paper addresses the concept of electronic editing as applied to the *Reference Corpus of Late Middle English Scientific Prose* in order to (a) describe the editorial principles and the theoretical implications adopted; (b) present the digital layout along with the tools implemented for information retrieval; and (c) evaluate our proposal for linguistic research.

2 The collection

The manuscripts have been primarily taken from two holders, the Hunterian Collection at Glasgow University Library, and the Wellcome Library. These two repositories have been chosen on account of (a) the number of scientific treatises from the 14th and the 15th centuries available; (b) their unedited status; and (c) their availability because they have provided us with digitised images of the manuscripts as well as with permission for on-line publication.¹ In its present form, the following items have been transcribed amounting to 471,143 running

¹The British Library and the Bodleian Library witnesses are offered without the digitised images as a result of their expensive copyright prices. The Ryland manuscript, in turn, presents the digitised images freely offered by the holder (<http://www.library.manchester.ac.uk/>).

words, which have also been annotated. The treatises are listed below:²

- Glasgow, University Library, MS Hunter 307, *System of Physic* (ff. 1r-166v), including an anonymous Middle English treatise on humours, elements, uroscopy, complexions, etc. (ff. 1r-13r); the *Middle English Gilbertus Anglicus* (ff. 13r-145v); an anonymous Middle English treatise on buboes (ff. 145v-146v); a gynaecological and obstetrical text (ff. 149v-165v); a Middle English version of Guy de Chauliac's *On bloodletting* (ff. 165v-166v); and an *Alphabetical List of Drugs with their Properties* (ff. 167r-172v).
- Glasgow, University Library, MS Hunter 328 including Gilles of Corbeil's *Treatise on Urines* (ff. 1r-44v); an *Alphabetical List of Remedies* (ff. 45r-62r); and *An Alphabetical List of Medicines* (ff. 62v-68v).
- Glasgow, University Library, MS Hunter 497, translation of Macer's *Herbary* (ff. 1r-92r).
- Glasgow, University Library, MS Hunter 509, *System of Physic* (ff. 1r-167v), including an anonymous Middle English treatise on humours, elements, uroscopy, complexions, etc. (ff. 1r-14r); and the *Middle English Gilbertus Anglicus* (ff. 14r-167v).
- London, British Library, MS Sloane 404, *Medicine: A General Pharmacopoeia* (ff. 3v-319v).
- London, British Library, MS Sloane 2463, *Antidotary* (ff. 154r-193v).
- London, Wellcome Library, MS Wellcome 397, *A Treatise of Powders, Pills, Electuaries and Plasters* (ff. 52v-68v).
- London, Wellcome Library, MS Wellcome 404, *Leechbook* (ff. 1r-44v).
- London, Wellcome Library, MS Wellcome 542, *Leechbook* (ff. 1r-20v).
- London, Wellcome Library, MS Wellcome 799, William de Congenis's *Chirurgia* (ff. 1r-23v).

²The list relies on the data and collation provided by Young and Aitken (1908), Moorat (1962), and Cross (2004), among others.

- London, Wellcome Library, MS Wellcome 5262, *Medical Recipe Collection* (ff. 3v-61v).
- London, Wellcome Library, MS Wellcome 8004, *Physician's Handbook* (ff. 113r-133v).
- Manchester University Library, MS Rylands 1310, Gilles of Corbeil's *Treatise on Urines* (ff. 1r-21r).
- Oxford, Bodleian Library, MS Rawlinson C. 81, *The Dome of Urines* (ff. 6r-12v).

The catalogue is being currently enlarged with the addition of the following treatises:

- London, British Library, MS Sloane 340, Henry Daniel's *Liber uricrisiarum* (ff. 39r-62v).
- London, Wellcome Library, MS Wellcome 226, Henry Daniel's *Liber uricrisiarum* (ff. 1r-70v).
- London, Wellcome Library, MS Wellcome 537, *The Practice on the Sight of Urines* (ff. 15r-46v).
- London, Wellcome Library, MS Wellcome 8004, *Bloodletting* (ff. 18r-32v); *Celestial Distances* (ff. 49r-96v).
- Oxford, Bodleian Library, MS Add. A.106 (ff. 244r-258v).

3 The platform

The *Reference Corpus of Late Middle English Scientific Prose* is hosted at <http://referencecorpus.uma.es>.³ The homepage contains three main items: a table (on the left), a door with the sign "Reading room" on it (on the right) and flags (on the right, above the door). The flags provide links to the websites of the institutions taking part in the project, each one represented by its emblem in the following order: University of Málaga, Junta de Andalucía (Autonomous Government of Andalusia) and University of Murcia.

On the table there are several objects: a picture, which supplies information about the members of the research team; documents, which include the transcription policy; several letters, which give access to the tool "Words & Phrases" (see 4.1); an envelope with a link to

³ The site is best viewed using Firefox (or a compatible web browser) and JavaScript. It requires Flash Player 10, with a resolution of 1280x1024 for best visualisation.

contact details; a diary with information about the project ("Description", "Copyright", "Acknowledgements", and "Sitemap"); and a globe, through which the "Guided tour"⁴ can be reached.

By opening the door the user enters the "Reading room", which holds the manuscripts and provides access to their digitised images, transcriptions, and brief physical descriptions. Five volumes, with the words "Hunter", "Rawlinson", "Rylands", "Sloane" and "Wellcome" (from left to right), are showcased on a table. By clicking on the spine of the desired collection,⁵ all the manuscripts belonging to that particular collection appear.⁶

Special characters for symbols such as runes and punctuation marks are employed, and so users need to have a Unicode compliant font in their computers (a link from which it can be freely downloaded is supplied).

As far as linguistic analysis is concerned, three different tools for the retrieval of linguistic information are available, namely *Word Search Tool*, *Text Search Engine (TexSEN)* and *Concordance Manager*. The first of them (see 4.1) allows creating word and lemma lists. *TexSEN* has been especially developed for the extraction of morpho-syntactic and statistical information from annotated corpora (see 4.2). The *Concordance Manager* (see 4.3), in turn, serves as an aid to view the concordances generated by *TexSEN*.

3.1 Editorial principles

More often than not, modern editions are guided by the editorial principles of publishers, which may deny the reader an immediate access to the source text as it was copied by the mediaeval writer. Aspects such as abbreviations,

⁴This supplies information about how to interact best with the icons, animations and visual elements displayed in the site, including the following points: (i) "Interaction", (ii) "Pop-up messages", (iii) "The cursor" and (iv) "Help messages".

⁵It must be borne in mind that the images of the spines arrayed do not correspond to the spines of any real manuscript.

⁶Only MS Hunter 328 (ff. 1r-68v), MS Wellcome 404 (ff. 1r-44v) and MS Wellcome 542 (ff. 1r-21r) are open for public consultation. The other manuscripts are freely available after registration, a required process to control the use and integrity of the resources offered.

marginalia, emendations or punctuation are usually a *desideratum* in modern editions, leaving aside physical details such as lemmata, decoration, illuminated capitals, etc. As the main goal here is the compilation of an annotated corpus of late Middle English scientific material in the vernacular, the principles of a semi-diplomatic edition have been adopted so as to provide an accurate reproduction of the source text. Our transcription follows the guidelines presented next, partially adapted from Petti (1977: 24-35):

- a) The spelling, capitalisation and word division of the original have been retained, including the use of <u> for <v> and <y> for <i>. The different spellings of a same consonant, however, have been regularised, particularly in the case of the letter <s>, which may be represented by <þ> and <ſ>, among others, since the occurrence or choice of separate graphs depends on the position of that letter within the word.⁷
- b) The punctuation and paragraphing of the original have been retained. Marks of punctuation have been represented by the symbols they stand for, e.g. the paragraph mark (¶), the virgule (/) and the caret (^).
- c) Lineation has been preserved and, for reference purposes, the lines have been numbered accordingly (every five lines).
- d) Abbreviations have been expanded with the supplied letter(s) italicised.
- e) Deletions are retained preserving the scribal practice.
- f) Insertions have been included in the body of the text, enclosed in square brackets. The caret mark, if used, has been placed immediately in front of the first bracket.
- g) Catchwords are given at the bottom of the page.

3.2 The electronic edition

The electronic editions of the manuscripts can be freely consulted once the user has registered. Each edition consists of the digitised images of

⁷A graphic transcription has been discarded on account of the research interests which lie behind the edition itself, as we pursue the compilation of an annotated corpus. A graphemic transcription has been adopted for linguistic reasons, as already noted by Robinson and Solopova (1993: 24–25), and Robinson (2009: 45).

folios of a given manuscript and additional features, including its transcription. This fully adheres to the principles of the semi-diplomatic editorial method, in which intervention is kept to a minimum (see 3.1). The only exception is the inclusion of the number of folio and lines, meant as a help to locate information.

The design recreates an environment that brings visitors as close as possible to the experience of consulting the original manuscripts without the need to move from their computers and with other added assets. In order to consult a specific manuscript, the volume that represents the collection housing it should be selected first. When this is done, a pop-up window, in which all the manuscripts available for that collection, together with their different treatises, are listed in alphabetical order, appears. Two icons appear next to the name of the treatise and the range of folios/pages in which it is held: one which shows “image & text” and one which displays information about the treatise selected. Concerning the latter, another pop-up window either with data extracted from the catalogues in which the manuscripts have been described (see footnote 2) or providing a link to the online description of the library catalogue, emerges. For consulting the desired treatise, two possibilities are given: either the images on their own or the images and their transcriptions. After loading, the original cover of the manuscript comes into view. The pages or folios can be browsed either manually or automatically and zoomed for a more exhaustive view by means of a magnifying glass (the icon is on the top-left side of the screen). A particular page or folio can be searched for, a helpful and time-saving option with long treatises. Those visitors unfamiliar with mediaeval scripts can read the transcribed text, which is displayed in a pop-up window next to the image –on the right-hand side in the case of versos and on the left-hand side in the case of rectos (see figure 1). In order to do so, the user only needs to click on the feather icon on the top-right side of the screen.

It is possible to go back to the “Reading room” at any time, just by clicking on the icon “Return to Library”, on the bottom-right side of the screen.

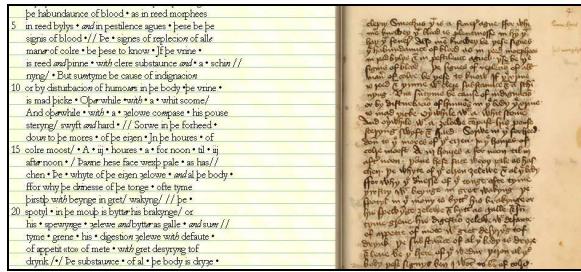


Figure 1. Digitised image of folio recto with transcription (f. 4r, MS Hunter 509)

3.3 Corpus annotation

Lemmatisation and morphological tagging are the two main stages followed in the linguistic annotation of our corpus, which currently comprises around 1,200,000 tokens. Once the manuscripts have been transcribed, the texts are pasted into spreadsheets and every individual item is assigned a row. Each word form occurring in the corpus is then paired with a corresponding base form or lemma. Thus, inflected forms of a word are identified as instances of the same lemma. Lemmatisation also helps to overcome the difficulties posed by orthographic variation, a salient characteristic of Middle English. For the selection of the lemmas, the main headword found in the online version of the *Middle English Dictionary* (*MED*) has been taken as the reference because it provides a standard form that can be used for all the texts. The task of lemmatisation is not exempt from complications, which include, for example, choosing an adequate lemma when a particular lexical item is not gathered in the *MED* (mainly Latin terminology) or deciding whether combinations of words should form part of the same or different lemmas.

Morphological tagging consists of tags including information about words, such as part of speech (noun, adjective, verb, etc.), tense, number, case, gender, etc., as shown in Figure 2. The texts have also been labelled so that each item includes reference to the folio and range of five lines in which it occurs.⁸

One of the most important advantages of such annotated corpora is that they allow extracting linguistic information easily. They

can also be used as input to be exploited by machines or computer-driven tools, which can provide the researcher with detailed analyses. Our corpus consists of full texts which have been transcribed following the original manuscripts closely and faithfully, and therefore constitutes a reliable resource for the study of Middle English (for instance, of aspects dealing with morphology, dialectology and punctuation, etc.).

The main disadvantage is that annotating is a time-consuming and taxing process, since the information in the tags for each particular item has to be manually recorded. This leads to another weakness: errors, which may compromise the accuracy of the results, can be made. For this reason, the annotated corpus has undergone several phases of revision.

4 Information retrieval

Various types of information may be retrieved from a corpus complying with the features explained in section 3. For the purpose, the tools offered in the platform may be used, namely *Word Search Tool*, *TexSEN* and *Concordance Manager*.

4.1 Word Search Tool

By clicking on the “Words and Phrases” icon, users access the *Word Search Tool*, which allows them to obtain a list of the variant spelling forms of the selected text, along with the number of occurrences and the reference where each of them can be found. It is even possible to perform a word- or a lemma-based search.

⁸Further specifications concerning the method of annotation are discussed in Moreno-Olalla and Miranda-García (2009), and Esteban-Segura and Marqués-Aguado (forthcoming).

	Word	Lemma	Class	Subclass	Type	Tense	Number	Person	Case	Gender	Folio	Line
1	Als[o]	alsō, b	Adve	Affirm							149	1
2	.	Pmark	Pmark								149	1
3	we	wē, r	Pron	Pers			Plur	1st	Nom		149	1
4	schulen	shulen, v (1)	Verb		Pret-P	PrsInd	Plur	1st			149	1
5	vndirsonde	understönden, v	Verb			Infin					149	1
6	þat	that, c	Conj	Compl							149	1
7	wymmen	wömmman, n	Noun				Plur				149	1
8	han	häven, v	Verb			PrsInd	Plur	3rd			149	1
9	lesse	lēs(se, a	Adje			Compa					149	1
10	hete	hēte, n (1)	Noun				Sing				149	1
11	in	in, p	Prep						RegDat		149	1
12	her	hēr(e, d	Dete	Poss			Plur	3rd		Fem	149	1
13	body	bōði, n	Noun				Sing				149	1
14	þan	than, c	Conj	Compa							149	1
15	men	man, n	Noun				Plur				149	1

Figure 2. Sample screenshot of morphological tagging (f. 149v, MS Hunter 307)

The first tab requests the selection of the text/treatise to be analysed. If the user is logged in, all the texts are displayed for analysis; otherwise, four texts are shown as samples. After choosing the text in the first tab, two options are possible, i.e. a list of words or a list of lemmas may be provided. For the purpose, the user should click on the preferred option under ‘Show word / lemma’, the default setting being ‘Words’. The second tab then presents all the units of analysis of the chosen text in alphabetical order, i.e. either all its words (taking into account spelling variation) or else the lemmas, which are in turn retrieved from the database including all the entries for all the texts. Since the word-class is shown alongside the lemma, the variant forms linked to each lemma can be distinguished according to their word-class (e.g. *either* is tagged as a pronoun and as a conjunction in the corpus).

The results are automatically shown as a KWIC (Key Word In Context) index with 6 lexical units preceding and following the unit under scrutiny, plus the reference. If there is more than one occurrence, the full list (arranged in order of appearance) is shown and, by clicking on the reference, the user may view the manuscript folio/page where that occurrence is found.

4.2. Text Search Engine

TexSEN, which is available at <http://texsen.uma.es> and linked to the research project presented in this paper, may perform both simple and complex (i.e. Boolean) searches using annotated corpora complying with its

requisites, together with several statistical calculations (Miranda-García and Garrido-Garrido, 2012).

Simple or non-Boolean searches comprise lists and indexes which, if the user is not logged in, will refer either to MS Hunter 503 or else to the ophthalmologic treatise or the antidotary housed in MS Hunter 513, all of which are used as samples.⁹ Once the text under analysis has been picked out from the list under ‘Bookshelf’, the requirements of the search have to be set by clicking on the corresponding tabs: ‘Selection’ refers to the page or folio range; ‘Item’ presents the units of analysis available (words or lemmas); ‘Output’ offers three possible types of lists (either a complete list with all the items; or else a reduced list with only the different items, allowing also for the addition of the number of hits); and, finally, ‘Save as/open’, which enables the user to select the filename for the list produced. Within a few seconds, the user is presented with the results in order of appearance according to the search criteria. The default configuration shows 50 results per screen, although the tab ‘Output’ allows changing this figure, along with the field used to sort the results (sequential order or alphabetical order of the unit of analysis, in ascending or descending order).

⁹This manuscript was used as a sample for the types of analyses that could be carried out under the previous project, *Desarrollo del corpus electrónico de manuscritos medievales ingleses de índole científica basado en la colección Hunteriana de la Universidad de Glasgow* (project reference FFI2008-02336/FILO), and has been maintained as such in this application.

Boolean searches include the options to generate complex searches, lemma-sorted KWIC concordances and glossaries.¹⁰ Complex searches work following the same procedure as simple searches, since the text has to be first selected from the list under ‘Bookshelf’ and then the choices regarding ‘Selection’, ‘Item’, ‘Output’ and ‘Save as/open’ must be specified. Additionally, the ‘Search parameters’ have to be set with the aim of determining the type of unit upon which the search is going to be performed. First, the type of unit under analysis or accidence (word, lemma, word-class, number...) must be picked from the ‘Column’ section. Then, the Boolean operator must be added in the ‘Condition’ tab, followed by the values for that unit/accidence under ‘Value’ (e.g. singular, plural, both and dual are the options provided). Another possibility is using several values together, hence increasing the potentialities of the complex search. The parameters have to be validated by clicking on the box icon, which changes from red to green when the former are suitably arranged, and then needs to be dragged to the suitable column in the KWIC arrangement, i.e. from six words preceding to six words following the keyword, the latter of which can also be specified, as shown in Figure 3. If the user is interested in typing in a particular word or lemma to carry out the search, a window with Unicode symbols is shown on the right, from which letterforms can be added by a simple click.

The tool lemma-sorted KWIC concordance builder replicates the same structure (‘Bookshelf’ to choose the text under scrutiny, and tabs to set the requirements concerning ‘Selection’, ‘Output’ and ‘Save as/open’). In the ‘Output’ tab users may choose the span of words preceding and following the keyword, which can range from 1 to 20 each, as well as the word types for analysis (nouns, verbs, adjectives, adverbs, all function words, all content words or all items). The results are shown on the screen using the default configuration, that is, showing the concordances arranged by lemmas, whose spelling forms (and therefore, the corresponding lines of concordance) are presented in order of appearance. Yet, this presentation may be modified by choosing a different parameter to

arrange all the results, or by altering the ascending/descending icons in some/all of the columns of the results page.

4.3. Concordance Manager

The *Concordance Manager* is an online application that allows viewing the concordances generated by *TexSEn*. As with the previous tool, if the user is registered, the concordances to a wider range of texts are offered (those in the Hunter manuscripts only); otherwise, only MS Hunter 503 is available for consultation as a sample.

In order to access the concordances of a particular lemma, the manuscript on which the search is to be performed first needs to be picked from the list provided, along with the lemma, selected in turn from a predictive list including all the lemmas for the text chosen.

The results screen displays the list of lines of concordances (i.e. rows) in tabular format, in which five words typically precede and other five follow the keyword, which is highlighted in red. Hence, the information rendered is structured into: a) lemma; b) preceding context; c) keyword; d) following context; and e) reference. The order in which the concordances are presented follows that of occurrence in the text (page or folio number, side and line span), as signalled in the column for the reference, rather than alphabetically. The default number of lines of concordance shown per page is 200, although this figure may be modified by the user. Likewise, the results can be re-arranged according to the reference (from end to beginning), to the word preceding or following the keyword, and to the keyword itself (thus grouping all the examples of particular spelling variants together, for instance).

¹⁰The tool Glossary builder is not available at the moment.

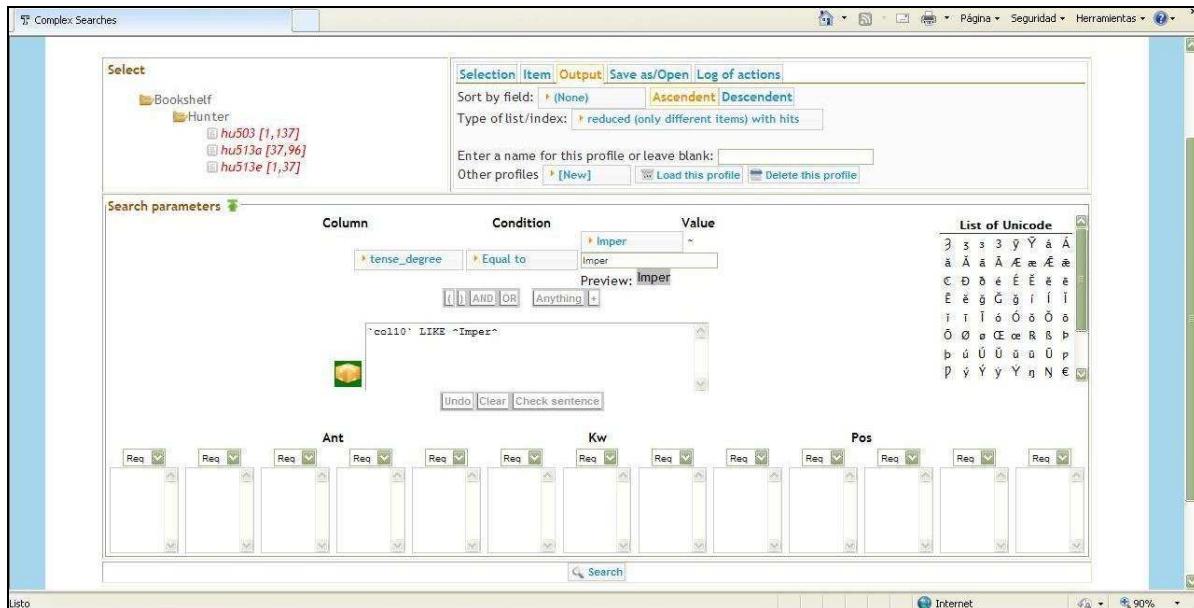


Figure 3. Sample screenshot of complex searches with *TexSEN*

Beyond purely statistical data, concordances lend themselves well to other types of studies. For instance, they are very helpful for the analysis of allomorphs of a particular inflection or morpheme, especially if the latter depend on the ensuing context. Likewise, comparative studies are possible if the same search is carried out on various texts with a view to analysing whether certain lemmas are common to all the texts under scrutiny, or if they are only used in some texts and alternatives are sought for other texts. These results may even suggest some kind of authorial fingerprint.

Conclusions

This paper has presented the highlights of the project *Reference Corpus of Late Middle English Scientific Prose*, discussing the motivations for electronic editing and the principles followed. It has also shown the potential of lemmatised and annotated corpora, especially when they are used in conjunction with specific-purpose tools. The ones presented in this paper are particularly aimed at extracting relevant data from annotated corpora in Middle English, hence allowing for a wide variety of quantitative studies on the language of the period. Furthermore, by being available online, most results can be made public easily. Future research will focus on the expansion of the corpus so as to include manuscripts from the

Early Modern period, which will eventually allow for more comprehensive diachronic studies.

Acknowledgments

The present research has been funded by the Spanish Ministry of Science and Innovation (grant number FFI2011-26492) and by the Autonomous Government of Andalusia (grant numbers P07-HUM2609 and P11-HUM7597). These grants are hereby gratefully acknowledged.

References

- Lou Burnard, Katherine O'Brien O'Keefe, and John Unsworth (eds.). 2006. *Electronic Textual Editing*. New York, Modern Language Association of America.
- Rowin Cross. 2004. *A Handlist of Manuscripts Containing English in the Hunterian Collection, Glasgow University Library*. Glasgow, Glasgow University Library.
- Marilyn Deegan and Kathryn Sutherland (eds.). 2009. *Text Editing, Print and the Digital World*. Surrey, Ashgate.
- Laura Esteban-Segura and Teresa Marqués-Aguado. Forthcoming. New Software Tools for the Analysis of Computerized Historical Corpora: GUL MSS Hunter 509 and 513 in the Light of *TexSEN*. In Vincent Gillespie and Anne Hudson

- (eds.). *Editing Medieval Texts from Britain in the Twenty-First Century*. Turnhout, Brepols.
- John Lavagnino. 1998. Electronic Editions and the Needs of Readers. In W. Speed Hill (ed.). *New Ways of Looking at Old Texts II*. Tempe, AZ, Renaissance English Text Society Publications:149–156.
- Robert E. Lewis *et al.* (eds.). 1952–2001. *Middle English Dictionary*. Ann Arbor: University of Michigan Press. Online version in Frances McSparran *et al.* (eds.). 2001–. *Middle English Compendium*. University of Michigan Digital Library Production Service. Available at <http://quod.lib.umich.edu/m/med>.
- Antonio Miranda-García and Joaquín Garrido-Garrido. 2012. *Text Search Engine (TexSEN)*. Málaga, Servicio de Publicaciones de la Universidad de Málaga. Available at <http://texsen.uma.es>.
- Samuel Arthur Joseph Moorat. 1962. *Catalogue of Western Manuscripts on Medicine and Science in the Wellcome Historical Medical Library. Vol. 1. Catalogues Written before 1650 AD*. London, Publications of the Wellcome Historical Medical Library.
- David Moreno-Olalla and Antonio Miranda-García. 2009. An Annotated Corpus of Middle English Scientific Prose: Aims and Features. In Javier E. Díaz Vera and Rosario Caballero (eds.). *Textual Healing: Studies in Medieval English Medical, Scientific and Technical Texts*. Bern, Berlin, etc., Peter Lang:123–140.
- Espen S. Ore. 2009. ... They Hid Their Books Underground. In Marilyn Deegan and Kathryn Sutherland (eds.). *Text Editing, Printing and the Digital World*. Surrey, Ashgate:113–125.
- Anthony G. Petti. 1977. *English Literary Hands from Chaucer to Dryden*. Cambridge, Mass., Harvard University Press.
- Elena Pierazzo. 2009. Digital Genetic Editions: the Encoding of Time in Manuscript Transcription. In Marilyn Deegan and Kathryn Sutherland (eds.). *Text Editing, Printing and the Digital World*. Surrey, Ashgate:169–186.
- Peter Robinson. 2009. What Text Really is Not, and Why Editors have to Learn to Swim. *Literary and Linguistic Computing*, 24(1):41–52.
- Peter Robinson and Elizabeth Solopova. 1993. Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue. In Norman Blake and Peter Robinson (eds.). *The Canterbury Tales Project. Occasional Papers Volume I*:19–52. Available at <http://www.canterburytalesproject.org/pubs/op1-transguide.pdf>.
- Kathryn Sutherland. 1997. *Electronic Text: Investigations in Method and Theory*. Oxford, Clarendon Press.
- Irma Taavitsainen. 2009. Early English Scientific Writing: New Corpora, New Approaches. In Javier E. Díaz Vera and Rosario Caballero (eds.). *Textual Healing: Studies in Medieval English Medical, Scientific and Technical Texts*. Bern, Berlin, etc., Peter Lang: 177–206.
- John Young and P. Henderson Aitken. 1908. *A Catalogue of the Manuscripts in the Library of the Hunterian Museum in the University of Glasgow*. Glasgow, Maclehose.

KONVENS 2012

**LexSem 2012: Workshop on Recent Developments
and Applications of Lexical-Semantic Resources**

September 21, 2012
Vienna, Austria

Preface

Lexical-semantic contents serve as valuable sources of linguistic information for research and applications in natural language processing. Recent years have seen different kinds of efforts to enhance the usability of these resources. Successful attempts to integrate and merge lexical resources have led not only to the mapping of lexical databases of the WordNet and FrameNet type but also to the creation of online dictionary networks of different sorts. Enhancing usability also means to provide data for the human user by new sophisticated access interfaces as well as to make these data available for different kinds of NLP applications. In addition, with the increase of statistical methods in linguistics, these resources have also become more and more interesting for various research issues. The methods to achieve this rise in functionality and usability of lexical resources is a major topic of the workshop.

The workshop was conceived as a forum for discussing recent developments and applications of various kinds of lexical-semantic resources including WordNets, ontologies, FrameNets, electronic dictionaries, internet lexicons, and other collaboratively built lexical semantic resources. It was a follow-up to a series of thematically related events: GSCL workshops in 2003, 2005, 2007, 2008, and a DGfS workshop in 2006.

The six papers presented at the workshop focus partly on the creation of new lexical resources and partly on applications in language processing that are connected to lexical resources.

Godoy and colleagues report on the construction of a lexical resource of Brazilian Portuguese verbs. Based on Levin's hypothesis on the dependency of a verb's argument structure on the semantic class it belongs to, the authors supply verbs with predicate decompositions which in turn define semantic verb classes. The decompositions are linked to syntactic structures exemplified by example sentences. The resulting resource will be available in printed form as well as a lexical database.

Klyueva's paper focuses on the detection of dissimilarities between Czech and Russian verb valencies. Building on two existing resources, a Czech-Russian bilingual dictionary and a Czech valency dictionary, she explores how far verb semantics and semantic class membership might account for the differences. The study will result in a digital lexical resource that contains those verb pairs in Czech and Russian that differ in their valency frames. This resource is supposed to support applications in automatic language processing such as machine translation.

The paper by *Lyngfelt and colleagues* presents the development of a constructicon for Swedish as part of the Swedish FrameNet. Based on principles of Construction Grammar, it contains partially schematic multi-word units. The constructicon can be used as a lexicographic resource per se, is supposed to support language technology applications (e.g., facilitation of automatic syntactic analysis), and shall stimulate L2 acquisition research. The resources described in this paper are the result of activities carried out within Språkbanken, the national center for research on Swedish language resources at the University of Gothenburg.

Pado and Utt present ongoing work on the creation of a Distributed Memory resource for German, aiming at the determination of lexical meaning from the word's distributional behavior in corpora. Based on a German web-corpus, the study follows Baroni and Lenci's original proposal (dealing with

English) closely and discusses the adaptations necessary for German. The model is evaluated with respect to a synonym detection task.

The paper by *Nicolas and colleagues* describes an unsupervised method for acquiring translation pairs for word sequences from parallel corpora (German - Italian). The corresponding sequences can have different length in the two languages and are based on a set of different strategies for choosing and discarding candidates. The results are intended to be used by tools or lexical resources in the domain of translations, e.g., multilingual WordNets, machine translation, or corpus alignment.

Panchenko's paper, which is included in the main proceedings of KONVENS 2012, presents a novel approach to the measurement of semantic similarity, based on lexico-semantic patterns. The results are compared to human judgements and evaluated with respect to the tasks of semantic relation ranking and extraction.

Stefan Engelberg, Claudia Kunze
Workshop Organizers

Organizers:

Stefan Engelberg (Institut für Deutsche Sprache, Mannheim)
Claudia Kunze (Qualisys GmbH, Langenfeld)

Program Committee:

Hans Boas (University of Texas, Austin)
Kurt Eberle (Lingenio GmbH, Heidelberg)
Christiane Fellbaum (Princeton University)
Lothar Lemnitzer (Berlin-Brandenburgische Akademie, Berlin)
Carolin Müller-Spitzer (Institut für deutsche Sprache, Mannheim)
Sanni Nimb (Society for Danish Language and Literature, Copenhagen)
Rainer Osswald (Heinrich-Heine-Universität, Düsseldorf)
Uwe Quasthoff (Universität Leipzig)
Bolette Sandford Pedersen (University of Copenhagen)
Angelika Storrer (Technische Universität Dortmund)
Marko Tadíć (University of Zagreb)

Invited Speaker:

Alexander Panchenko (Université catholique de Louvain)

Workshop Program

Friday, September 21, 2012

- 09:00–09:10 Opening
- 09:10–09:50 *The construction of a catalog of Brazilian Portuguese verbs*
Márcia Cançado, Luisa Godoy and Luana Amaral
- 09:50–10:30 *Comparing Czech and Russian valency on the material of VALLEX*
Natalia Klyueva
- 10:30–11:00 Coffee break
- 11:00–11:40 *Adding a constructicon to the Swedish resource network of Språkbanken*
Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg and Sofia Tingsell
- 11:40–12:20 *A distributional memory for German*
Sebastian Padó and Jason Utt
- 12:20–13:50 Lunch break
- 13:50–14:30 *Towards high-accuracy bilingual sequence acquisition from parallel corpora*
Lionel Nicolas, Egon Stemle and Klara Kranebitter
- 14:30–15:10 Invited talk: *A semantic similarity measure based on lexico-syntactic patterns*
Alexander Panchenko
- 15:10–15:45 Coffee break

The construction of a catalog of Brazilian Portuguese verbs

Márcia Cançado

Faculdade de Letras, UFMG
Antônio Carlos, 6627
Belo Horizonte, Brazil
mcancado@ufmg.br

Luisa Godoy

Faculdade de Letras, UFMG
Antônio Carlos, 6627
Belo Horizonte, Brazil
luisagodoy@gmail.com

Luana Amaral

Faculdade de Letras, UFMG
Antônio Carlos, 6627
Belo Horizonte, Brazil
luanalopes@ufmg.br

Abstract

In this paper we present the construction of a lexical-semantic resource for the study of Brazilian Portuguese (BP): a “catalog” of BP verbs. The purpose of this catalog is to serve as a complete source of data, in which we present a large amount of verbs and sentences (over 800 verbs in over 5500 sentences in a first volume). An important characteristic of our project is that it is not a mere listing of verbs. Besides actually listing them, we group the verbs into semantically and syntactically coherent classes, adopting the hypothesis presented in Levin (1993) that semantic properties of verbs determine the syntactic realization of their arguments. We propose representations for the verbs using predicate decomposition, so that predicate decomposition structures define the verb classes. The syntactic properties of the verbs are presented in sample sentences and are related to the predicate decomposition structures. The relevance of our catalog is to be a complete resource for lexical-semantic studies in BP, not only a listing of verbs, but also a classification and an exhaustive exemplification. This catalog will take the forms of a printed version (Cançado, Godoy and Amaral, to appear - first volume) and a digital database.

1 Introduction

In this paper we present and comment on an empirical project that has been developed in the field of lexical semantics, taking Brazilian Portuguese (BP) as a language to be described. This project is the construction of a lexical-semantic resource

for the study of verbs in BP. We call it a “catalog” of the BP verbal lexicon, and it will take the forms of a printed version and a digital database.

The catalog is divided into a number of volumes, each one presenting verbs grouped according to semantic properties. Here, we present the first volume, which is ready to be published (Cançado, Godoy and Amaral, to appear). Subsequent volumes will be released in the future. This first volume contains **861 verbs of change**¹ from BP. All these verbs share a semantic property: they all denote some type of change. They are divided into 4 semantically and syntactically coherent classes *à la* Levin (1993): change of state, change of locative state, change of location, also known as “location verbs”, and change of possession, also known as “locatum verbs”. The class of change of state is the most representative one, with 685 verbs. We chose the verbs of change to start the construction of our catalog because we noted that they are very common verbs in the language, and they are the most studied verbs in the literature as well. Taking into account that there are actually approximately 6000 verbs in use in BP (Borba, 1990), we believe that the first volume of our catalog contains a significant amount of data.

2 The form of the catalog

Let us now describe and illustrate the form that our catalog takes. For each class of verbs, we present a listing of all the verbs we could find (the

¹This notion is related with the primitive predicate BE-COME. We will talk about that when we present the theoretical background of the catalog, in section 3.

verbs were collected mostly from an extensive BP verb dictionary: Borba, 1990) and of the semantic and syntactic properties of those verbs, as we illustrate below with the verb *quebrar* ‘break’, which belongs to the change of state class:

- (1) O João quebrou o vaso.
the John broke the vase
'John broke the vase.'
(*transitive causative with an agent subject*)
- (2) A queda quebrou o vaso.
the fall broke the vase
'The fall broke the vase.'
(*transitive causative with a cause subject*)
- (3) O vaso ficou quebrado.
the vase became broken
'The vase became broken.'
(*change of state entailment*)
- (4) O vaso (se) quebrou.
the vase (SE) broke
'The vase broke.'
(*intransitive inchoative with the optional se*)
- (5) O vaso (se) quebrou com a queda.
the vase (SE) broke with the fall
'The vase broke from the fall.'
(*intransitive inchoative with the optional se and the cause in adjunct position*)
- (6) O João quebrou o vaso com um martelo.
the John broke the vase with a hammer
'John broke the vase with a hammer.'
(*transitive causative with an instrument in adjunct position*)
- (7) O vaso foi quebrado (pelo João).
the vase was broken (by-the John)
'The vase was broken by John.'
(*passive construction*)

- (8) O menino quebrou o braço (com a queda).
the boy broke the arm (with the fall)
'The boy's arm broke from the fall.'
(*patient-possessor alternation with an optional cause in adjunct position*)

- (9) O João quebrou as paredes da casa com um ótimo pedreiro.
the John broke the walls of-the house with an excellent bricklayer
'John had the walls of the house broken by an excellent bricklayer.'
(*agent-possessor alternation*)².

We show that some specific construction is either allowed or prohibited by all members within a class. For example, for the classes which do not participate in the causative-inchoative alternation, we list ungrammatical examples of that alternation. The verb *aparelhar* ‘equip’, a verb of change of possession (also known as locatum), does not participate in the causative-inchoative alternation and it cannot have a cause as subject, as we show in the examples marked with * below:

- (10) O dentista aparelhou o consultório.
the dentist equipped the office
'The dentist equipped the office.'
(*transitive causative with an agent subject*)
- (11) *A necessidade de modernizar aparelhou o consultório.
the need to modernize equipped the office
(*transitive causative with a cause subject*)
- (12) O consultório ficou com aparelho.
the office became with equipment
'The office became with equipment.'
(*change of possession entailment*)
- (13) O dentista aparelhou o consultório com aparelhos de aço inox.
the dentist equipped the office with equipment of steel stainless

²The alternations agent-possessor and patient-possessor are very productive in BP. See Cançado (2010). They are also known in the literature as “possessor raising”.

'The dentist equipped the office with stainless steel equipment.'

(*cognate/instrument construction*)

- (14) *O consultório (se) aparelhou.
the office (SE) equipped
(*intransitive inchoative with the optional se*)
- (15) O consultório foi aparelhado (pelo dentista).
the office was equipped (by-the dentist)
'The office was equipped by the dentist.'
(*passive construction*)
- (16) O dentista se aparelhou.
the dentist SE equipped
'The dentist provided himself with equipment.'
(*reflexive*)
- (17) O dentista aparelhou o consultório com a melhor loja de aparelhos odontológicos da cidade.
the dentist equipped the office with the best store of equipment orthodontic of-the city
'The dentist had the office equipped by the best store of orthodontic equipment in the city.'
(*agent-possessor alternation*)

The syntactic properties reflexive, patient-possessor alternation and agent-possessor alternation are not exemplified when they do not occur (for example, we do not present a reflexive sentence for the verb *quebrar* in (1)-(9), nor an example of the patient-possessor alternation for the verb *aparelhar* in (10)-(17)) because they are not classificatory of the classes, they do not occur with all the verbs in a class and they also have pragmatic or other semantic restrictions (see section 3.2). The change entailment is an important semantic characteristic of the verbs, which distinguishes the semantics of the classes (e.g., distinguishes change of state from change of possession).

We provide the classes with an exhaustive exemplification, presenting carefully crafted examples of each one of the syntactic properties picked

up for observation, for each verb. So, for each verb belonging to the class of change of state, for example, we present an example of a sentence with the semantic and syntactic properties of the class, just like was done for *quebrar* and *aparelhar* above. In the first volume of the catalog we present over 5500 examples.

Since our purpose is to describe the BP lexicon as a system, we test the limits of grammaticality and deal with negative and intuitive data. So, as illustrated above, we present examples carefully crafted by us, not empirical data collected from corpus. Our methodology is very close to the one adopted in the research done by Maurice Gross (Gross, 1975, 1981), although the theoretical frame differs a lot. Research based on intuitive/introspective examples has been much criticized with the ascendancy of corpus linguistics, but we believe that this is the adequate methodology for the kind of research we develop and it holds many advantages, as argued for by Laporte (2008).

Our catalog contains, in the first volume dedicated to verbs of change, 861 verbs divided into 4 classes. The class of change of state, the most representative one, contains 685 verbs; the class of change of locative state contains 68 verbs; the class of change of location contains 15 verbs; and the class of change of possession contains 93 verbs³. For each verb we list syntactic and semantic properties, as illustrated above, totalizing over 5500 sentences.

Besides the presentation of the data, which is the most important part of our catalog, in the printed version we also present a theoretical explanation of our analysis (which will not be available in the digital database).

³We do not explain why there is such difference in the number of members of the classes of verbs of change, but the main hypothesis is that BP has a preference for lexicalizing change of state. The verbs from the class of change of location appear to be diachronically going to the class of verbs of change of locative state, another clue that BP likes having change of state verbs. For example, the verb *acomodar* 'place in a comfortable position' comes from the name *cômodo* 'room' (Cunha, 2010), so the original meaning of this verb would be 'place in room', but when forming a sentence with *acomodar* we can add any locative, not only one that is a room. For example, *a mãe acomodou a filha na cama* 'the mother placed the daughter in bed in a comfortable position'.

3 Theoretical Background

Let us now take a look at the theoretical background of our work. Our catalog starts from Levin's (1993) original insights about the importance of lexical description and her methodological ideas, taking the hypothesis of lexical semantic determinants of syntactic behavior seriously. We assume a strong version of Levin's (1993) methodology, admitting that if two verbs or groups of verbs behave alike, they must belong to the same class. For example, the verbs *quebrar* 'break' and *abrir* 'open' are transitive, participate in the causative-inchoative alternation, and can be passivized. This similar syntactic behavior, according to Levin (1993), is a clue that these verbs belong to the same class and share some grammatically relevant semantic property. So, an important part of our work is to divide the verbs in semantically and syntactically coherent classes, showing that semantic properties of verbs determine their syntactic behavior. It is important to note that our catalog is not a transposition of Levin's (1993) classes into BP. Our project is not a translation and all the verbs were collected directly from BP.

Another very important characteristic of our catalog is the metalanguage we use for representing the meaning of the verbs and their relevant semantic components: the semantic decomposition of verbs in primitive predicates, known as "predicate decomposition". Predicate decomposition is a more formal metalanguage to represent meaning than natural language, and it is also less problematic and more complete than a lexical representation in terms of thematic grids (Levin and Rappaport Hovav, 2005). Predicate decomposition structures can accommodate distinct types of semantic information, both thematic and aspectual, contained in a verb's meaning within a single representation. So it is not restricted to representing the semantic function of the arguments (like in thematic grids), but it represents also, and most importantly, the semantics of the event and its subparts. An important characteristic of the verbs of change is the presence of the predicate BECOME in their representations. So the common semantic feature that these verbs share is very clear in the semantic representations, as well

as the differences between them, as we will see below.

An important point in the predicate decomposition representations is the opposition between structural semantic information and idiosyncratic semantic information. The introduction of the concept of a "root" or "constant" was a major step in lexical representation (Levin and Rappaport Hovav, 2005) and is now widely accepted. We assume (along with Grimshaw, 2005 and others) that every grammatically irrelevant semantic information belongs to the root, so the relevant semantic information is represented in the structural part⁴. The structural part is shared by the verbs within the same class, so it does not only represent grammatically relevant semantic information, but also identifies a class.

In the catalog, we articulate the syntactic properties listed with the predicate decomposition structures proposed for each class. We motivate the existence of each component within a structure, in order to represent only those aspects of meaning which are relevant. So we try to propose well-motivated and economic structures (we represent in the structural part the relevant features leaving all the rest for the root). In the printed version of the catalog we provide explanation and independent evidence for the structures in (18)-(21).

Below, we present the representations for each verb class together with a short list of examples:

The structures of the verbs of change:

(18) [[X_(VOLITION)] CAUSE [BECOME Y
<STATE>]]

(change of state: *quebrar* 'break', *abrir* 'open', *machucar* 'hurt')

(19) [[X_{VOLITION}] CAUSE [BECOME Y
<STATE> IN Z]]

(change of locative state: *acomodar* 'place in a comfortable position', *esconder* 'hide', *pendurar* 'hang')

⁴We are aware, however, that some properties of the root may be relevant to some syntactic properties, which we call "non-classificatory". For example, the reflexive construction only occurs with verbs that select animate objects (Godoy, 2012). See section 3.2.

(20) [[XVOLUTION] CAUSE [BECOME Y IN
<PLACE>]]

(change of location: *engaiolar* ‘cage’, *engarrafar* ‘bottle’, *ensacar* ‘bag’)

(21) [[XVOLUTION] CAUSE [BECOME Y WITH
<THING>]]

(change of possession: *aparelhar* ‘equip’,
amanteigar ‘butter’, *colorir* ‘color’)

All the structures contain the predicate BECOME, which represents the change. But the structures that occur with BECOME are different, signaling that each class denotes a different kind of change. All verbs of change are transitive and causative, so the representations show the change (the substructure which contains BECOME) and the agent/cause responsible for bringing that change about (X), as well as the relation between them, represented by the predicate CAUSE. Let us now present what kind of semantic and syntactic properties we can predict for the verb classes by observing the predicate decomposition structures.

3.1 Semantic properties

There are three kinds of semantic information that we can derive from those structures: the thematic roles of the arguments, lexical aspect, and the kinds of changes denoted by the verbs. All that information is given for each class of verbs in the catalog.

Thematic roles are defined by Jackendoff (1990) as argument places in the predicate decomposition structure. X as the first argument of CAUSE receives the cause thematic role, X modified by VOLITION receives the agent thematic role, Y as the argument of BECOME receives the patient thematic role, and Z, argument of IN, receives the locative thematic role. All the verbs of change have a patient object, the verbs of change of state may have cause or agent subjects, the verbs of change of locative state, change of location and change of possession have agent subjects, and the verbs of change of locative state also have a third argument, which is a location.

Lexical aspect can also be defined in terms of predicate decomposition structures, as done by Dowty (1979). All the verbs of change denote

accomplishments, since they denote causative events, with two subevents related by the predicate CAUSE in the predicate decomposition structure. When the verbs participate in the causative-inchoative alternation (only verbs of change of state), the inchoative counterparts denote achievements, since they only denote the final result, the change of state.

Each kind of root, together with other elements, like the prepositions, will indicate if the change in question is a change of state, locative state, location or possession. For each verb in the catalog we show the sentence with the change entailment, so the verbs of change of state ([BECOME Y <STATE>]) entail the sentence *become state*, verbs of change of locative state ([BECOME Y <STATE> IN Z]) entail *become state in Z*, the verbs of change of location ([BECOME Y IN <PLACE>]) entail *become in location*, and the verbs of change of possession ([BECOME Y WITH <THING>]) entail *become with thing*.

3.2 Syntactic properties

Taking the hypothesis that the semantic properties of verbs determine their syntactic behavior, we are able to link certain predicate decomposition structures with certain syntactic properties.

The causative-inchoative alternation only occurs with the verbs of change of state, so it is sensible to the substructure [BECOME Y <STATE>]. Passive and instrument adjunction occur with agentive verbs, so these constructions are sensible to the substructure [XVOLUTION]. The adjunction of a cause will only occur with verbs that permit a cause subject, so verbs which have the substructure [X(VOLUTION)].

The other syntactic properties, reflexive, patient-possessor alternation and agent-possessor alternation, are sensible to the structure, but also have pragmatic and other kind of semantic restrictions. For example, the reflexivization occurs with agentive verbs ([XVOLUTION]), but it also requires that the verb selects an animate object, as proposed by Godoy (2012). For those different kinds of restrictions, the verbs within a class do not have exactly the same behavior in respect to these syntactic properties. We are aware of that and that fact is accounted for in our catalog. Syntactic properties that we call “non-classificatory”

are sensible to semantic properties or pragmatic properties that crosscut verb classes or that do not belong to all the verbs in a class. This does not invalidate the hypothesis that the verb classes are syntactically coherent, as we explain in length in the printed version of the catalog.

The verbs of change of state show a particularity in respect with their X argument. While all other classes have agent subjects, verbs of change of state may have agents or causes as subject, as observed by Cançado (2005, 2010). That is why we represent VOLITION between parentheses in the representation of that class ($[X_{(VOLITION)}]$). This property actually divides the class of change of state verbs into four subclasses: strictly volitive, optionally volitive, non-volitive (a class pointed out by Cançado and Franchi, 1999), and inchoative. The verbs of change of state belong to a broad class because all of them participate in the causative–inchoative alternation and entail *become state* as showed by Cançado and Godoy (2012). Subclasses arise because these verbs differ with respect to the agentivity of the external argument, and also with respect to the occurrence of the clitic *se* in the intransitive-inchoative sentences. As we showed in example (4), optionally volitive verbs like *quebrar* ‘break’, strictly volitive verbs (like *legalizar* ‘legalize’) and non-volitive verbs (like *deprimir* ‘depress’) may or may not occur with *se* in intransitive sentences. Inchoative verbs (like *apodrecer* ‘rot’), however, never occur with *se* (**a maçã se apodreceu*). The optionality of *se* in the three first subclasses of verbs of change of state seems to be dialectological, implying no meaning difference. The prohibition of *se* in intransitive sentences with inchoative verbs, however, seems to be a consequence of the fact that these verbs are basically intransitive, as argued for by Cançado and Amaral (2010).

The classes of verbs of change of locative state, change of location and change of possession are also differentiated by syntactic properties. The class of change of locative state is the only one with three arguments: *a mãe acomodou a menina na cama* ‘the mother placed the daughter in bed in a comfortable position’, as stated in Godoy (2012). The verbs of the class of change of location form sentences with a cognate location, like:

o homem engaiolou os pássaros em uma gaiola de ouro ‘the man caged the birds in a golden cage’, and form a different kind of reflexive, which we call “middle reflexive” (also known as “naturally reflexive”; Reinhart and Reuland, 1993). The verbs of the class of change of possession form sentences with a cognate possession, like: *o dentista aparelhou o consultório com aparelhos de aço inox* ‘the dentist equipped the office with stainless steel equipment’ and form canonical reflexives.

4 Some important observations

Here, we should mention a work similar to ours, the database called VerbNet⁵, which is a current descriptive project that was also inspired by Levin’s (1993) work. At first sight, our catalog might be seen as a translated version of it, not only because VerbNet is inspired in Levin’s verbal classes, but also because it uses predicate decomposition as semantic information. However, our catalog bears important differences in relation to that database. First, in VerbNet, the only syntactic property listed in the verbs entries is a sort of a subcategorization frame, and our catalog presents several syntactic properties that group the verbs into classes. Besides, we assume that predicate decomposition structures play a much more important role: first, they define the verb classes; second, thematic and aspectual properties can be derived from them (Jackendoff, 1990; Dowty, 1979). VerbNet does not assume a theoretical background in proposing representations – thematic roles and predicate decomposition structures are used only as semantic descriptions.

Another departure from VerbNet, and also from Levin (1993), is an obvious but not as trivial aspect as one might think – the fact that our catalog describes BP, not English. We take on the hypothesis that even when translation can occur, lexicalization might differ, if one is observing two different languages. Levin and Rappaport Hovav (1995) present the example of the two verbs, English *blush* and its Italian version *arrossire*, which are good translations of each other (e.g., they may have the same truth conditions), however they seem to belong to different classes in each lan-

⁵<http://verbs.colorado.edu/mpalmer/projects/verbnet.html>

guage, *blush* behaving as an unergative in English and *arrossire* behaving as an unaccusative in Italian. The analysis provided by VerbNet of the English verbal lexicon, therefore, cannot be trivially applied to the BP verbal lexicon. Plus, BP is an interesting alive language, whose grammar differs much from European Portuguese (for example, it was shown in Raposo, 1998 that the pronominal system of BP is very different from that of the other European Romance languages), which is something that is not accounted for in many studies that refer to Romance languages as a whole.

5 Applications in linguistic research

After presenting the project we developed for BP, we point out some of the applications of our catalog in linguistic research. First of all, any researcher who wants to study verbs in BP would benefit from a listing of the verbs of the language and their properties already classified into semantic classes. Also, our catalog will be the base for building an electronic database. We plan on building an online verb database where researchers could search verbs by their names, their thematic roles, their syntactic properties, their representations, and so on. For example, if one wants to study the passive construction in BP, he/she can go to our online database and search for verbs that occur in the passive construction. The result of the search will show all the verbs of change (in this first volume) that have this syntactic property as well as sample sentences. Another important application of our catalog is that it can be used by researchers who are not native BP speakers. With a list of verbs and sentences at hand one could, for example, compare BP with his/her own language in terms of lexicalization patterns, syntactic constructions available, classes of verbs, and so on. It would also be very helpful for the study of language typology, since the researcher could easily acquire a great amount of data without having to collect it.

6 Conclusion

In this paper, our purpose was to present the construction of a lexical-semantic resource for the study of verbs in BP. We wanted to show how we use semantic and syntactic properties of verbs to

group and to separate them in classes and subclasses, relating this grouping and some cross-classifications that might occur to components in a semantic predicate decomposition representation. The listing of syntactic properties exhaustively exemplified, along with the central role played by well-motivated and parsimonious predicate decomposition structures are the two central characteristics of our project. The main purpose of an exhaustive exemplification is to provide a complete and straightforward source of data of the BP verbal lexicon, which we hope is going to be very helpful for researchers. In doing this exhaustive exemplification, we also try to leave little space for the refutation, on an empirical ground, of the existence of the classes we propose and of the hypothesis of grammatically relevant components of lexical meaning that determine syntactic behavior.

It is worth noting that the catalog is not an isolated project. The group of lexical semantics, coordinated by Cançado, has been doing descriptions of the BP verbal lexicon and studying lexical representations for over 17 years now, so the catalog contains a sort of a compilation of the main results of her work, both individual and advising students, in the field of lexical semantics.

Acknowledgments

We thank CNPq and FAPEMIG the financial support for Márcia Cançado; we thank CNPq the financial support for Luisa Godoy; and we thank CAPES the financial support for Luana Amaral. Besides, we must thank Diógenes Evangelista for help with formatting the paper.

References

- Borba, Francisco. 1990. *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil* (Grammatical Dictionary of Verbs of Contemporary Brazilian Portuguese). Unesp, São Paulo, Brazil.
- Cançado, Márcia. 2005. Posições argumentais e propriedades semânticas (Argument positions and semantic properties). *DELTA*, 21(1):23-56.
- Cançado, Márcia. 2010. Verbal alternations in Brazilian Portuguese: a lexical semantic approach. *Studies in Hispanic and Lusophone Linguistics*, 3(1):77-111.

- Cançado, Márcia and Amaral, Luana. 2010. Representação lexical de verbos incoativos e causativos no português brasileiro (Lexical representation of inchoative and causative verbs in Brazilian Portuguese). *Revista da Abralin*, 9(2):123-147.
- Cançado, Márcia and Franchi, Carlos. 1999. Exceptional Binding with Psych-Verbs? *Linguistic Inquiry*, 30(1):33-143.
- Cançado, Márcia and Godoy, Luisa. 2012. Representação lexical de classes verbais do PB (Lexical representation of verb classes in BP). *ALFA*, 56(1):109-135.
- Cançado, Márcia; Godoy, Luisa and Amaral, Luana. to appear. Catálogo de verbos do português brasileiro, volume 1 (A catalog of Brazilian Portuguese verbs, volume 1). (printed version of the catalog presented in this paper)
- Cunha, Antônio. 2010. *Dicionário Etimológico da Língua Portuguesa (Etymological Dictionary of the Portuguese Language)*. Lexicon, Rio de Janeiro.
- Dowty, David. 1979. *Word Meaning and Montague Grammar*. D. Reidel, Dordrecht.
- Godoy, Luisa. 2012. *A reflexivização no PB e a decomposição semântica de predicados (Reflexivization in BP and predicate semantic decomposition)*. Phd dissertation, Faculdade de Letras, UFMG, Belo Horizonte, Brazil.
- Grimshaw, Jane. 2005. *Words and Structure*. CSLI Publications/University of Chicago Press, Stanford.
- Gross, Maurice. 1975. *Méthodes en syntaxe*. Hermann, Paris.
- Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 53:7-52.
- Jackendoff, Ray. 1990. *Semantic Structures*. MIT Press, Cambridge.
- Laporte, Éric. 2008. Exemplos atestados e exemplos construídos na prática do léxico-gramática (Attested examples and crafted examples in the practice of lexicon-grammar). *Revista (Con)textos Lingüísticos*, 2:26-51.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago.
- Levin, Beth and Rappaport Hovav, Malka. 2005. *Argument Realization*. Cambridge University Press, Cambridge.
- Levin, Beth and Rappaport Hovav, Malka. 1995. *Unaccusativity: at the syntax-lexical semantics interface*. MIT Press, Cambridge.
- Raposo, Eduardo. 1998. Some Observations on the Pronominal System of Portuguese. *CatWPL*, 6: 59-93.
- Reinhart, Tanya and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry*, 24 (4): 657-720.

Comparing Czech and Russian Valency on the Material of Vallex

Natalia Klyueva

Institute of Formal and Applied Linguistics

Charles University in Prague

klyueva@ufal.mff.cuni.cz

Abstract

In this study we have compared Czech and Russian valency frames based on monolingual and bilingual data. We assume that Czech and Russian are close enough to have, for the majority of their verbs, similar valency structures. We have exploited Vallex as a source of valency frames and have used a Czech-Russian dictionary to automatically translate Czech verbs into Russian. Afterwards, we have manually checked whether the Czech frame fits the Russian verb and, in case it was different, we have added the verb to the set that will be described in our paper. We suggest that there is a connection between the semantic class of some verbs and the type of difference between their Czech and Russian valency frames.

1 Introduction

Verbal valency is an important topic in Natural Language Processing which has been broadly studied within various linguistic branches - theoretical and practical. Bilingual research on valency is crucial for practical fields such as Machine Translation, or second language acquisition. There are many sources of information on both valency and word classes - WordNet(Fellbaum, 1998), FrameNet(Baker et al., 1998), VerbaLex(Hlaváčková and Horák, 2006) and Vallex(Lopatková et al., 2006) to name some of them. The central resource for our research has been the Czech Valency Lexicon Vallex. The Resource for Russian Valency, Explanatory Combi-

natorial Dictionary of Modern Russian(Mel'čuk and Zholkovsky, 1984), which is comparably big and rich in terms of language information is not available on-line, so we can not make a straightforward comparison. Instead, we have looked at the Russian verbal valency through the prism of the Czech one.

Czech and Russian are Slavic languages which are related and therefore share many morphological and syntactic features. A valency frame of a Czech verb is, in the majority of cases, similar to that of a Russian one. The focus of our study is on the verbs which have a different valency structure between the two languages. It seemed interesting for us not just to collect the set of those verbs, but rather to find out whether those Czech and Russian verbs that present some dissimilarity between their valency have some regularity or rule, or if this discrepancies are merely coincidental. Our hypothesis is that these differences have something to do with the semantic class of a verb.

There is a resource of valency bilingual data for Czech and Russian - the dictionary Ruslan (Oliva, 1989) that contains this information. But it is not big and it can only give us 'absolute' numbers - the percentage of verbs with different valency structure (Klyueva and Kuboň, 2010), without a insight into the nature of these dissimilarities. Vallex enables us to browse various verbs classes and see the underlying connections between the semantics of these verbs and the difference of frames in the two languages.

The idea of using data from Czech language in order to create new data for Russian was exploited by (Hana and Feldman, 2004), who constructed a

morphological tagger for Russian language upon Czech data and tools. In (Benešová and Bojar, 2006) authors compare the similarity between the automatically extracted valency frames and the manually annotated frames.

2 Vallex and Verb Classes

Vallex - a manually created Lexicon of Czech Verbs - is based on the valency theory in the Functional Generative Description (Panová, 1994), (Sgall et al., 1986). It provides an information on valency frames of the most frequent verbs (in version Vallex 2.5 over 2.700 lexeme multiplied by different senses of the verbs). The frame consists of a slot that reflects the number of complements the verb may govern. A slot includes a functor (a deep semantic role, written after a period attached to the word) and the surface realization of it (mostly morphosyntactic case, written in brackets). The main deep semantic roles that have frequently been used in our work are:

ACT:Actor, ex. I.ACT love peaches.

PAT:Patient, ex. Cats love rats.PAT

ADDR:Addressee(a person or an object to whom/to which the action is performed - more in the paper below), ex. He gave him.ADDR a book

DIFF:Difference measure, ex. Prices have fallen twice.DIFF

Verbs are classified into verb classes according to their meaning, which we have used in our research as well. Vallex distinguishes 22 verb classes, among them are communication, exchange, motion, perception, transport, psych verb, just to mention some. Naturally, words that belong to the same semantic field or share some component of meaning will have a similar valency frame. Vallex entry also provides other valuable information on aspect, reciprocity, reflexivity etc. that we have not used in our work, so it will not appear in our examples. Here is an example of a Czech verb frame that belongs to the Mental Action verb class:

apelovat Act(Nom) Pat(na+Acc)-(on+Acc). This means that the verb *to appeal* governs two arguments: an Actor in the Nominative case and a Patient in prepositional phrase on+Accusative. The case systems in Czech and Russian are very similar and prepositions have almost identical surface form which simplifies the process of comparison.

3 Czech Vallex for Russian verb frames

We made a comparison of Czech and Russian frames based on the Czech Lexicon in the following way. We took a Czech verb and said if its frame fits the frame of a Russian equivalent verb as well. At this stage, it was impossible to evaluate a big amount of verb frames (totally 2,903 lexical units have a verb class assigned), so we took the selection of the most frequent verbs distributed among the following verb classes: motion, communication, change, exchange and mental action, as the most representative ones.

This verb set contains frequent verbs of various semantic types. Our assumption is that the difference in valency frames might be related to a verb class, in other words, verbs from certain classes might have tendency to have different valency in Czech and Russian. In our study we focus on morphemic forms of **noun complements**, leaving aside verb complements and sentence complements of verbs. Within a semantic class for each Czech verb we state whether or not a Russian equivalent has the same valency structure.

For example, (1) shows the verb with the same valency frame and the verb in example (2) has two discrepancies in it.¹

(1cz)*obhajovat* ACT(Nom) PAT(Acc), to defend

(1ru)*zaščiščat'* ACT(Nom) PAT(Acc), to defend
The frame is the same in both languages

(2cz)*blahopřál mu.ADDR(Dat) k narozeninám*
'congratulated him.ADDR(Dat) to birthday'

(2ru)*pozdravljal ego(Acc) s dnem roždenija*
'congratulated him.ADDR(Acc) with birthday(with+Ins)'

In the example (2) it is illustrated that in Czech and Russian different prepositions and different cases are used to express the same semantic roles - Patient and Addressee. Especially diverse in this case is the surface realization of Patient as a prepositional phrase across the languages: Czech

¹There are 6 cases in Russian and 7 cases in Czech (7th, Vocative, is not relevant for our study) and case endings are very similar in both languages. Czech and Russian prepositions are almost identical as well. All this makes it rather easy to detect differences in valency frames.

- congratulate to, Russian - congratulate with, English - congratulate on/upon.

We consider the Russian frame to be similar to the Czech one if it has the same number complements, the same semantic roles and if these semantic roles have the same surface realization. All the verbs we observed met the first two conditions because we tried hard to find the closest translation equivalent in Russian. It was always the surface form that was different in two languages. If a surface form is represented by a preposition with some case, we judge the default translation of prepositions as the similar realization.

Further on, to simplify the examples, we will leave only the slot of the frame that is different in the languages and leave out the slots that are irrelevant to our comparison. So the example verb (2cz) will be shortened to *blahopřát* PAT(k+Dat) ADDR(Dat) and (2ru)*pozdravljat'* PAT(s+Ins) ADDR(Acc) 'leaving aside the functor ACT(Nom) which is almost always the same in Czech and Russian. The examples in this paper are either taken from corpus, invented or taken straightly from Vallex examples.

4 Differences According to the Verb Classes

While analyzing Czech and Russian frames, it became evident that the differences between Valency frames can be either regular or occasional. In this paper we will present the description of the differences according to the semantic classes of the verbs. Some groups of verbs that have some regular discrepancy in a valency frame may belong to different classes, as it will be illustrated below.

4.1 Class of Change

Verbs of the class Change often have the complement DIFF, and we observed that it often has different realization in Czech and Russian, namely the slot cz:DIFF(o+Acc)-(about+Acc) generally corresponds to ru:(na+Acc)-(on+Acc) in Russian (other variations are possible), see examples (3) and (4).

(3cz)*ceny klesly o 20%* 'prices fall **about** 20%'
(3ru)*ceny upali na 20%* 'prices fall **on** 20%'

- (4cz) *Administrace zkrátila dovolenou o 2 dny*
'administration cut off the holiday **about** 2 days'
(4ru) *Administracija sokratila otpusk na 2 dnja*
'administration cut off the holiday **on** 2 days'

For the functor DIFF, we should mention, that the form (about+Acc) is typical of Czech while Russian language uses the preposition 'o' (about) mainly with mental predicates like English(forget about+Loc) or communication verbs (tell about+Loc) and does not occur with the Accusative case at all.

4.2 Class of Motion

We have not found many dissimilarities in Czech and Russian valency frames within the class of Motion verbs. One most evident is that verbs of classes motion with the semantic component of 'going away from somewhere' in the case they have the surface realization of PAT as (před+Ins)-(before+Ins) in Czech are translated into Russian with the respective verb plus the prepositional phrase (ot+Gen)-(from+Gen), not the expected ru:(pered +Ins): *prchat*, *ujízdět*, *unikat*.

- (5cz)*prchat před* policii- 'run before police'
(5ru)*ubegat' ot* policii 'run from police'

In other words, Russian prefers the preposition 'from' whereas Czech uses 'before' in this context. Verbs of other semantic classes with the similar component of meaning, ex. class location - share this rule as well(cz:schovat před+Ins - 'to hide before' vs. ru:spryatat' ot+Gen - 'to hide from').

The following example illustrates a coincidental difference in verb frame :

- (6cz)*trefit* PAT(**Acc**) 'to hit smth'
(6ru)*popast'* PAT(v+Acc) 'to hit into+Acc'

4.3 Verbs of Exchange

One of the regular and rather evident differences between Czech and Russian frames was described in (Lopatková and Paněnová, 2006). This is the case of some exchange verbs with the meaning of removing something from someone, ex. *sebrat*(take away), *krást*(steal), *brát*(take) etc. The addressee here is a person or an object from

whom/which something is taken.

- (7cz)brát ADDR(**Dat**)-'take +Dat'
bere dítěti hračku -'takes baby.Dat toy'
(7ru)brat' ADDR(**u+Gen**)-'take (u+Gen)'
beret u rebenka igrushku 'He takes of baby.Gen
toy'
(7en)'He takes a toy from a baby'

- (8cz)zabírat ADDR(**Dat**)-take(time)+Dat
studium mi zabírá hodně času
'study me.Dat takes many time'
(8ru) otnimat' ADDR(**u+Gen**)
ucheba otnimaet u menja mnogo vremeni
'study takes from me many time'
(8en)'Study takes me a lot of time.'

In these cases if the sentence (8cz) was translated into Russian according to the Czech valency pattern, they would have the reverse meaning in Russian, because the Dative case of the noun in this context is understood as Benefactor (taken TO someone), not Addressee (taken FROM someone). Especially this difference causes big problems to learners of foreign languages: they project the known pattern from their native language onto the phrase in the foreign language and, given that the surface form of the preposition is the same, they make a mistake.

This scheme does not work for all words with this meaning in this class, for example a semantically related word 'odpírat'(to deny) in Russian has the same surface form of Addressee in Russian(ADDR(Dat)) as in Czech, yet another non-direct realization of Patient:

- (9cz)odpírat ADDR(Dat) PAT(**Acc**)
odpíral mu pomoc
'denied him help'
(9ru)otkazyvat' ADDR (Dat) PAT(**v+Loc**)-
(in+Loc)
on otkažal emu v pomošči
'denied him in help'
(9en)'he denied to help him'

On the example of this verb class we can see that the semantically related group of words has different surface realizations of a functor (ADDR in this case) in Czech and Russian. This makes

us believe that difference in valency frames can depend on the semantic class. Only two words of this class with different valency framedo not belong to the group described, and we consider them to be occasional discrepancies. The number of occasional discrepancies in the verb classes is not so big in comparison with ones that have some regular difference.

4.4 Class of Communication

Czech and Russian verbs belonging in this class have many differences with respect to valency. Here we could not observe some of the leading difference present in the previous classes. Differences may concern several functors and several surface forms. They may be considered coincidental, but we can allocate several groups of verbs with some dissimilarity in valency frames.

1. The functor Addressee with the surface form ADDR(**na+Acc**)-(on+Acc) in Czech is presented in another way in Russian

- (10cz)*mluvit* (**na+Acc**)-'speak on smb(Acc)'
(10ru)*obraščat'sja* ADDR(**k+Dat**)-'speak to
smb(Dat)'

- (11cz)*zavolat* (**na+Acc**)-'call on smb(Acc)'
(11ru)*pozvat'* (**Acc**) -'to call smb(Acc)'

2. Patient with the surface form (**na+Loc**)-(on+Loc) in Czech corresponds to another realization in Russian, generally the morphemic form is (**o+Loc**)-(about+Loc) for such verbs used for 'asking question' as (ze)ptát se, tázat se etc.:

- (12cz)*ptát se* PAT(**na+Acc**) *zdraví* -'ask on
health'
(12ru)*sprosít'* PAT(**o+Loc**) *zdorov'je* -'ask about
health'

Other verbs with a frame slot PAT(**na+Loc**)-(on+Loc) are also very similar to the above sample:

- (13cz)*domlouvat se* PAT(**na+Loc**) - 'to agree on'
(13ru)*dogovorit'sja* PAT(**o+Loc**) 'to agree about'

3. Addressee in Dative case for the following verbs corresponds to Accusative in Russian:

(14cz)*poblahopřát* ADDR(**Dat**) 'congratulate +Dat'
(14ru)*pozdravit'* ADDR(**Acc**) 'congratulate +Acc'

(15cz)*děkovat* ADDR(**Dat**) 'thank +Dat'
(15ru)*blagodarit'* ADDR(**Acc**) 'to thank + Acc'

4. Similar to the verbs of Exchange class, some Czech communication verbs with surface form (o+Acc)-(about+Acc) will be translated in another manner in Russian due to the fact that, unlike in Czech, the preposition 'o'- 'about' does not combine with the Accusative:

(16cz)*hlásit se* PAT(**o+Acc**):
hlásí se o slovo 'ask about word'
(16ru)*prosít'* PAT(**Gen**)
Ona prosít slova
'She ask word.gen'
(16en)'ask for a word'

Coincidental differences occurring only once or twice are not going to any scheme:

(17cz)*doznávat se* PAT(**k+Dat**) 'confess to smth'
(17ru)*priznavat'sja* PAT(**v+Loc**) 'to confess in smth'
(18cz)*konzultovat* PAT(**Acc**) 'to consult +Acc'
(18ru)*konsultirovat'* PAT(**po+Loc**)-(about+Loc)

4.5 Class of Mental Action

Verbs of this class often have differences in valency frames, but they are rather coincidental and we have found only one regular difference - when Czech PAT(na+Acc)-(on+Acc) corresponds to Russian PAT (o+Loc)-(about+Loc), (pro+Acc)-(about+Acc) or (k+Dat)-(to+Dat). The surface form (na+Acc) in Czech is also different for verbs belonging in the class Communication, but for that class it was regularly translated as (o+Loc)-(about+Loc) whereas for the class of Mental Action no common translation equivalent exists.

(19cz)*pamatovat* PAT(**na+Acc**) 'to remember on'

(19ru)*pomnit'* PAT(**pro+Acc**) 'to remember about'
(20cz)*myslet* PAT(**na+Acc**) 'to think on'
(20ru)*dumat'* PAT(**o+Loc**) 'to think about'

(21cz)*zvykat si* PAT(**na+Acc**) 'get used on'
(21ru)*privykat'* PAT(**k+Dat**) 'to get used to'

The structure of the following verb coincides a lot with that from ex. (14) and (15) though the functor is PAT, not ADDR:

(22cz)*rozumět* PAT(**Dat**) 'understand'
(22ru)*ponimat'* PAT(**Acc**) 'understand'

Other coincidental differences:

(23)*pohrdat* PAT(**Ins**)
(23)*prezirat'* PAT(**Acc**) 'to despise'
(24cz)*mrzet* ACT(**Acc**)
(24ru)*sožalet* ACT(**Nom**) 'to be sorry for'
The example (24) is the one of a very few verbs with different surface realization of ACTor.

4.6 Overall results

We have compared Czech and Russian valency frames of verbs from 5 semantic classes, totally 1473 lexical entries. 111(7.5%) of them were different in Czech and Russian. The comparison was rather straightforward because of the relatedness of the languages. If some more distant languages were compared, more complicated method of evaluation should be chosen. From the examples above we can make the following observations:

- most dissimilarities occur in prepositional phrases.
- the regular discrepancies are more frequent than the coincidental ones.
- Within a verb class we can find some typical valency patterns of Czech verbs which correspond regularly to the different Russian pattern.

The table 1 presents the distribution of verbs with different frames according to the verb classes.

From this table we can see that verbs of physical activity(change, motion, exchange) have in

Verb class	same frame	different frame	# of verbs
Change	309(95%)	14(5%)	323
Exchange	166(92%)	13(8%)	179
Motion	305(99%)	3(1%)	308
Communication	312(88%)	42(12%)	354
Mental Action	270(87%)	39(13%)	309
Total	1362(92%)	111(8%)	1473

Table 1: Differences according to the verb classes

some sense less complicated valency structures than verbs of mental activity(communication, mental action) and that in most cases, their valency structure corresponds to that of Russian verbs.

5 Conclusion

In this paper we have described the dissimilarities in Czech and Russian Valency based on the material of the Czech lexicon. Our main hypothesis was that the differences in valency structure might be explained by the semantics of verbs , so we have exploited the classification of the semantic classes provided by Vallex. In almost in each verb class we have found some regular dissimilarity that is typical of this class. Still, there are some cases when verbs from other classes are subjected to this regularity as well, so other aspects (such as surface realization) should also be taken into consideration. A practical result of our paper is that we have made a draft version of a small bilingual Czech-Russian lexicon with different frames in the Vallex format.

Acknowledgments

The research is supported by the grants P406/2010/0875 GAČR and GAUK 639012.

References

- Ch. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press
- C. Baker, C. Fillmore and J. Lowe. 1998. The Berkeley FrameNet Project. *Proceedings of the 17th international conference on Computational linguistics - Volume 1 (COLING '98)*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 86-90.
- V. Benešová and O. Bojar. 2006. Czech Verbs of Communication and the Extraction of their Frames. *Proceedings of the 9th International Conference, TSD 2006*, pages 29-36.
- J. Hana and A. Feldman. 2004. Portable Language Technology: Russian via Czech. *Proceedings of the Midwest Computational Linguistics Colloquium*, June 25-26, 2004, Bloomington, Indiana.
- D. Hlaváčková and A. Horák. 2006. VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. *Computer Treatment of Slavic and East European Languages*. Bratislava, Slovakia: Slovenský národný korpus, p. 107-115.
- N. Klyueva and V. Kuboň. 2010. Verbal Valency in the MT Between Related Languages. *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features* Universita di Pisa - Dipartimento di Lingistica, Pisa, Italy, pp. 160-164.
- M. Lopatková and J. Panevová. 2006. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. *In María Šimková, editor, Insight into Slovak and Czech Corpus Linguistic*, pages 83-92. Veda Bratislava, Slovakia.
- M. Lopatková, Z. Žabokrtský and V. Benešová. 2006. *Valency Lexicon of Czech Verbs VALLEX 2.0*. Technical Report 34, UFAL MFF UK.
- I. Mel'čuk and A. Zholkovsky. 1984. *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary*. Vienna: Wiener Slawistischer Almanach.
- K. Oliva. 1989. *A parser for Czech implemented in systems Q*. Praha:MFF UK
- J. Panevová. 1994. Valency Frames and the Meaning of the Sentence. *The Prague School of Structural and Functional Linguistics* (ed. Ph. L. Luelsdorff), Amsterdam-Philadelphia, John Benjamins, pp. 223-243.
- P. Sgall, E. Hajičová and J. Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague,Czech Republic/Dordrecht, Netherlands

Adding a Constructicon to the Swedish resource network of Språkbanken

Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice,
Rudolf Rydstedt, Emma Sköldberg, Sofia Tingsell

Department of Swedish, University of Gothenburg
first-name.last-name@svenska.gu.se

Abstract

This paper presents the integrated Swedish resource network of Språkbanken in general, and its latest addition – a constructicon – in particular. The constructicon, which is still in its early stages, is a collection of (partially) schematic multi-word units, constructions, developed as an addition to the Swedish FrameNet (SweFN). SweFN and the constructicon are integrated with other parts of Språkbanken, both lexical resources and corpora, through the lexical resource SALDO. In most respects, the constructicon is modeled on its English counterpart in Berkeley, and, thus, following the FrameNet format. The most striking differences are the inclusion of so-called collostructional elements and the treatment of semantic roles, which are defined globally instead of locally as in FrameNet. Incorporating subprojects such as developing methods for automatic identification of constructions in authentic text on the one hand, and accounting for constructions problematic for L2 acquisition on the other, the approach is highly cross-disciplinary in nature, combining various theoretical linguistic perspectives on construction grammar with language technology, lexicography, and L2 research.

1 Introduction

Large-scale linguistic resources typically consist of a lexicon and/or a grammar, and so do the linguistic components in language technology (LT) applications. Lexical resources mainly account for words, whereas grammars focus on general

linguistic rules. Arguably, this holds both for knowledge-driven and data-driven language processing, since what counts as a “word” (or “token”) is determined *a priori* in both cases. Consequently, patterns that are too general to be attributed to individual words but too specific to be considered general rules are peripheral from both perspectives and hence have tended to be neglected. Such constructions are not, however, peripheral to *language*, and neither are they a trivial phenomenon that can simply be disregarded. On the contrary, semi-productive, partially schematic multi-word units are highly problematic for language technology (Sag et al., 2002), L2 acquisition (Prentice and Sköldberg, 2011), and, given that idiosyncracies are typically attributed to the lexicon, lexicography. They are also quite common (cf. e.g., Jackendoff 1997, 156). Accordingly, constructions have received more attention in recent years, but resources with large-scale coverage are still lacking.

In response to this situation, we are currently building a Swedish constructicon, a collection of (partially) schematic multi-word units, based on principles of Construction Grammar and developed as an addition to the Swedish FrameNet (SweFN). It will be integrated with other resources in Språkbanken by linked lexical entries. The constructicon project is a collaboration involving experts on (construction) grammar, language technology, lexicography, phraseology, second language research, and semantics.

The resource environment of Språkbanken is treated in section 2, and the work on integrating the resources in section 3. Constructions and

Construction Grammar are introduced in section 4 and the Swedish constructicon is presented in section 5, followed by an outlook in section 6.

2 Språkbanken

Språkbanken (the Swedish Language Bank)¹ is a research and development unit at the University of Gothenburg, which was established with government funding already in 1975 as a national center for research on Swedish language resources, in particular corpora and lexical resources. The main focus of Språkbanken's present-day activities is in the development and refinement of language resources and LT tools, and their application as research and teaching tools in various fields outside LT itself – several areas of linguistics: descriptive, typological, historical and genetic linguistics (e.g., Saxena and Borin 2011; Rama and Borin 2011), Swedish as a second language (e.g., Johansson Kokkinakis and Magnusson 2011; Voldina and Johansson Kokkinakis 2012), computer-assisted language learning, text complexity and lexical semantics (e.g., Borin 2012); other humanities disciplines: comparative literature (e.g. Borin and Kokkinakis 2010; Oelke et al. 2012) and history (e.g. Borin et al. 2011); and medicine and medical informatics (e.g., Kokkinakis 2012; Heppin 2011) – i.e., activities that can be broadly characterized as LT-based eScience.

Språkbanken's LT research activities and in-house LT tools are characterized by a strong reliance on linguistic knowledge encoded in rich and varied lexical resources. The present focus is on the creation of a highly interlinked resource infrastructure informed by current work on LT resource standardization (e.g., in CLARIN, ISO TC37/SC4, and META-SHARE), as well as by work on linked open data (see, e.g., Chiarcos et al. 2012). This is the Swedish FrameNet++ project described in the next section.

3 Swedish FrameNet++

The goal of the Swedish FrameNet++ project (Borin et al., 2010a) is to create a large integrated lexical resource for Swedish – so far lacking – to be used as a basic infrastructural component in Swedish LT research and in the development of

LT applications for Swedish. The specific objectives of the project are

- integrating a number of existing free lexical resources into a unified lexical resource network;
- creating a full-scale Swedish FrameNet with 50,000 lexical units;
- developing methodologies and workflows which make maximal use of LT and other tools in order to minimize the human effort needed in the work.

3.1 The lexical resource network

The lexical resource network has one primary lexical resource, a pivot, to which all other resources are linked. This is SALDO² (Borin and Forsberg, 2009), a large (123K entries and 1.8M wordforms), freely available morphological and lexical-semantic lexicon for modern Swedish. It has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other resources are linked.

The standard scenario for a new resource to be integrated into the network is to (partially) link its entries to the sense PIDs of SALDO. This typically has the effect that the ambiguity of a resource becomes explicit: the bulk of the resources associate lexical information to PoS-tagged baseforms, information not always valid for all senses of that baseform. This is natural since most of the resources have initially been created for human consumption, and a human can usually deal with this kind of underspecification without problem. Some of these ambiguities can be resolved automatically – especially if information from several resources are combined – but in the end, manual work is required for complete disambiguation.

The network also includes historical lexical resources (Borin et al., 2010b; Borin and Forsberg, 2011), where the starting point is four digitized paper dictionaries: one 19th century dictionary, and three Old Swedish dictionaries. To make these dictionaries usable in a language technology setting, they need morphological information, a work that has been begun in the CON-

¹<http://spraakbanken.gu.se/eng/start>

²<http://spraakbanken.gu.se/saldo>

PLISIT project for 19th century Swedish (Borin et al., 2011) and in a pilot project for Old Swedish (Borin and Forsberg, 2008).

Linking SALDO to the historical resources is naturally a much more complex task than linking it to the modern resources, especially when moving further back in time. The hope is that a successful (but possibly partial) linking introduces the possibility to project the modern lexical-semantic relations onto the historical resources, so that, e.g., a wordnet-like resource for Old Swedish becomes available for use.

3.2 Swedish FrameNet

The Swedish FrameNet builds on the Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2010), using its frame inventory and accompanying semantic roles. The current size of the Swedish FrameNet is 632 frames, 22,548 lexical units (+1,582 suggested lexical units), and 3,662 annotated examples, which may be compared with the size of Berkeley FrameNet with 1,159 frames, 12,601 lexical units, and 193,862 annotated examples. Some new frames have been created, but none of them are motivated by differences between English and Swedish, and they could just as well have been present in the Berkeley FrameNet.

3.3 Methodology development

The methodological work is conducted within the lexical infrastructure of Språkbanken, described by Borin et al. (2012b). Some of the features of the infrastructure are: daily publication of the resources, both through search interfaces and for downloading; a strong connection to the corpus infrastructure (Borin et al., 2012c); formal test protocols; statistics; and change history.

An important methodological task is the development of automatic methods for locating good corpus examples for the Swedish FrameNet. The task has been explored by, e.g., Kilgarriff et al. (2008) and Didakowski et al. (2012), but the notion of what constitutes a good example is still under active research. An important step has been taken by Borin et al. (2012a), where a tool has been developed that enables ranking of corpus examples based not only on a (tentative) measure of goodness but also on diversity (ideally, the exam-

ples should cover the full usage range of a linguistic item).

4 Constructions

Language consists to a quite large extent of semi-general linguistic patterns, neither general rules of grammar nor lexically specific idiosyncrasies. Such patterns may be called constructions (cx). Peripheral from the view-point of grammar as well as lexicon, they have a long history of being neglected, despite being both numerous and common. For the last few decades, however, the study of constructions is on the rise, due to the development of Construction Grammar (CxG; Fillmore et al. 1988; Goldberg 1995, and others) and other cx-oriented models. Furthermore, cx have also been gaining increased attention from some lexicalist perspectives, e.g., Head-Driven Phrase Structure Grammar (HPSG; Pollard and Sag 1994), especially through the CxG-HPSG hybrid Sign-Based Construction Grammar (SBCG; Sag 2010; Boas and Sag to appear). Still, these approaches have mostly been applied to specific cx, or groups of such. To date, there are few, if any, large-scale constructional accounts.

Cx are typically defined as conventionalized pairings of form and meaning/function. Hence, linguistic patterns of any level, or combination of levels, from the most general to the most specific, may be considered cx and therefore relevant for a constructicon. Since our goal is a constructicon of wide applicability we do not wish to exclude any types of cx beforehand; it may well turn out that many relevant cx are of types that have been previously overlooked. Indeed, one of the expected benefits of this project is coverage of cx not yet accounted for. On the other hand, we do not have infinite resources. We will therefore focus on semi-general cx in the borderland between grammar and lexicon, since this is where better empirical coverage is most sorely needed.

There are also some cx types that we are particularly interested in. These include cx of relevance for L2 acquisition, e.g., date expressions, which can display surprising complexities and idiosyncrasies (Karttunen et al., 1996). Although time adverbials are usually expressed as PPs in Swedish, this is not the case if the time is a date: *Hon åker (*på) 7 maj* ‘She will leave on May 7th’, as op-

posed to *Hon åker på måndag* ‘She will leave on Monday’. In L2 Swedish, incorrect inclusion of the preposition is not uncommon: **Jag är född på 2 mars* ‘I was born on March 2nd’ (Prentice, 2011).

Of general theoretical interest are argument structure cx, which concern matters of transitivity, voice, and event structure, and are at the heart of discussions on the relationship between grammar and lexicon. Argument structure is usually assumed to be determined by lexical valence, but there are good reasons to assume that syntactic constructions also play a role (Goldberg, 1995). Consider, for instance, the (Swedish) Reflexive Resultative Cx (Jansson, 2006; Lyngfelt, 2007), as in *äta sig mätt* ‘eat oneself full’, *springa sig varm* ‘run oneself warm’, and *byta sig ledig* ‘swap oneself free’. Its basic structure is Verb Reflexive Result, where the result is typically expressed by an AP, and its meaning roughly ‘achieve result by V-ing’. (Hence, an expression like *känna sig trött* ‘feel tired’ is not an instance of this cx, since it does not mean ‘get tired by feeling’.) This pattern is applicable to both transitive and intransitive verbs, even when it conflicts with the verb’s inherent valence restrictions. Notably, in the case of transitive verbs, the reflexive object does not correspond to the object role typically associated with the verb; for example, the *sig* in *äta sig mätt* does not denote what is eaten. Such cx raise theoretically interesting questions regarding to what extent argument structure is lexically or constructionally determined.

From a structural perspective, a cx type of high priority are so-called partially schematic idioms (cf. Fried to appear; Lyngfelt and Forsberg 2012), i.e., cx where some parts are fixed and some parts are variable. Typical examples are conventionalized time expressions like [minuttal] *i/över* [timtal] ‘[minutes] to/past [hour]’ and semi-prefab phrases such as *i ADJEKTIV-aste laget* ‘of ADJECTIVE-superlative measure’. The latter cx basically means ‘too much of the quality expressed by the adjective’: *i hetaste laget* ‘too hot for comfort’, *i minsta laget* ‘a bit on the small side’, *i senaste laget* ‘at the last moment’. These cx are somewhat similar to fixed multi-word expressions and are fairly close to the lexical end of the cx continuum. They should be easier to iden-

tify automatically than fully schematic cx, and are therefore a natural initial target for the development of LT tools. Also, these cx are the ones closest at hand for integration into lexical resources.

Parallel to Construction Grammar, Fillmore (1982) and associates have also developed Frame Semantics, which in turn constitutes the base for the FrameNet resource (Baker et al., 1998; Ruppenhofer et al., 2010). Frame Semantics and FrameNet treat meaning from a situation-based perspective and put more emphasis on semantic roles and other cx-related features than other lexical resources usually do. By its historical and theoretical connections to Construction Grammar, FrameNet is well suited for inclusion of constructional patterns. There is also a growing appreciation for the need to do so. Accordingly, an English constructicon is being developed as an addition to the Berkeley FrameNet (Fillmore, 2008; Fillmore et al., to appear). In a similar fashion, the Swedish constructicon will be an extension of SweFN (Lyngfelt and Forsberg, 2012). Furthermore, there are plans to add constructicons to the Japanese and the Brazilian Portuguese FrameNet.

5 The Swedish constructicon

The Swedish constructicon is still in its early stages of development, so far numbering only a few sample cx, but it is growing and getting more refined by the day. Its format is for the most part modeled on Berkeley’s English constructicon, and thus on FrameNet. The core units in a constructicon, however, are not frames but cx; and instead of frame elements, there are cx elements, i.e. syntactic constituents.

As in the Berkeley constructicon, the cx are presented with definitions in free text, schematic structural descriptions, definitions of cx elements, and annotated examples. We try to keep the analyses simple, to make the descriptions accessible and reduce the labor required. This goes against common practice in linguistic research, where depth and detail usually get higher priority than simplicity. In the words of Langacker (1991, 548), “the meaning of linguistic expressions cannot be characterized by means of short, dictionary type definitions”.

While Langacker is of course right about linguistic meaning being complex and multi-faceted,

reflexiv_resultativ

type	Cx
category	vbm
evokes	Causation_scenario
definition	[Någon] _{Actor} eller [något] _{Theme} utför eller undergår [en aktion] _{Activity} som leder (eller antas leda) till att [aktören] _{Actor} / [temat] _{Theme} , uttryckt med reflexiv, uppnår [ett tillstånd] _{Result} .
structure	vb refl AP
cee	refl
coll	{äta ¹ : mätt ¹ } {supa ¹ : full ² } {skrika ¹ : hes ¹ } springa ¹
internal construction elements	<ul style="list-style-type: none"> ■ role: name=Activity cat=vb ■ role: cx=refl name=Actor ■ role: cx=refl name=Theme ■ role: name=Result cat=AP
external construction elements	<ul style="list-style-type: none"> ■ role: name=Actor cat=NP ■ role: name=Theme cat=NP
examples	<ul style="list-style-type: none"> ■ [Vi åskådare]_{Actor} [springer]_{Activity} [oss]_{Actor} inte [varma]_{Result}]_resultativ_reflexiv direkt. ■ [Kornet och havren]_{Theme} får [frysa]_{Activity} [sig]_{Theme} [mogen]_{Result}]_resultativ_reflexiv. ■ [Drick]_{Activity} [dig]_{Actor} [smal]_{Result}]_resultativ_reflexiv i vår.
comment	Det finns också en PP-variant med resultativ betydelse, t.ex. "träna sig i form", som ev. bör inkorporeras här - alt. betraktas som en metaforisk utvidgning av någon rörelse-cx.
reference	Jansson, Håkan (2006): Har du ölat dig odödig? En undersökning av resultativkonstruktioner i svenska. (D-uppsats, även publicerad som MISS 57) http://hdl.handle.net/2077/19000 Lyngfelt, Benjamin (2007): Mellan polerna. Reflexiv- och deponenskonstruktioner i svenska. Språk och stil NF 17: 86–134. http://hdl.handle.net/2077/21731

Figure 1: The reflexive-resultative construction

the approximations presented in dictionaries have after all turned out to be quite useful in many respects. Our expectation is that a corresponding level of complexity will work for a construction as well. More detailed analyses are both space and especially time consuming and therefore difficult to conduct on a large scale. Hence, simplicity is a main priority. Still, it is necessary to add some complexity compared to lexical definitions, since descriptions of syntactic cx also must contain constituent structure. Therefore, initially the core of the cx descriptions consists of a simple structural sketch and a free text definition of dictionary type. The intention is to refine and extend the description formalism incrementally as

needed to reflect the complexity and variability of constructions as we come across them in our work, while still striving to keep it as simple as possible, not least in order for it to be usable in LT applications.

An example constructicon entry, the reflexive resultative cx (cf. section 4), as represented in the current preliminary interface, is shown in figure 1. Like in other FrameNet based models, semantic roles and other cx elements are explicitly included in the definitions and annotated in the accompanying examples. The treatment of the roles themselves, however, is somewhat different. In FrameNet, semantic roles are locally defined for each frame, which has led to 125 different defi-

nitions of Agent, for example. Instead, we define roles etc. globally – generalizing where we can and maintaining specific roles where we must, but the same role label always has the same definition. Accordingly, cx elements are represented as sets of features, where each feature is a database entry of its own. In addition to the linguistic value of a consistent treatment of semantic roles, global role definitions will be helpful to LT applications (cf. Johansson 2012). The treatment of semantic roles is therefore an important subproject, based on the model in Rydstedt (2012).

As in the Berkeley constructicon, fixed cx elements are specifically indexed (*cee*, construction evoking element). In addition, the Swedish constructicon also lists collostructional elements, i.e., words that are not fixed, but significantly frequent in a certain cx (cf. Stefanowitsch and Gries 2003), *coll* in figure 1. Such information is useful for LT, and likely also for educational purposes. For reasons of time, this will not be based on full-fledged collostructional analyses; we will simply note salient common elements.

To enable cross-linguistic compatibility, much of the English terminology from FrameNet and the Berkeley constructicon is maintained. However, for constructicons to be cross-linguistically useful, additional information is required. In FrameNet, the frames serve as a technical lingua franca. Lexical units are language specific, but by assuming the same frames across languages, FrameNet resources for different languages may still be connected. In a constructicon, on the other hand, the central units are not frames but cx, and cx are typically language specific. Therefore, some form of common metalanguage is needed.

Initially, however, the constructicon is primarily designed for Swedish users. Hence, cx names and definitions are all in Swedish. This makes things easier for us, but the main reason is that the descriptions are eventually intended to be usable in an interface for non-linguists. An international representation may be added later on and should reasonably be developed in collaboration with the other constructicon projects under way. Awaiting that, we indicate the frame closest to the meaning of a certain cx, whenever applicable, as an approximation (cf. *evokes* in figure 1). A cx with causative meaning, such as the reflexive re-

sultative, may thus be associated with a frame like Causation_scenario.

The constructicon is usage-based, i.e., cx are identified and characterized according to authentic usage, as perceived from corpora. Such studies will chiefly be conducted using Korp, the main corpus tool of Språkbanken, where several corpora of different types are integrated and searchable by a common interface (Borin et al., 2012c). Korp gives access to around 1 billion running words (and growing), annotated for lexical unit, part of speech, morphosyntactic category, textual properties etc. This annotation is a vital feature for this project, since a cx may be defined by constraints on different levels: word, word-form, part of speech, morphosyntactic category, grammatical function, information structure etc. (as illustrated by the examples in the preceding paragraphs). The Korp interface can also present statistic information about grammatical and lexical contexts, as well as text type.

Up until now, we have mainly relied on linguistic methods for the identification of cx, but we will also develop tools to identify cx automatically. As a first step, we will explore methods for the identification of unknown cx, or rather cx candidates. For this purpose, StringNet (Wible and Tsao, 2010) is one of many possible research directions. StringNet identifies recurring n-gram patterns of two or more units, where every unit is classified on three levels – word form, lemma, and grammatical category – potentially revealing patterns of lexical units and form classes in combination. Narrowing down the search by combining the result of StringNet with methods for automatic morphology induction (Hammarström and Borin, 2011) and word segmentation (Hewlett and Cohen, 2011), should make it possible to identify likely cx candidates, which must then be judged manually, but the heuristic work process should be greatly simplified using these kinds of methods. Another possible research direction is the type of methods used to locate multiword expressions and terminology (see, e.g., Pecina 2010), which need to be further developed to cater for the identification of cx, where a position might have schematic content rather than being a specific word. For the latter, the morphosyntactic and syntactic information provided in the Korp anno-

tations will be used (cf. Baroni and Lenci 2010; Pitulainen 2010).

6 Outlook

Developing tools for automatic identification of cx is both a methodological approach and a highly relevant research objective in its own right. If we are able to automatically identify cx in authentic text, the ambiguity that has always plagued automatic syntactic analysis, where even relatively short sentences may have tenths or even hundreds of analyses, can be greatly reduced. Kokkinakis (2008) has shown that the identification of complex terminology and named entities simplifies a subsequent syntactic analysis considerably. Also, Attardi and Dell'Orletta (2008) and Gadde et al. (2010), and others, have shown how pre-identification of different types of local continuous syntactic units may improve a subsequent global dependency analysis. Our hypothesis is that cx can be used in the same way, and exploring this would be a valuable contribution to LT research. The cx primarily targeted in the project are largely language-specific, partly by virtue of containing lexical material. However, on a more abstract level, many of the classes of cx – and consequently the methods both for their discovery in corpora and for their use in LT applications – are expected to be cross-linguistically relevant. Hence our research on Swedish will be relevant to LT in general.

The constructicon is meant to be a large-scale, freely available electronic resource for linguistic purposes and language technology applications, in the first place. As already mentioned, it will be integrated in the SweFN network and, of course, benefit the network enriching it with cx. But the constructicon can also be regarded as a lexicographic resource per se, and of relevance for lexicography/lexicographers in general (cf. Hanks 2008). Cx have traditionally been neglected in dictionaries. Some cx can be found in the information given on valency, and many cx are indirectly presented in the usage examples (cf. Svensén 2009, 141ff., 188ff.). The coverage, however, is only partial, since the dictionaries tend to favor colorful fixed phrases at the expense of more anonymous cx with variable component slots. This is a problem, as many such

cx are arguably more relevant for language learners than, for example, idioms, which by comparison are used quite rarely (Farø and Lorentzen, 2009). Furthermore, paper dictionaries are inherently limited, partly due to their size, partly due to their structure; they are mainly based on headwords/lemmas. Electronic dictionaries, on the other hand, offer new opportunities through alternative search paths and (more or less) unlimited amount of space. Hence, in a longer perspective, the constructicon can be further developed and adapted as an extension to a future, general language e-dictionary of Swedish.

The improved coverage of constructional patterns provided by the constructicon should also be a valuable contribution to the fields of second language research and second language learning. As mentioned above, it is a special priority to account for cx that are problematic for second and foreign language acquisition. Besides such cx in particular, the constructicon in general should be highly relevant for L2 research and teaching. Usually, L2-learners do not acquire cx to any larger extent, except for the most general types. On the contrary, even advanced L2 learners have to rely on grammatical rules in their language production – in contrast to native speakers, who use prefabricated cx-templates extensively. This results in unidiomatic L2 production. It also adds a cognitive strain on the L2 speaker, since combinatorial language production is more taxing for the processing memory (Ekberg, 2004, 272), which makes L2 production more difficult than it needs to be. Adding the aspect of cx to L2 teaching situations would facilitate L2 learning for advanced students as well as for those who find traditional grammar an obstacle.

In summary, the constructicon is not only a desirable and natural development of the FrameNet tradition, it is also potentially useful in a number of areas, such as language technology, lexicography and (L2) acquisition research and teaching. In addition to these practical uses, we hope that this work will lead to theoretically valuable insights about the relation between grammar and lexicon.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project Swedish

Framenet++, VR dnr 2010-6013), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, by the European Commission through its support of the METANORD project under the ICT PSP Programme, grant agreement no 270899, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldberg, sponsored by the Knut and Alice Wallenberg Foundation.

References

- Giuseppe Attardi and Felice Dell'Orletta. 2008. Chunking and dependency parsing. In *LREC Workshop on Partial Parsing*, pages 27–32, Marrakech.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morris town, NJ, USA.
- Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36:673–721.
- Hans Boas and Ivan Sag, editors. to appear. *Sign-Based Construction Grammar*. CSLI, Stanford.
- Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech.
- Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.
- Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language technology for cultural heritage*, pages 41–61. Springer, Berlin.
- Lars Borin and Dimitrios Kokkinakis. 2010. Literary onomastics and language technology. In Willie van Peer, Sonia Zyngier, and Vander Viana, editors, *Literary education and digital learning. Methods and technologies for humanities studies*, pages 53–78. Information Science Reference, Hershey / New York.
- Lars Borin, Dana Danélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010a. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden.
- Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010b. Diabase: Towards a diachronic BLARK in support of historical studies. In *Proceedings of LREC 2010*, Valletta, Malta.
- Lars Borin, Markus Forsberg, and Christer Ahlberger. 2011. Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In *NODALIDA 2011 Conference Proceedings*, pages 58–65, Riga.
- Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, and Annika Kjellandsson. 2012a. Search result diversification methods to assist lexicographers. In *Proceedings of the 6th Linguistic Annotation Workshop*, pages 113–117.
- Lars Borin, Markus Forsberg, Leif-Jörän Olsson, and Jonatan Uppström. 2012b. The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, Istanbul.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012c. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, Istanbul.
- Lars Borin. 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, Krister Lindén, and Wanjiku Ng'ang'a, editors, *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, pages 53–65. Springer.
- Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.
- Jörg Didakowski, Lothar Lemnitzer, and Alexander Geyken. 2012. Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings of the 15th EURALEX International Congress*, pages 343–349, Department of Linguistics and Scandinavian Studies, University of Oslo.
- Lena Ekberg. 2004. Grammatik och lexikon i svenska som andraspråk på nästan infödd nivå [‘Grammar and lexicon in Swedish as a second language at almost native level of proficiency’]. In Kenneth Hyltenstam and Inger Lindberg, editors, *Svenska som andraspråk – i forskning, undervisning och samhälle*. Studentlitteratur, Lund.
- Ken Farø and Henrik Lorentzen. 2009. De oversete og mishandlede ordforbindelser – hvilke, hvor og hvorfor? [‘The overlooked and mistreated word combinations – which, where, and why?’]. *LexicoNordica*, 16:75–101.

- Charles Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomticity in grammatical constructions: the case of *let alone*. *Language*, 64:501–538.
- Charles Fillmore, Russell Lee Goldman, and Russell Rhodes. to appear. The FrameNet constructicon. In Hans Boas and Ivan Sag, editors, *Sign-Based Construction Grammar*. CSLI, Stanford.
- Charles Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co, Seoul.
- Charles Fillmore. 2008. Border conflicts: FrameNet meets Construction Grammar. In Elisenda Bernal and Janet DeCesaris, editors, *Proceedings of the XIII EURALEX International Congress*, pages 49–68, Barcelona.
- Mirjam Fried. to appear. Construction grammar. In A. Alexiadou and T. Kiss, editors, *Handbook of Syntax (2nd ed.)*. de Gruyter, Berlin.
- Phani Gadde, Karan Jindal, Samar Husain, Dipti Misra Sharma, and Rajeev Sangal. 2010. Improving data driven dependency parsing using clausal information. In *HLT: The 2010 Conference of the NAACL*, pages 657–660, Los Angeles.
- Adele Goldberg. 1995. *Constructions. A Construction Grammar approach to argument structure*. University of Chicago Press, Chicago & London.
- Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.
- Patrick Hanks. 2008. The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, 21:219–229.
- Karin Friberg Heppin. 2011. MedEval – a Swedish medical test collection with doctors and patients user groups. *Journal of Biomedical Semantics*.
- Daniel Hewlett and Paul Cohen. 2011. Word segmentation as general chunking. In *Proceedings of CoNLL 2011*, pages 39–47, Portland, Oregon.
- Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.
- Håkan Jansson. 2006. Har du ölat dig odödlig? En undersökning av resultativkonstruktioner i svenska [‘Have you aled yourself immortal? A study of resultative constructions in Swedish’]. Technical report, Dept. of Swedish, University of Gothenburg.
- Sofie Johansson Kokkinakis and Ulrika Magnusson. 2011. Computer based quantitative methods applied to first and second language student writing. In Roger Källström and Inger Lindberg, editors, *Young urban Swedish. Variation and change in multilingual settings*, pages 105–124. University of Gothenburg, Dept. of Swedish.
- Richard Johansson. 2012. Non-atomic classification to improve a semantic role labeler for a low-resource language. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 95–99, Montréal.
- Lauri Karttunen, Jean-Pierre Chanod, Gregory Grefenstette, and Anne Schiller. 1996. Regular expressions for language engineering. *Natural Language Engineering*, 2(4):305–328.
- Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pages 425–432, l’Institut Universitari de Lingüística Aplicada (IULA) dela Universitat Pompeu Fabra.
- Dimitrios Kokkinakis. 2008. Semantic pre-processing for complexity reduction in parsing medical texts. In *Proceedings of the 21th Conference on the European Federation for Medical Informatics (MIE 2008)*.
- Dimitrios Kokkinakis. 2012. The Journal of the Swedish Medical Association – a corpus resource for biomedical text mining in Swedish. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop*, Istanbul.
- Ronald Langacker. 1991. *Foundations of Cognitive Grammar. Volume II: Descriptive application*. Stanford University Press, Stanford, CA.
- Benjamin Lyngfelt and Markus Forsberg. 2012. Ett svenskt konstruktikon. Utgångspunkter och preliminära ramar [‘A Swedish constructicon. Points of departure and preliminary guidelines’]. Technical report, Dept. of Swedish, University of Gothenburg.
- Benjamin Lyngfelt. 2007. Mellan polerna. Reflexiv- och deponenskonstruktioner i svenska [‘Between the poles. Reflexive and deponent constructions in Swedish’]. *Språk och stil NF*, 17:86–134.
- Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. In *Proceedings of LaTECH 2012*.
- Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.
- Jussi Piitulainen. 2010. *Explorations in the distributional and semantic similarity of words*. University of Helsinki.
- Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago & London.
- Julia Prentice and Emma Sköldberg. 2011. Figurative word combinations in texts written by adolescents in multilingual school environments. In Roger

- Källström and Inger Lindberg, editors, *Young urban Swedish. Variation and change in multilingual settings*, pages 195–217. University of Gothenburg, Dept. of Swedish.
- Julia Prentice. 2011. “Jag är född på andra november”. Konventionaliseraade tidsuttryck som konstruktioner – ur ett andraspråksperspektiv. [‘I was born on second November. Conventionalized time expressions as constructions – from a second language perspective’]. Course paper on Construction Grammar, Dept. of Swedish, Univ. of Gothenburg.
- Taraka Rama and Lars Borin. 2011. Estimating language relationships from a parallel corpus. A study of the Europarl corpus. In *NODALIDA 2011 Conference Proceedings*, pages 161–167, Riga.
- Josef Ruppenhofer, Michael Ellsworth, R. L. Miriam Petrucc, R. Christopher Johnson, and Jan Scheffczy. 2010. *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.
- Rudolf Rydstedt. 2012. *En matchningsdriven semantisk modell. Mellan ordboken och den interna grammatiken*. [‘A match-driven semantic model. Between the dictionary and the internal grammar’]. University of Gothenburg, Dept. of Swedish.
- Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multi-word expressions: A pain in the neck for NLP. In *Proceedings of CICLING-2002*.
- Ivan Sag. 2010. English filler-gap constructions. *Language*, 86:486–545.
- Anju Saxena and Lars Borin. 2011. Dialect classification in the Himalayas: a computational approach. In *NODALIDA 2011 Conference Proceedings*, pages 307–310, Riga.
- Anatol Stefanowitsch and Stefan Gries. 2003. Collocations: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8:209–243.
- Bo Svensén. 2009. *A handbook of lexicography. The theory and practice of dictionary-making*. Cambridge University Press, Cambridge.
- Elena Volodina and Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of LREC 2012*, Istanbul.
- David Wible and Nai-Lung Tsao. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 25–31, Los Angeles.

A Distributional Memory for German

Sebastian Padó

Institut für Computerlinguistik

Universität Heidelberg

pado@cl.uni-heidelberg.de

Jason Utt

Institut für maschinelle Sprachverarbeitung

Universität Stuttgart

uttjn@ims.uni-stuttgart.de

Abstract

This paper describes the creation of a Distributional Memory (Baroni and Lenci 2010) resource for German. Distributional Memory is a generalized distributional resource for lexical semantics that does not have to commit to a particular vector space at the time of creation. We induce a resource from a German corpus, following the original design decisions as closely as possible, and discuss the steps necessary for a new language. We evaluate the German DM model on a synonym selection task, finding that it can compete with existing models.

1 Introduction

Distributional semantics is a paradigm for determining a word’s meaning by observing its occurrence in large corpora. It builds on the Distributional Hypothesis (Harris, 1954; Miller and Charles, 1991) which states that words occurring in similar contexts are similar in meaning. Vector space models, the most widely used incarnation of distributional semantics, represent words as vectors in a high-dimensional space whose dimensions correspond to features of the words’ contexts (Turney and Pantel, 2010). These features can be chosen in different ways; popular choices are context words (Schütze, 1992), dependency relations or paths (Lin, 1998; Padó and Lapata, 2007), or subcategorization frames (Schulte im Walde, 2006).

The notion of graded semantic similarity that vector space models provide has been used in various applications, including word sense disambiguation (McCarthy et al., 2004), representation of selectional preferences (Erk et al., 2010), verb class induction (Schulte im Walde, 2006), ana-

logical reasoning (Turney, 2006), or alternation discovery (Joanis et al., 2006).

Such different applications however tend to require different types of semantic spaces. In traditional cases like WSD, the objects whose similarity we are interested in are words. In analogical reasoning, however, we need to compare word pairs, and for alternation discovery we need to compare verbal argument positions. Baroni and Lenci (2010) addressed this fragmentation by proposing a model called Distributional Memory (DM). It captures distributional information at a more abstract level and can be mapped onto various vector spaces to perform different tasks. Baroni and Lenci show that DM is competitive with state-of-the-art models with dedicated spaces on many NLP tasks.

Baroni and Lenci’s original work only considered English. In this paper, we describe ongoing work on the construction and evaluation of a corresponding Distributional Memory resource for German. Section 2 provides background on Distributional Memory. Section 3 assesses potential strategies for inducing a German DM-style resource. Section 4 describes the concrete steps, and discusses properties of German that require particular attention. Section 5 evaluates the German DM on a synonym selection task, and Section 6 concludes.

2 Distributional Memory

Distributional Memory (DM, Baroni and Lenci 2010) is a recent multi-purpose framework in distributional semantics. Contrary to traditional studies, which directly constructed task-specific vector spaces, DM extracts a tensor, i.e. a three-dimensional matrix, of weighted *word-link-word* tuples. Each tuple is mapped onto a number by a

scoring function $\sigma: W \times L \times W \rightarrow \mathbb{R}^+$. For example, $\langle \text{pencil} \text{ obj} \text{ use} \rangle$ is assigned a higher weight than $\langle \text{elephant} \text{ obj} \text{ use} \rangle$.

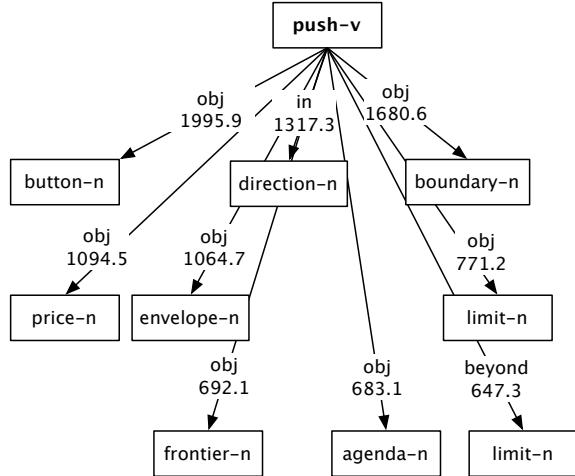


Figure 1: The DepDM tensor as labeled graph: *push* with its nine highest-scoring co-occurrences

This DM tensor can be visualized as a directed graph whose nodes are labeled with lemmas and whose edges are labeled with links and scores. As an example, Figure 1 shows the nine highest-scoring context co-occurrences for the verb *push* together with their scores. All of them happen to be arguments and adjuncts of *push*, although DM also models ‘inverse’ relations. Seven of the nine co-occurrences are objects, and two are prepositional adjuncts. The benefit of this tensor representation is that it is applicable to many tasks in computational linguistics. Once a task is selected, a dedicated semantic space for this task can be generated efficiently from the tensor by matricization. For example, the *word by link-word* space ($W \times LW$) contains vectors for words w and its dimensions are labeled with pairs $\langle l, w \rangle$ of a link and a context word. This space models similarity among words, e.g. for thesaurus construction (Lin, 1998). Another space is the *word-word by link* space ($WW \times L$). It contains co-occurrence vectors for word pairs $\langle w_1, w_2 \rangle$. Its dimensions are labeled with links l . This space can be used to model semantic relations.

DM does not assume any specific source for the tuples. However, since dependency parses are the most obvious source of such relational information, DM falls into the category of dependency-

based semantic space models. DM does not presuppose a particular vocabulary of link (relation) types. Baroni and Lenci present three variants of DM that differ in the relations that they assume.

The simplest variant is Dependency DM (DepDM), as shown in Figure 1. It uses a set of relations among verbs, nouns, and adjectives most of which are adopted directly from the dependency parses. (*obj*, *iobj*). The subject relation is extended with subcategorization information (*subj_tr*, *subj_intr*) to better handle diathesis alternations, and prepositions are turned directly into links ($\langle \text{walk} \text{ on} \text{ street} \rangle$).

The second variant, Lexicalized DM (LexDM), uses more complex links. The links encode rich information about the two words in the triple, such as their part of speech, definiteness, modification, as well as lexicalized information about the dependency path between the words. An example is the tuple $\langle \text{soldier} \text{ use}+\text{n}+\text{the}+\text{n-a} \text{ gun} \rangle$ which is constructed from the sentence *The soldier used a gun* and indicates the verb linking the two nouns, their POS tags and articles. The definition of the LexDM links is highly language-specific and is based on WordNet information as well as semantic classes extracted from corpora with the help of various patterns.

Finally, the third variant, TypeDM, builds on LexDM but modified the scoring function, following the intuition that the frequency of a link is less informative (since it is influenced by a large number of factors) than the number of its surface varieties: links with a high number of varieties are likely to express prominent semantic relations.

3 Strategies for Creating a German DM

In situations like ours, where some resource is available in language A, but not in another language B, two strategies can be followed. The first one is *cross-lingual transfer*: The existing resource in language A is ported onto language B. The second one is *parallel induction*: The schema for creating the resource in language A is replicated, as closely as possible, for language B.

Cross-lingual transfer is an attractive proposition since it takes optimal advantage of the existing resource. It falls into two subcases: If the resource mainly contains information at the token (instance) level, this can be achieved via annota-

tion projection in parallel corpora (Yarowsky and Ngai, 2001; Bentivogli and Pianta, 2005). If the resource concentrates on the type (lemma) level, bilingual dictionaries can be used to simply translate the resource (Fung and Chen, 2004; Peirsman and Padó, 2011).

The translation-based strategy would be applicable in the case of DM which records and scores lemma-link-lemma triples. For the language pair English–German, it is also the case that reliable, high-coverage bilingual dictionaries are publicly available. However, we decided against adopting this strategy. The reason is that simple translation runs into serious ambiguity problems. The problem can be visualized easily in terms of the labeled graph view on DM (Figure 1). Translating this graph involves replacing the original node labels, English lemmas, with German lemmas.¹ This is unproblematic for one-to-one translations where nodes are just relabeled (*sitzen* → *sit*). There is also a solution for cases where two English lemmas correspond to a single German lemma (*cup*, *mug* → *Tasse*): the two English nodes can be collapsed. However, a significant number of English words have more than one translation in German. Where these translations are not just synonyms but express different senses of the English word (*wood* → *Holz*, *Wald*), the English node would need to be split and all its incoming and outgoing edges assigned to either of the two new German nodes. This, however, is a full-fledged sense disambiguation problem which appears difficult to solve, in particular given the very limited amount of context information available in bilingual dictionaries.

For this reason, we decided to adopt the second strategy, parallel induction, and create a DM resource from German text. In the case of distributional semantic models, this choice is made possible by the fact of the relatively good corpus situation for German: there is a very large web corpus available for German, as well as fairly good dependency parsers. The remainder of this section discusses the different steps involved in this task and the difficulties that arise.

¹We assume for this discussion that syntactic relations show a high degree of parallelism for the language pair English–German, which we found to be a reasonable assumption in previous experiments (Peirsman and Padó, 2011).

4 Inducing a DM Resource from German corpora

4.1 DepDM vs. LexDM vs. TypeDM

Our first decision was which of the three DM variants to implement for German. We observed above that the patterns in LexDM and TypeDM are more complex than those in DepDM. A number of (semi-)manual annotations were utilized to construct the former, such as a list of high-frequency verbs selected as part of the lexicalized edge labels. While the more elaborately designed TypeDM nearly always performs best (Baroni and Lenci, 2010), the much simpler DepDM performs at a similar level for many tasks. We therefore opted to implement DepDM.

4.2 Defining patterns to extract links

The types of German links we use correspond fairly directly to the simple syntactic patterns of DepDM. We obtained these patterns by inspecting the most frequent syntactic configurations in a large German corpus (cf. Section 4.3). The patterns can be categorized into two groups: unlexicalized and lexicalized patterns.

Unlexicalized patterns. We use 7 patterns which extract information at the verb phrase and sentence levels. subjects for transitive and intransitive verbs (*subj_tr*, *subj_intr*); direct objects, indirect objects, and phrasal complements of verbs (*obj*, *iobj*, *vcomp*); noun modification (*nmod*) and the relation between a subject and an object of a verb (*verb*). These patterns are unlexicalized, that is, the patterns correspond directly to link types.

Lexicalized patterns. Three more patterns are lexicalized. This means two things: (a), they may contain fixed lexical material (marked in boldface); (b), they may incorporate some of the lexical information on the dependency paths that they match into the resulting link (marked in square brackets below). These patterns apply mostly to NPs and PPs. The first pattern is $n_1 [prep] n_2$, which captures phrases like *Recht auf Auskunft* which is turned into the triple $\langle \text{Recht} \text{ auf } \text{Auskunft} \rangle$. The second pattern is $adj n_1 \text{ von } [n_2]$ which extracts the second linking noun as a link and captures such phrases as *heutige Größe von der Sonne* which

results in the triple $\langle \text{heutige Sonne GröÙe} \rangle$. The third lexicalized pattern is $n_1 [\text{verb}] n_2$ where the particular verb combining subject n_1 and object n_2 is used as the link. For example, the sentence *Hochtief sieht Aufwind* results in the tuple $\langle \text{Hochtief sehen Aufwind} \rangle$. Due to the presence of lexical material in the links, these patterns give rise to a very large number of link types.

Note that we currently ignore prepositional arguments of verbs, adverbs, and relative clauses. This is partly a frequency-based decision (the patterns above account for the majority of links in the corpus), but also one motivated by minimalism: we wanted to find out how well such a minimal set of patterns is able to perform.

We also found that as opposed to English, German offers a number of difficulties when extracting word relations from text. Particle verbs for example often possess a detachable prefix e.g. *mit|geben*, *weiter|reichen*, which at the surface level can be realized at a large distance from the verb stem, e.g. *Er gab ihr das Buch, nach dem sie am Vortag gefragt hatte, nur ungern mit*. We reconstruct such verbs from the parser output by looking for words with the PTKVZ (verbal particle) POS tag which stand in a SVP (verbal particle) or MO (modifier) relation to a full verb.

Additional issues are the very productive compounding (e.g. *Wasserstoffbetankungseinrichtung*) and derivation (e.g. *Pappkärtchen*) of nouns. This gives rise to many more noun types that need to be integrated into the system than in English. In the design of the English DM, Baroni and Lenci set a limit for nouns, including only the 20k most frequency nouns into their DM. In German this would give too sparse a model (cf. Section 5). We thus chose to not use a cutoff for any part of speech even though this makes the tensor and its resulting matrices even sparser than usual in dependency-based representations. We considered splitting compound nouns, but left this problem for future work, given the problem of deciding whether compounds are semantically transparent or not (cf. *Schule – Baumschule*).

4.3 Step 2: Corpus counts

The corpus we used to extract the co-occurrence counts was the SDEWAC web corpus (Faaß et al.,

2010) parsed with the MATE German dependency parser (Bohnet, 2010). SDEWAC is based on the DEWAC corpus (Baroni and Kilgarriff, 2006), a large corpus collected from the German WWW (.de top-level domain). SDEWAC performs various preprocessing steps on DEWAC such as identifying sentences and then filtering out duplicate sentences from the same URL. It contains 9M different word types and 884M word tokens.

The advantage of using a web corpus is size and availability. At the same time, it includes text from many different domains, in contrast to most other corpora which contain mostly news text. On the other hand, it contains much noise such as spelling and grammatical errors that part-of-speech taggers and parsers must deal with. At the current stage, we are more interested in preserving the information in the data. While this means we include nouns such as *Kurz-vor-Ladenschluss-noch-schnell-einkaufenden-Kassen-Warteschlange*, we intend to address these issues in future work.

4.4 Step 3: Scoring

The weighting in our German DM follows the scheme for LexDM and DepDM, which is Local Mutual Information (LMI), Evert (2005)). The following equation defines the LMI of a word-link-word combination (i, j, k) via its observed frequency in the corpus O_{ijk} and its expected frequency E_{ijk} :

$$\text{LMI}(i, j, k) = O_{ijk} \cdot \log \frac{O_{ijk}}{E_{ijk}}.$$

The logarithmic term stems from the definition of pointwise mutual information: It captures the degree to which the observed frequency of the word-link-word combination differs from the expected frequency under an independence assumption, i.e. $P(i, j, k) = P(i) \cdot P(j) \cdot P(k)$. This means E_{ijk} assumes that (i, j, k) occurs due to chance rather than a latent relatedness. Then, if $O_{ijk} = E_{ijk}$, the mutual information will be equal to 0, since i , j , and k are statistically independent of one another.

In contrast to PMI, though, LMI contains the raw frequency O_{ijk} as a factor that discounts infrequent combinations, addressing PMI's well-known tendency to overestimate the informativeness of very infrequent combinations. We follow the construction of the English DM and set negative Lo-

calMI values to zero, which is equivalent to excluding them from the tensor.

5 Evaluation

5.1 Statistics of German DM

The German DM resulting from these steps contains over 78M links for 3,5M words (nouns, verbs and adjectives). It contains about 220K link types, almost all of which stem from surface patterns. On average, a lemma has 22 links. This makes our German DM considerably more sparse than its English counterpart, which contains about 131M links for just 31K lemmas and has just over 25K link types.

Table 1 shows information about the link types. As described in Section 4.2, there are 7 unlexicalized link types and more than 220K lexicalized link types. Table 2 shows an example of the extracted co-occurrences for the verb *sehen*.

The current version of our German DM (DM.de) can be obtained at <http://goo.gl/GNbMb>.

5.2 Task-based evaluation: Synonym selection

As we have discussed above, the main benefit of the DM model is its ability to inform various tasks in lexical semantics. Baroni and Lenci (2010) evaluate their English DM on a number of tasks, including semantic similarity, relational similarity, and relation extraction. A comprehensive evaluation is outside the scope of this article; we focus on synonym selection.

Data. Our evaluation is based on the German Reader’s Digest Word Power (RDWP) dataset (Wallace and Wallace, 2005) which contains 984 word choice problems.² This dataset is similar to the English TOEFL dataset (Landauer and Dumais, 1997). Each problem consists of one target word and a set of four possible candidates, which are either a synonym or a phrase defining the word. For example, for the word *Prozedur*, the candidates are *Gerichtsverfahren*, *Vorbild*, *Vorgang*, and *Demonstration*.

Model. As this is a word similarity task, we base our model on the word-by-link-word $W \times LW$

²The dataset is available from: <http://goo.gl/PN42E>

matricization of the DM tensor which represents words in a space whose dimensions are labeled with pairs of a link and a context word. For each problem, we compute the cosine between its vector and the candidates’ vectors, predicting the most similar candidate.

Our base model (‘base’) excludes problems that include phrases as candidates, since “plain” DM focuses on representing the meaning of individual words. Given the high number of phrasal candidates in the dataset, however, we also experiment with a very simplistic model for phrase meaning (‘phrases’) which defines the similarity between a target word and a phrasal candidate as the maximum similarity between the target and any of the constituent words of the phrase. This corresponds to the oversimplification that the meaning of a phrase is determined by its most informative word. Finally, we combine the two models (‘combined’), which is trivial since they make predictions for disjoint sets of problems.

We compare two versions of our DM model that differ in the amount of sparsity that they tolerate. The first model ($DM.de^\alpha$) excludes items for which at least one candidate has a zero similarity to the target (i.e. there is not one single context in which both words were observed). This model is conservative and abstains from any experiments items where sparsity problems can be expected. The second model, $DM.de^\beta$, excludes only those items where all candidates have a zero similarity to the target; this generally indicates that the target was seen never or almost never.

Evaluation. We evaluate analogously to Mohammad et al. (2007), defining the Score as a weighted sum of correct predictions, while discounting ties: $Score = A + .5 \cdot B + .3 \cdot C + .25 \cdot D$, where A is the number of correctly predicted items with no ties, B those with a 2-way tie, C with a 3-way tie, and D with a 4-way tie. Note that Score does not take the number of covered problems into account. For this reason, Mohammad et al. also define a precision-oriented Accuracy measure which is defined as the average Score per covered problem: $Accuracy = Score / Covered$. Note that a random baseline would perform at an accuracy of .25.

We compare our results with those reported by

Unlexicalized links		
7,833,635	$n \ (i)\text{obj} \ v$	noun n is (in)direct object of verb v
7,478,550	$n \ \text{subj_}(in)\text{tr} \ v$	noun n is subject of verb v in an (in)transitive construction
677,397	$v_1 \ \text{vcomp} \ v_2$	subcategorization of verb v_1 by a verb v_2 (excluding modals and auxiliaries)
1,575,516	$x \ \text{nmod} \ n_2$	noun n_2 modified by $x \in \{\text{adj}, \text{n}\}$
3,304,045	$n_1 \ \text{verb} \ n_2$	noun n_1 is subject and n_2 object of a verb
Lexicalized links		
220,269	link types	e.g. in, von, an, als, vor, um, gegen, stellen, machen, bieten, geben
18,180,604	links	2,462,927 $n_1 \ \text{in} \ n_2$; 1,424,398 $n_1 \ \text{von} \ n_2$; 1,170,609 $n_1 \ \text{mit} \ n_2$; ...

Table 1: Statistics for the German DM tensor.

	sbj_intr	sbj_tr	obj	iobj
<i>Realität</i>	2,339.3	<i>Entwurf</i>	4,455.5	<i>Chance</i>
<i>Wirklichkeit</i>	1,343.0	<i>Mensch</i>	3,555.2	<i>Film</i>
<i>Sache</i>	1,204.1	<i>Senat</i>	3,244.1	<i>Bild</i>
<i>Welt</i>	802.4	<i>Zuschauer</i>	3,147.2	<i>Möglichkeit</i>
				<i>Zukunft</i>
				324.8

Table 2: Highest-LMI co-occurrences for the verb *sehen*. The intransitive subjects and indirect objects are mainly due to misparses (*‘die X sieht ... aus’*) and idiomatic expressions, respectively (*‘der Y ins Gesicht sehen’*).

Mohammad et al. (2007) who evaluate a number of monolingual models based on GermaNet, a large lexical resource for German. They fall into two categories: (a) gloss-based methods which are variants of the Lesk (1986) algorithm applied to GermaNet glosses³; and (b), hierarchical methods which compute similarity with various graph-based similarity measures in GermaNet. In contrast, our DM-based model is purely distributional and does not use any structured semantic knowledge for German.

Results. Table 3 shows the results. In terms of accuracy, Mohammad et al.’s gloss-based models are the clear winners: the very specific “context” provided by a definition provides an excellent basis for synonym detection where it is available; but these models at the same time have a low coverage of below 30%. Hierarchical GermaNet models

show a higher coverage (up to 40%), but also a substantially lower precision and accuracies of around .5.

Our ‘base’ models already perform at the same level as the hierarchical models, with accuracies of .5 and above. The ‘combined’ models (both DM.de^α and DM.de^β) show the highest coverage of all models considered: the more cautious DM.de^α covers more than 40%, and the DM.de^β even over 80% of all items. This result underscores the benefit of distributional models in terms of coverage.

Somewhat surprisingly, the ‘phrases’ model reliably outperforms the ‘base’ model even though it uses a simplistic heuristic. The reason can be found in the properties of the phrasal candidates. As Table 4 shows, they often resemble definitional phrases or glosses. Consequently, our model basically attempts a simplified Lesk-style disambiguation task, similar to Mohammad et al.’s gloss-based models. In contrast, the single-word candidates treated by the ‘base’ model are often rare or otherwise difficult synonyms of the target. For instance,

³The Lesk algorithm is a word sense disambiguation method which uses dictionary glosses for a word’s different senses and checks their overlap with the given context. The sense whose gloss has the highest overlap is taken to be the correct sense.

Measure	Covered	Correct	Ties	Score	Accuracy
<i>gloss-based</i> (Mohammad et al. 2007)					
HPG	222	174	11	171.5	.77
RPG	266	188	15	184.7	.69
<i>distributional</i> (our work)					
DM.de ^α (base)	174	96	0	96	.55
DM.de ^α (phrases)	228	135	4	132	.58
DM.de ^α (combined)	402	231	4	228	.57
DM.de ^β (base)	358	178	0	178	.50
DM.de ^β (phrases)	466	261	4	258	.55
DM.de ^β (combined)	824	439	4	436	.53
<i>hierarchical</i> (Mohammad et al. 2007)					
JC	357	157	1	156.0	.44
Lin _{GN}	298	153	1	152.5	.51
Res	299	154	33	148.3	.50

Table 3: Performance on the Reader’s Digest Word Power data set, DM.de^α includes only items for which all candidates have a non-zero similarity, while DM.de^β only requires one similarity be non-zero.

Golfwägelchen is so uncommon that we get no similarity to the target *Caddie*, while *Golfschläger* is frequent enough that the shared topic leads to a certain degree of similarity. On the other hand, the common word *Witz*, while not an exact synonym, would be used in defining the target word *Kalauer*. The difference between the ‘base’ and ‘phrases’ models remains fairly small, though.

The non-zero ties are all phrasal items that include the same phrase with subtle variations, e.g. *Sommertag* as a target has the candidate phrases: *[Höchsttemperatur] von mindestens 25/20/28/30°C*. This clearly shows the limits of our simple phrasal comparison method which fails to distinguish between candidates when the most similar word is shared.

The ‘combined’ models, which combine all predictions of the ‘base’ and ‘phrases’ models, end up with the partial models’ averaged accuracies (.53 and .57 for DM.de^α and DM.de^β, respectively). They outperform Mohammad et al.’s hierarchical models but not the gloss-based models. They show however the two highest score numbers, 228 and 436, which is a direct result of their high coverage. We see these results as encouraging and promising for an unoptimized and sparse representation like our current DepDM.de.

word	candidate	similarity
	<i>Golfwägelchen</i>	0
target:	<i>Golfschläger</i>	.008
<i>Caddie</i>	<i>Golfplatz</i>	0
	<i>Golf-Abschlag</i>	0
	<i>billiger [Witz]</i>	.146
target:	<i>leichte [Kutsche]</i>	.055
<i>Kalauer</i>	<i>steifer [Hut]</i>	.036
	<i>[Merkspruch]</i>	0

Table 4: Examples of ‘base’ items and ‘phrase’ items. Words in brackets are those with the highest similarity to the target.

The difference in accuracies between DM.de^α and DM.de^β indicates that requiring all candidates to have non-zero similarities to the target, as DM.de^α does, does indeed reduce the number of sparsity-related problems. However, the modest increase in accuracy comes with a massive loss in coverage of roughly 50%. Thus, even the large SDEWAC appears not to be large enough to provide sufficiently rich representations for many infrequent targets and candidates.

Finally, we investigated the nature of the 160 items which were not covered by any of the DM.de models. We discovered that the majority of them are adjectives, which we attribute to the fewer number of link types associated with adjectives as opposed to nouns. In the current version of DM.de, adjectives are only “counted” when they occur in nmod contexts, i.e., when they are in an attributive construction. In the next iteration of DM.de, we will include predicative uses as well.

6 Conclusion

This paper has reported on ongoing work to create a German Distributional Memory resource. Using a fairly simple set of patterns and collecting counts from a German web corpus, we have been able to perform competitively on a German synonym selection task. Our work can provide a blueprint for the induction of similar resources in other languages and indicates that interesting results can be achieved with relatively little effort.

Nevertheless, we are aware of a number of shortcomings in the model. Some of them relate to preprocessing. In our web corpus, tagging errors are a frequent source of problems. The tagger has difficulties with tokens it was not trained on, e.g. ‘*I*’ is tagged either as a verb and an adjective. One remedy would be to filter out likely tagging errors with simple heuristics such as excluding nouns that are lowercase. Misparses of NP-PP sequences form another problem: in ‘*während lange Zeit von Säkularisierung gesprochen wurde*’, the *von*-PP belongs to the phrasal verb *von etwas sprechen* but it is matched by the adjective - noun (*of*) - noun pattern.

Another general problem that we have mentioned repeatedly is sparsity. Our German DM is considerably more sparse than the English DM. We plan to extend the pattern definitions to other

syntactic constructions and will consider more sophisticated ways of computing similarity that take advantage of DM’s graph structure, such as graph walks (Toutanova et al., 2004).

Finally, we want to revisit our original decision not to use a translation-based approach and plan to combine monolingual and cross-lingual evidence (Naseem et al., 2009) to bootstrap a more reliable DM for new languages. This will, however, require formulating a probabilistic framework for the induction task to weigh the relative evidence from both languages.

Acknowledgments. This work was partially supported by the EC-funded project EXCITEMENT (FP7 ICT-287923) and the German Research Foundation (SFB 732).

References

- Marco Baroni and Adam Kilgarriff. 2006. Large Linguistically-Processed Web Corpora for Multiple Languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 87–90.
- Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1–49.
- Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Journal of Natural Language Engineering*, 11(3):247–261.
- Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.
- Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763, December.
- Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, IMS, Universität Stuttgart.
- Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR as an Example of Morphological Evaluation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta.
- Pascale Fung and Benfeng Chen. 2004. BiFrameNet: Bilingual Frame Semantics Resources Construction

- by Cross-Lingual Induction. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 931–935, Geneva, Switzerland.
- Zelig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.
- Eric Joanis, Suzanne Stevenson, and David James. 2006. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.
- Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.
- Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation, SIGDOC ’86*, pages 24–26, New York, NY, USA. ACM.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL*, pages 768–774, Montreal, QC.
- Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 279–286, Barcelona, Spain.
- George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.
- Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-Lingual Distributional Profiles of Concepts for Measuring Semantic Distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 571–580, Prague, Czech Republic.
- Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual Part-of-Speech Tagging : Two Unsupervised Approaches. *Journal of Artificial Intelligence Research*, 36:1–45.
- Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.
- Yves Peirsman and Sebastian Padó. 2011. Semantic relations in bilingual vector spaces. *ACM Transactions in Speech and Language Processing*, 8(2):3:1–3:21.
- Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.
- Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing ’92*, pages 787–796, Minneapolis, MN.
- Kristina Toutanova, Christopher D. Manning, and Andrew Y. Ng. 2004. Learning random walk models for inducing word dependency distributions. In *Proceedings of ICML*, Banff, BC. ACM Press.
- Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
- Peter Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.
- DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader’s Digest, das Beste für Deutschland*. Verlag Das Beste, Stuttgart.
- David Yarowsky and Grace Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 200–207, Pittsburgh, PA.

Towards high-accuracy bilingual phrase acquisition from parallel corpora

Lionel Nicolas and Egon W. Stemle and Klara Kranebitter

{lionel.nicolas, egon.stemle, klara.kranebitter}@eurac.edu

Institute for Specialised Communication and Multilingualism

European Academy of Bozen/Bolzano

Abstract

We report on on-going work to derive translations of phrases from parallel corpora. We describe an unsupervised and knowledge-free greedy-style process relying on innovative strategies for choosing and discarding candidate translations. This process manages to acquire multiple translations combining phrases of equal or different sizes. The preliminary evaluation performed confirms both its potential and its interest.

1 Introduction

This paper reports on work in progress that aims at acquiring translations of phrases from parallel corpora in an unsupervised and knowledge-free fashion. The process described has two important features. First, it can acquire multiple translations for each phrase. Second, no restrictions is set on the size of the phrases covered by the translations, phrases can be of equal or different sizes. The process is a greedy-style one: it constructs a set of candidate translations and iteratively selects one and discards others. The iteration stops when no more candidate translations remain.

The main contributions of this paper are:

- a metric that evaluates a candidate translation by taking into account the likeliness in frequency of two phrases in a candidate translation (see sect. 6.3),
- a metric that evaluates a candidate translation by taking into account the number of occurrences of a candidate translation

and the significance of each occurrence (see sect. 6.4),

- an approach that discards candidate translations by enforcing coherence with the ones validated in previous iterations.

This paper is organized as follows. In section 2 we introduce the terminology we use in this paper whereas in section 3 we describe the state of the art. Section 4 briefly introduces the research subjects for which the generated data is relevant for. Section 5 explains in an abstract fashion the ideas which implementations are later detailed in section 6. In section 7, we present and discuss a preliminary evaluation. Finally in section 8 and 9, we highlight possible future works and conclude.

2 Definitions

A *bitext* is composed of both source- and target-language versions of a sequences of tokens. The sequences are usually sentences, paragraph or documents. A *phrase* is a sequence of tokens. A *translation* is said to cover two phrases from two different languages when they are translation of one another. The *size of a phrase* corresponds to the number of tokens it contains. The *size of a translation* th is designated by $size(th)$ and corresponds to the sum of the sizes of the phrases it covers. A phrase *includes* another one if it includes all the tokens of the included phrase. A translation includes another one if the phrases it covers include the phrases covered by the included translation. An occurrence of a phrase in a bitext is called a *slot*. The value $slots(ph, b_n)$ is the number of slots of an phrase ph in a bitext b_n .

A candidate translation ct covering two phrases ph_i and ph_j is said to claim slots in a bitext b_n when $slots(ph_i, b_n) \neq 0$ and $slots(ph_j, b_n) \neq 0$. The number of slots claimed by ct is designated by the value $claims(ct, b_n)$ and is initially set to $\min(slots(ph_i, b_n), slots(ph_j, b_n))$. A candidate translation ct is said to occur in a bitext b_n when $claims(ct, b_n) \neq 0$. Slots of a phrase ph_i in a bitext b_n are said to be *locked* when they cannot be claimed any more by any candidate translations. The number of locked slots of a phrase ph_i is designated by $locks(ph_i, b_n)$ and is initially set to 0.

3 Previous Work

The closest related work with fairly equivalent objectives, we found so far is the one of Lavecchia et al. (2008) where mutual information is used to extract translations of small phrases which quality is evaluated through the performance of a machine translation tool.

In a more indirect fashion, the method presented here can be related to phrase alignment and bilingual lexicon extraction.

Phrase alignment, a key aspect to improve machine translation tool performances, is for most methods such as Koehn et al. (2003), Zhang and Vogel (2005) or Deng and Byrne (2008) the task of acquiring a higher level of alignment from a corpus originally aligned on the word level. Even though it can allow to perform phrase translation extraction in a later stage, the two subjects are similar but not equivalent in the sense since input data and objectives are different. The evaluation protocol usually involves studying the performances of a machine translation tool such as Moses (Koehn et al., 2007) taking as input data the alignment.

Because word forms are the smallest type of phrases, the work presented is related to bilingual lexicon extraction. Many early approaches for deriving such lexicon from parallel corpora use association measures and thresholds (Gale and Church, 1991; Melamed, 1995; Wu and Xia, 1994). The association measures are meant to rank candidate translations and the threshold allow to decide which ones are kept or discarded. Although most association measures focus on recurrent occurrences of a candidate translation, other methods like Sahlgren and Karlsgren (2005) and Wid-

dows et al. (2002) use semantic similarity.

As it has been later explained later in Melamed (1997) and Melamed (2000), such strategy keeps many incorrect translations because of indirect associations, i.e., pairs of phrases that often co-occur in bitexts but are not mutual translations. Nevertheless, since translations tend to be naturally more recurrent than the indirect associations, the counter-measure is generally to discard in a bitext a candidate translation if it covers a phrase covered by another one with a greater score (Moore, 2001; Melamed, 1997; Melamed, 2000; Tsuji and Kageura, 2004; Tufis, 2002).

In Tsuji and Kageura (2004), an extension of the method described in Melamed (2000) has been designed to cope with translations with very few occurrences.

4 Applicability

The extracted phrase translations can be used by tools or resources that deal directly or indirectly with translations.

The most direct application is to use such data as input for machine translation or memory-based translation systems. Another interesting use would be to exploit the phrase translations to directly perform phrase alignment.

Because word forms are the smallest type of phrases, the data can also be adapted and used for subjects that take advantage of bilingual lexicons. For example, the acquired translations could be used for extending multilingual resources such as Wordnets or help perform word sense disambiguation (Ng et al., 2003).

Because the process can cope with multiple translations, many homonyms or synonyms/paraphrases could also be derived (Bannard and Callison-Burch, 2005) by studying phrases that can be translated to a same phrase.

5 Design Goals

An abstract algorithm for extracting translations could be summarized by the following three objectives: (1) generate a set of candidate translations that includes the correct ones. (2) classify them and ensure that the correct ones are the best ones. (3) decide how many are to be kept.

The process we designed implements that abstract strategy in a greedy style way: it iterates

over a set of candidate translations and, at the end of each iteration, it validates one and discard others. The process finally stops when no candidate translations are left. The first objective thus remains the same. The second and third objectives are however final results: the classification of the best candidate translations and the number that are to be kept is only be established when the process stops. By being a greedy-style process, the task of deciding what are the correct translations is split into less-difficult sub-tasks, one for each iteration, where the process “just” needs to have as its best candidate translation a correct one.

5.1 Design criteria applied

The process should be able to acquire translations covering phrases of any sizes, i.e. strict restrictions on the size are to be avoided. The process should be able to acquire multiple (n to m) translations, i.e. strict restrictions on the number of translations for each phrase are to be avoided.

5.2 Abstract strategies for choosing

Local and global significance. All occurrences of a candidate translation should not have the same significance. The significance of an occurrence should take into account the number of other candidate translations also occurring in the same bitex with which it conflicts. In other words, the fewer are the candidate translations covering a same phrase in a bitext, the more interesting should be a candidate translation covering it. The process should also favour the candidate translations with a larger number of occurrences, i.e. the more recurrent a candidate translation is, the more interesting it is. The process should thus take into account both the number of occurrences and the significance of each, i.e. the significance of the occurrences of a candidate translation should be evaluated on “quality” and “quantity”.

Frequencies likeliness Since we deal with translated texts, the vast majority of phrases in a bitext have a translation. Two phrases that can be translated one to the other should therefore have similar frequency. However, since the process should also cope with multiple translations, i.e. the process should also consider that occurrences can be divided among several translations.

5.3 Abstract strategy for discarding

The process should maintain coherence with previously validated candidate translations. Thus, previously validated ones should allow to discard the remaining ones that are not compatible with them. One can think of it as a sudoku-like strategy, i.e. taking a decision for one box/phrase allow to reduce the options for other boxes/phrases.

6 Detailed Description

6.1 Candidate generation

For both texts in each bitext, we generate the set of every phrases occurring and count how many times they occur, i.e. how many slots they have. We produce candidate translations by computing the Cartesian product between the two sets of phrases of every bitext and rule out most of them by applying the following permissive criteria:

- (1) both covered phrases should occur at least \min_occ times in the corpora,
- (2) the covered phrases should co-occur in at least \min_co_occ bitexts.
- (3) both covered phrases covered should be among the \max_cand phrases they co-occur the most with.

6.2 Choosing the best candidate

The process keeps at each iteration the candidate translation ct maximizing the following score:

$$size(ct) * like_freq(ct) * significance(ct)$$

where $like_freq(ct)$ is the evaluation of the likeliness of the frequencies of the phrases covered by ct and $significance(ct)$ is a score representing the significance of its occurrences (see below).

6.3 Evaluation of frequencies likeliness

As briefly sketched in 5.2, phrases covered by a correct candidate translation should have similar frequencies. In order to illustrate the idea, let’s imagine that we only detect 1 to 1 translation and we classify phrases into three categories: low-frequency, medium-frequency, high-frequency. A metric trying to evaluate frequency likeliness will thus aim at giving a high score to candidate translation that cover phrases both classified in the same category and a lower one when classified in two different categories.

In practice, we do not classify the phrases into categories but assign to each phrase ph a frequency degree $fdeg(ph) \in [0, 1]$. This degree represents how frequent it is with regards to the other phrases of the same language. The most frequent ones receive a degree close to 1 and the less frequent ones a degree close to 0. We compute the frequency degree of a phrase ph as

$$fdeg(ph) = \frac{(nb_inf+1)}{nb_phrases}$$

where nb_inf is the number of phrases less frequent than ph and $nb_phrases$ is the total number of phrases of the language. Then, for each candidate translation ct covering two phrases ph_i and ph_j , we compute a score

$$like_freq(th) = 1 - abs(fdeg(ph_i) - fdeg(ph_j)).$$

Two aspects are to be considered. The first is the reason for computing the value $fdeg$ with the rank and not directly using the frequency. The reason is that it is not possible to know if a difference in x occurrences matters the same at different levels of frequency and with different languages. However, since we deal with translated texts, ordering them by frequency should stand from one language to the other and two phrases that are translations of one another should receive a similar $fdeg$.

The second aspect to consider is the fact of dealing with multiple translations. Therefore, occurrences can be divided among several translations. Since there is no reason for all translations to be equally balanced in frequency, one translation should dominate the others¹. Every time a candidate translation is validated, we decrease the frequency of both covered phrases by the number of slots claimed by the validated candidate translation. If a phrase has multiple translations, validating the dominating one allows to “re-synchronize” the frequency with the next dominating one. We thus recompute the $like_freq$ value for the remaining candidate translations covering one of the two covered phrases.

6.4 Significance score

As briefly explained before in 5.2, we aim at evaluating the significance of a candidate translation

¹especially if enforced for enhancing translation standardization.

on both the “quantity” and the “quality” of its occurrences.

For every phrase ph_i having slots available in a bitext b_n , we compute a value

$$claimed(ph_i, b_n) = \sum_{k=0}^n claims(ct_k, b_n)$$

of all ct_k candidate translations covering it.

Then, for each candidate translation ct occurring in a bitext b_n and covering two phrases ph_i and ph_j , we compute a local score of significance

$$local(ct, b_n) = \frac{claims(ct, b_n)}{claimed(ph_i, b_n)} * \frac{claims(ct, b_n)}{claimed(ph_j, b_n)}$$

The value of $local(ct, b_n)$ will be equal to 1 if ct is the only one covering ph_i and ph_j and drop towards 0 as the number of candidate translations covering one of them raises.

Finally, we compute at every iteration the score

$$significance(ct) = \sum_{i=0}^n local(ct_k, b_n)$$

6.5 Updating candidate translations

We apply a strategy that maintain coherence with the previously candidate translations. For every occurrence of a validated candidate translation, two types of restrictions, strict and soft ones, are dynamically build so as to inflict handicap to the remaining candidate translations that are in conflict with the validated one.

Whenever a candidate translation ct conflicts with a restriction set in a bitext b_n its value $claims(ct, b_n)$ and thus its significance score and global score are re-evaluated for the next iteration. If the value $claims(ct, b_n)$ falls to 0, its occurrence is removed. If a candidate translation does not fulfil any more the original criteria of occurring in at least min_co_occ different bitext (see 6.1), it is discarded.

6.5.1 Strict restrictions

Strict restrictions lock the slots of the phrases covered by the previously validated candidate translations, i.e. some slots become not “claimable” any more. For two phrases ph_i and ph_j covered by a validated candidate translation ct and each bitext b_n in which it occurs, the values $locks(ph_i, b_n)$ and $locks(ph_j, b_n)$ are incremented by $claims(ct, b_n)$.

6.5.2 Soft restrictions

Soft restrictions impact candidate translation that cover phrases that are included or are included by a validated candidate translation. To

each soft restriction $soft_m$ set on an phrase ph_i is associated a number of slots $num(soft_m)$ that a candidate translation cannot claim if it does not fulfil the condition.

Whenever a phrase PH_1 covered by a validated candidate translation ct includes a phrase ph_1 , we consider that the translation of ph_1 should be included by the second n-gram PH_2 also covered by ct . We associate to such soft restriction $soft_m$ a $num(soft_m) = claims(ct, b_n)$. In other words, if PH_1 includes ph_1 , we consider that for $claims(ct, b_n)$ slots of ph_1 its translation should be included in PH_2 . For example if “la bella casa” in Italian is validated as the translation of “das schöne Haus” in German then, for any bitext containing both, phrases included in “la bella casa” should translate to phrases included in “das schöne Haus” and vice-versa.

Also, whenever a phrase PH_1 covered by a validated candidate translation ct is included in a phrase ph_1 and $slots(ph_1, b_n) = slots(PH_1, b_n)$, we consider that the translation of ph_1 should include the other phrase PH_2 covered by ct . We associate to such soft condition $soft_m$ a $num(soft_m) = claims(ct, b_n)$. In other words, if PH_1 is included in ph_1 and both phrases have the same original number of slots then PH_2 should be included by the translation of ph_1 at least $claims(ct, b_n)$ time(s). For example if “bella” is validated as the Italian translation of “schöne” in German then phrases including “bella” and having the same number of slots should translate into phrases including “schöne” and vice-versa.

6.5.3 Combining restrictions and updating the remaining candidate translations

Since we do not try to align phrases, combining the restrictions violated by a candidate translation must take into account that some restrictions may apply on slots that overlap between one another.

Regarding strict restrictions, we can ensure that two restrictions concern a set of slots that don't overlap even if we don't explicitly affect a given slot to a given strict restriction. For example, for a phrase ph_i with $m + n$ slots in a given bitext that is covered by two validated candidate translations ct_e and ct_h , we can tell that m slots have been locked by ct_e and n slots by ct_h and cannot

be claimed by other candidate translations without stating explicitly which slot is locked by ct_e or ct_h .

Whenever a soft restriction is involved, simply adding the number of slots covered by the restrictions would be incorrect because we cannot establish if the restrictions violated do not overlap on a same set of slots. For example, let's consider a bitext containing both one occurrence of “la bella casa” in Italian and “das schöne Haus” in German with only one occurrence of “bella” and “schöne” in the whole bitext and two validated candidate translations ct_i and ct_j that associate “la bella” with “das schöne” and “bella casa” with “schöne Haus”. A candidate translation ct_k that covers “bella” but does not associate it with “schöne” would violate both soft restrictions set by ct_i and ct_j . Simply adding the number of slots covered by the soft restrictions set by ct_i and ct_j would prohibit ct_k to claims two slots when only one is actually available. The same reasoning can be extended to phrases having more than one slot and to the combination soft and strict restrictions.

We thus look for the maximum number of slots that a remaining candidate translation ct occurring in a bitext b_n and covering two phrases ph_i and ph_j can claim. For each covered phrase ph , we compute a value $max_soft(ph, ct, b_n)$ corresponding to the maximum of the $num(soft_m)$ values of the soft restrictions violated by ct for covering ph in b_n .

We then compute the value $sub_claims(ph, ct, b_n)$ corresponding to the number of slots originally available $slots(ph, b_n)$ minus the maximum value between the number of slots locked by strict restrictions $locks(ph, b_n)$ and $max_soft(ph, ct, b_n)$,

$$sub_claims(ph, ct, b_n) = slots(ph, b_n) - max(locks(ph, b_n), max_soft(ph, ct, b_n)).$$

Finally we update the $claims(ct, b_n)$ value in a similar manner as it has been first initialized

$$claims(ct, b_n) = min(sub_claims(ph_i, ct, b_n), sub_claims(ph_j, ct, b_n))$$

It is important to note that generally only one slot is available for most phrases in a bitext. Therefore, conflicting with just one restriction in a bitext, be it a strict or a soft, is enough for most candidate translations to loose their occurrence.

7 Preliminary Evaluation

7.1 Input Corpora and configuration

To perform the evaluation, we extracted 50 000 bitexts from the Catex Corpus (Streiter et al., 2004). This bilingual corpus is a collection of Italian-language legal texts with the corresponding German translations. The bitexts in this corpus are composed of two sentences. The average length for the Italian sentences was 15,3 tokens per sentence and 13,9 for the German ones.

We have set the *min_occ* and *min_co_occ* variables to 3 and the *max_cand* variable to 20 (see Sect. 6.1). The maximum size of a candidate translation was set to 12 tokens, i.e 6 for each phrase covered. A total of 57 406 candidate translations have been generated.

7.2 Evaluation protocol

As explained in section 3, comparing the methods to the state-of-the-art is not straightforward. The closest method in terms of input data and objectives is the one described in Lavecchia et al. (2008). However, the results are evaluated according to the performance of a machine translation tool which is a task out of the reach of the preliminary evaluation we wanted to performed. We thus decided to establish an evaluation protocol as close as possible to the bilingual extraction methods such as those described in Melamed (2000) and Moore (2001). In these papers, the authors classify the precision of candidate translations into three categories: wrong, correct and “near misses”. Even though these notions are quite straightforward for translations covering phrases of one or two tokens, they are more difficult to apply to larger ones. We thus report results obtained with several strategies for evaluating precision.

All evaluation strategies performed start from a manual evaluation that states the minimum number of tokens $errors(ct)$ that are to be added or deleted in both phrases covered by a candidate translation ct so as to obtain a fully valid translation. For each candidate translation, we thus start by evaluating how close it is from a perfect translation. For example, a candidate translation linking “landesgesetz vom 8 november” with “legge provinciale 8 novembre” is fully valid and receives a perfect score $errors(ct) = 0$ whereas an-

other one linking “landesgesetz vom 8 november” with “provinciale 8 novembre” requires to add “legge” before “provinciale” and thus receives a score $errors(ct) = 1$. 6 samples of 500 candidate translations at different ranking have been manually evaluated by a trained interpreter.

We then applied the following two strategies to evaluate the precision of each candidate translations and compute average precisions over the 6 samples. The first strategy, called hereafter *Scalable precision*, assigns a precision score equal to $(size(ct) - errors(ct))/size(ct)$. The second strategy, called hereafter *Strict precision*, is a generic one that is instantiated with a threshold value *thresh*. It classifies a candidate translation ct as “correct” or “wrong” depending on whether or not $errors(ct)$ is under or above *thresh*: ct receives a precision score of 1 if $errors(ct) \leq thresh$ and 0 otherwise. The thresholds chosen are not static but dynamically adjusted according to the size of the candidate translation evaluated. For example, if we set the threshold to $(3 * size)/12$ then a candidate translation ct_1 with $size(ct_1) = 6$ needs $errors(ct_1) \leq 1.5$ to be considered correct whereas a candidate translation ct_2 with $size(ct_2) = 12$ needs $errors(ct_2) \leq 3$.

7.3 Results

Table 1 provides some statistics about the validated candidate translations between two ranks: their average size, significance score and number of occurrences. Table 2 provides the results in terms of precision. The results are provided by evaluation criteria, sample and size of candidate translations. In each cell, the left value is the precision for the sample whereas the right one is a cumulative precision of all candidate translations with the same size. In table 3, each cell contains the number of candidate translations generated between two ranks according to their size and the proportion it represents among the ones generated between these two ranks.

As one can observe from table 2, precision remain stable and fairly high for the first two thirds. It then start decreasing noticeably and crumble in the last sample. An interesting result is that more than 50% of the candidate translations (30 000 over 57 406) have a close-to perfect precision

($> 98\%$ in scalable precision) and around 70% (40 500 over 57 406) have a reasonably high one ($> 95\%$ in scalable precision).

When analyzing the last part of the list we could observe that most errors are either random occurrence of a candidate translation covering frequent phrases or, as explained in Melamed (2000), indirect associations.

A first obvious observation is that the quality of the candidate translations does improve with their score, i.e. the higher the score is the better is the candidate translation.

When looking at table 1, apart from the first 10 000 candidate translations, we can see that the average frequency remains quite stable when compared with the average frequency, i.e. some less frequent candidate translations do receive a greater score than more frequent ones and therefore the average frequency remains stable. This behaviour meets our expectations since we wanted the process to not only consider the number of occurrences to decide whether or not a candidate translation was better than another one.

As said before, because the input corpora, the type of translations, and the form of evaluation are different, comparing our results to the state-of-the-art is challenging. Nevertheless, we noticed two aspects in favour of our results. For methods such as in Moore (2001) or Sahlgren and Karlsgren (2005), the average number of occurrences of the candidate translations they reported is rather high. It is worth highlighting that our method achieves very high precision with frequent candidate translations. For other methods such as in Tufis (2002), the number of candidate translations seems relatively small when compared with the number of sentences provided. It is worth noting that we outputted more candidate translations than the total number of sentences we gave as input.

8 Future Work

By the very nature of the approach, the implementation process is heavy in terms of memory and computations. As it starts with an exponential number of phrases and thus an exponential number of candidate translations, running the process with the above configuration consumed up to 26 Gbyte during its first iterations, and ran for 5 days on a recent computer. For this technical reason,

we had to limit our experiment and set the configuration to be more restrictive than originally intended. There is therefore a need to decrease the search space or dynamically adapt it.

Another evaluation using the generated translations in a machine translation system has been postponed as it implies external tools and additional knowledge. To do so, we could reproduce the evaluation performed in Lavecchia et al. (2008). As the method is very recent, we wanted first to have an preliminary overview of its potential to be able to consider further developments and evaluations. Another interesting evaluation would be to compare the phrases translations to the ones we could extract from the phrases alignment methods such as the one used in the Moses toolkit (Koehn et al., 2003).

As it has been designed as a greedy-style process, converting it into a beam-search process seems a viable option.

Like most natural language processing methods, this process would benefit from reducing the data sparsity. As it is currently designed, we could add a pre-processing that converts an input form-based parallel corpus into a lemmatised one.

Finally, several future works can be considered by reusing the data generated for other subjects that could take advantage of it (see sect. 4).

9 Conclusion

In this paper, we have presented an unsupervised and knowledge-free greedy-style process to derive multiple translations of phrases of equal or different sizes.

As it is still a recent and an on-going work, it has still much room for improvement. Several tracks towards this objective have been provided.

Finally, the preliminary evaluation performed has confirmed both its relevance and its potential.

Rank	Avg. size	Avg. signif	Avg. occ
< 10000	6.99	44.67	26.02
10000-19999	7.18	6.50	8.35
20000-29999	5.48	4.38	7.06
30000-39999	4.61	3.02	6.38
40000-49999	3.77	1.45	7.66
≥ 50000	2.24	0.23	5.67

Table 1: Statistics on average size, significance and occurrences depending on rank.

Scalable precision												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	0.95/0.95	1.00-1.00
10001-10500	1.00/1.00	0.80/0.80	0.99/0.99	0.96/0.97	0.99/1.00	0.87/0.92	0.98/0.99	0.96/0.97	0.99/0.99	1.00/1.00	0.98/0.98	0.98-0.99
20001-20500	0.96/0.98	0.92/0.87	0.97/0.99	0.96/0.96	0.96/0.98	0.92/0.92	0.98/0.98	0.93/0.96	0.97/0.99	0.93/0.95	0.93/0.96	0.96-0.98
30001-30500	0.97/0.98	0.84/0.86	0.98/0.98	0.91/0.95	0.92/0.98	0.88/0.92	0.88/0.98	0.89/0.95	0.97/0.99	-	0.83/0.96	0.97-0.98
40001-40500	0.83/0.92	0.85/0.85	0.83/0.96	0.77/0.87	0.88/0.97	0.75/0.86	0.89/0.97	0.91/0.94	0.85/0.99	0.91/0.95	0.25/0.95	0.83-0.95
50001-50500	0.16/0.55	0.15/0.49	0.16/0.93	0.40/0.85	0.47/0.96	0.33/0.84	0.38/0.97	0.44/0.93	-	-	-	0.17-0.82
Strict precision, thresh = (0 * size) / 12												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	0.99/0.99	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	0.95/0.95	1.00/1.00	0.43/0.43	0.98-0.98
10001-10500	1.00/1.00	0.40/0.40	0.97/0.99	0.82/0.88	0.97/0.98	0.42/0.65	0.93/0.96	0.71/0.81	0.95/0.95	1.00/1.00	0.85/0.76	0.92-0.95
20001-20500	0.95/0.98	0.75/0.62	0.92/0.96	0.78/0.82	0.85/0.94	0.56/0.61	0.88/0.93	0.43/0.66	0.80/0.93	0.68/0.76	0.50/0.68	0.84-0.91
30001-30500	0.96/0.97	0.53/0.57	0.95/0.96	0.65/0.77	0.67/0.92	0.33/0.57	0.75/0.93	0.50/0.65	0.67/0.93	-	0.00/0.67	0.90-0.91
40001-40500	0.79/0.90	0.68/0.65	0.66/0.92	0.33/0.58	0.57/0.87	0.23/0.45	0.45/0.88	0.46/0.60	0.25/0.92	0.00/0.71	0.00/0.65	0.62-0.85
50001-50500	0.14/0.54	0.12/0.38	0.07/0.88	0.40/0.57	0.33/0.86	0.00/0.43	0.00/0.87	0.00/0.59	-	-	-	0.13-0.73
Strict precision, thresh = (1 * size) / 12												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	0.99/0.99	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	0.95/0.95	1.00/1.00	1.00/1.00	0.99-0.99
10001-10500	1.00/1.00	0.40/0.40	0.97/0.99	0.82/0.88	0.97/0.98	0.42/0.65	0.93/0.96	0.71/0.81	0.95/0.95	1.00/1.00	0.96/0.97	0.93-0.96
20001-20500	0.95/0.98	0.75/0.62	0.92/0.96	0.78/0.82	0.85/0.94	0.56/0.61	0.88/0.93	0.43/0.66	0.80/0.93	0.68/0.76	0.79/0.91	0.84-0.92
30001-30500	0.96/0.97	0.53/0.57	0.95/0.96	0.65/0.77	0.67/0.92	0.33/0.57	0.75/0.93	0.50/0.65	0.67/0.93	-	0.00/0.90	0.90-0.92
40001-40500	0.79/0.90	0.68/0.65	0.66/0.92	0.33/0.58	0.57/0.87	0.23/0.45	0.45/0.88	0.46/0.60	0.25/0.92	0.00/0.71	0.00/0.88	0.62-0.86
50001-50500	0.14/0.54	0.12/0.38	0.07/0.88	0.40/0.57	0.33/0.86	0.00/0.43	0.00/0.87	0.00/0.59	-	-	-	0.13-0.74
Strict precision, thresh = (2 * size) / 12												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	0.99/0.99	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00-1.00
10001-10500	1.00/1.00	0.40/0.40	0.97/0.99	0.82/0.88	0.99/0.99	0.67/0.80	0.96/0.98	0.93/0.95	0.99/0.99	1.00/1.00	1.00/1.00	0.96-0.98
20001-20500	0.95/0.98	0.75/0.62	0.92/0.96	0.78/0.82	0.93/0.97	0.89/0.84	0.95/0.97	0.93/0.94	0.92/0.98	0.88/0.91	0.93/0.98	0.92-0.96
30001-30500	0.96/0.97	0.53/0.57	0.95/0.96	0.65/0.77	0.86/0.96	0.83/0.84	0.75/0.96	0.50/0.92	1.00/0.98	-	1.00/0.98	0.92-0.95
40001-40500	0.79/0.90	0.68/0.65	0.66/0.92	0.33/0.58	0.78/0.94	0.68/0.79	0.73/0.94	0.85/0.90	0.25/0.97	1.00/0.92	0.00/0.96	0.69-0.90
50001-50500	0.14/0.54	0.12/0.38	0.07/0.88	0.40/0.57	0.33/0.93	0.00/0.75	0.00/0.93	0.00/0.88	-	-	-	0.13-0.77
Strict precision, thresh = (3 * size) / 12												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	0.99/0.99	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00-1.00
10001-10500	1.00/1.00	0.40/0.40	0.99/0.99	1.00/1.00	0.99/0.99	0.67/0.80	0.96/0.98	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	0.97-0.98
20001-20500	0.95/0.98	0.75/0.62	0.97/0.98	1.00/1.00	0.93/0.97	0.89/0.84	0.98/0.98	1.00/1.00	0.96/0.99	0.94/0.96	1.00/1.00	0.95-0.97
30001-30500	0.96/0.97	0.53/0.57	0.97/0.98	0.94/0.98	0.86/0.96	0.83/0.84	0.75/0.97	1.00/1.00	1.00/0.99	-	1.00/1.00	0.94-0.97
40001-40500	0.79/0.90	0.68/0.65	0.79/0.95	0.79/0.90	0.78/0.94	0.68/0.79	0.95/0.97	0.85/0.96	1.00/0.99	1.00/0.96	0.00/0.98	0.78-0.93
50001-50500	0.14/0.54	0.12/0.38	0.07/0.91	0.40/0.87	0.33/0.93	0.00/0.75	0.50/0.97	0.00/0.94	-	-	-	0.14-0.80
Strict precision, thresh = (4 * size) / 12												
Rank \ Size	2	3	4	5	6	7	8	9	10	11	12	all
1-500	1.00/1.00	-	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00-1.00
10001-10500	1.00/1.00	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	0.99-0.99
20001-20500	0.95/0.98	1.00/1.00	0.97/0.98	1.00/1.00	0.98/0.99	1.00/1.00	0.98/0.98	1.00/1.00	1.00/1.00	0.97/0.98	1.00/1.00	0.98-0.99
30001-30500	0.96/0.97	1.00/1.00	0.97/0.98	0.94/0.98	1.00/0.99	1.00/1.00	0.75/0.97	1.00/1.00	1.00/1.00	-	1.00/1.00	0.97-0.98
40001-40500	0.79/0.90	0.93/0.95	0.79/0.95	0.79/0.90	0.96/0.99	0.77/0.92	0.95/0.97	1.00/1.00	1.00/1.00	1.00/0.98	0.00/0.98	0.84-0.96
50001-50500	0.14/0.54	0.13/0.53	0.07/0.91	0.40/0.87	0.33/0.98	0.00/0.88	0.50/0.97	0.00/0.98	-	-	-	0.14-0.82

Table 2: Precision depending on size, rank and evaluation criteria.

Rank / Size	2	3	4	5	6	7	8	9	10	11	12
0-9999	934-9%	45-0%	1928-19%	218-2%	1764-18%	485-5%	1340-13%	699-7%	977-10%	1031-10%	579-6%
10000-19999	682-7%	100-1%	1662-17%	307-3%	1649-16%	447-4%	1947-19%	629-6%	1299-13%	611-6%	667-7%
20000-29999	1169-12%	243-2%	2969-30%	699-7%	2416-24%	666-7%	748-7%	289-3%	360-4%	265-3%	176-2%
30000-39999	2223-22%	787-8%	3082-31%	948-9%	1160-12%	571-6%	489-5%	303-3%	260-3%	71-1%	106-1%
40000-49999	4191-42%	1545-15%	1690-17%	689-7%	636-6%	375-4%	369-4%	189-2%	194-2%	45-0%	77-1%
Total	15288-27%	3830-7%	11449-20%	2879-5%	7650-13%	2557-4%	4906-9%	2117-4%	3097-5%	2025-4%	1608-3%

Table 3: Number and distribution over size of the translation hypotheses generated.

References

- Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Y. Deng and W. Byrne. 2008. Hmm word and phrase alignment for statistical machine translation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):494–507.
- William A. Gale and Kenneth W. Church. 1991. Identifying word correspondence in parallel texts. In *Proceedings of the workshop on Speech and Natural Language, HLT '91*, pages 152–157, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Caroline Lavecchia, David Langlois, and Kamel Smaili. 2008. Phrase-Based Machine Translation based on Simulated Annealing. In European Language Resources Association (ELRA), editor, *Sixth international conference on Language Resources and Evaluation - LREC 2008*, Marrakech, Maroc.
- I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *In Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198.
- I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*.
- I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26:221–249.
- Robert C. Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the workshop on Data-driven methods in machine translation - Volume 14*, DMMT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.
- M. Sahlgren and J. Karlsgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Nat. Lang. Eng.*, 11(3):327–341, September.
- O. Streiter, M. Stuflesser, and I. Ties. 2004. Cle, an aligned tri-lingual ladin-italian-german corpus. corpus design and interface. *First Steps in Language Documentation for Minority Languages*, page 84.
- Keita Tsuji and Kyo Kageura. 2004. Extracting low-frequency translation pairs from japanese-english bilingual corpora. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 23–30, Geneva, Switzerland, August 29. COLING.
- Dan Tufis. 2002. A cheap and fast way to build useful translation lexicons. In *In Proceedings of the 19th International Conference on Computational Linguistics, COLING2002*, pages 1030–1036.
- Dominic Widdows, Beate Dorow, and Chiu ki Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *In Third International Conference on Language Resources and Evaluation*, pages 240–245.
- Dekai Wu and Xuanyin Xia. 1994. Learning an english-chinese lexicon from a parallel corpus. In *In Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213.
- Ying Zhang and Stephan Vogel. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *In Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT-05)*, pages 30–31.

KONVENS 2012

SFLR 2012: Workshop on Standards for Language Resources

September 21, 2012
Vienna, Austria

Preface

The need for automatically generated and processed linguistic resources has increased dramatically in recent years, due to the fact that linguistic resources are used in a wide variety of applications, including machine translation, information retrieval and extraction, sentiment analysis, lexicon creation, linguistic applications of machine learning, and so on.

Building linguistic resources is generally expensive, time-consuming, and requires highly specialized skills. It is important, therefore, for these resources to be reusable and inter-operable, for the benefit of the commercial and research communities. This would allow the same resources to be applied in different applications and possibly in different research areas. To achieve this, one fundamental problem is the issue of standardization of linguistic resources.

Language resource standards define, for example, how linguistic data can be created, imported and integrated in a platform-independent way. The relevant standardization activities are currently conducted by ISO/TC 37/SC 4, an international working group ('sub-committee') of the International Organization for Standardization (ISO).

The objective of subcommittee ISO/TC 37/SC 4 is to develop international standards and guidelines for effective language resource management in mono- and multilingual applications. It also develops principles and methods for representations and annotations of data, for the creation of categories for thesauri, ontologies, morphological and syntactic analysis and so on.

Subcommittee ISO/TC 37/SC 4 has so far published standards such as LAF (Linguistic Annotation Framework, ISO 24612:2012), SynLAF (Syntactic Annotation Framework, 24615:2010), LMF (Lexical Markup Framework, ISO 24613:2008). These standards define metamodels for data representation or the termsterminologies for data description and specify general requirements for linguistic resources. Other standards such as MAF (Morpho-syntactic Annotation Framework, ISO/CD 24611) and MLIF (Multilingual Information Framework, ISO 16642:2003) are still under development. Although a lot of work has been accomplished in over the past few years, there are still many goals to accomplish and much work to be done.

The goal of this workshop is to present the current status of ISO standards in the domain of language resources and language technology, and to discuss current and future applications of these standards. A number of academics and experts from industry will present their work in the field of standardization and application of linguistic resources.

The workshop includes tutorials, research presentations, use-cases and reports that allow participants to get acquainted with the standards. The workshop organizers Andreas Witt and Ulrich Heid currently chair the German DIN group for language resource standards. Andreas Witt also convened the ISO working group Linguistic Annotation (ISO/TC 37/SC 4/WG 6).

Andreas Witt, Ulrich Heid
Workshop Organizers

Workshop Program

Friday, September 21, 2012

- 09:15–09:30 Workshop Opening
Andreas Witt and Ulrich Heid
- 09:30–10:05 *Standards for language technology – Relevance and impact*
Gerhard Budin
- 10:05–10:30 *Standards for the technical infrastructure of language resource repositories:
Persistent identifiers*
Oliver Schonefeld, Andreas Witt
- 10:30–10:55 *Standards for the formal representation of linguistic data: An exchange format for
feature structures*
Rainer Osswald
- 10:55–11:15 Coffee break
- 11:15–11:40 *Towards standardized descriptions of linguistic features: ISOcat and procedures for
using common data categories*
Menzo Windhouwer
- 11:40–12:05 *Standardizing metadata descriptions of language resources: The Common Metadata
Initiative, CMDI*
Thorsten Trippel and Andreas Witt
- 12:05–12:30 General discussion
- 12:30–13:50 Lunch break
- 13:50–14:15 *Standardizing lexical-semantic resources – Fleshing out the abstract standard LMF*
Judith Eckle-Kohler
- 14:15–14:40 *A standardized general framework for encoding and exchange of corpus
annotations: The Linguistic Annotation Framework, LAF*
Kerstin Eckart
- 14:40–15:05 *Standard for morphosyntactic and syntactic corpus annotation: The Morpho-
syntactic and the Syntactic Annotation Framework, MAF and SynAE*
Laurent Romary
- 15:05–15:30 *Annotation specification and annotation guidelines: Annotating spatial senses in
SemAE-Space*
Tibor Kiss
- 15:30–15:50 Discussion: Using standard annotation in lexical and corpus resources

- 15:50–16:15 Coffee break
- 16:15–16:40 *Towards standards for corpus query: Work on a Lingua Franca for corpus query*
Elena Frick, Piotr Bánki and Andreas Witt
- 16:40–17:05 *Getting involved into language resource standardization: The map of standards and ways to contribute to ongoing work*
Gottfried Herzog
- 17:05–17:30 General discussion and closing of the workshop

Standards for language technology – Relevance and impact

Gerhard Budin
University of Vienna

Abstract

While there is common concensus that IT industry at large hardly works without international standards, it is far less obvious that language technologies for practical (industrial) as well as for research purposes should also heavily rely on international standards. In addition, there are several communities and organisations that do work on standards for language industry and for computational linguistics, which has led to some degree of fragmentation and lack of cooperation. The paper reviews the state-of-the-art of global standardization efforts for language resources and language technologies and identifies most urgent work items and most pressing needs for inter-organisational and cross-community collaboration efforts in order to achieve a higher degree of interoperability of formats, schemata, web services, data models, etc. for the wide spectrum of language technologies, the main goal of these efforts being to achieve a much higher impact of language technology standards in research as well as in industry.

Standards for the technical infrastructure of language resource repositories: Persistent Identifiers

Oliver Schonefeld and Andreas Witt

IDS Mannheim

Abstract

This presentation addresses the standard on Persistent Identification and Sustainable Access (PISA). It explains the usage of the international standard for persistent reference to and management of electronic language resources (ISO 24619:2011) prepared by ISO's Technical Committee TC37 (Terminology and other language and content resources), Subcommittee SC4 Language resource management). Its application allows to uniquely and sustainably identify language resources by issuing so-called persistent identifiers (PIDs) that can be assigned to single resources rendering them identifiable, referenceable and interoperable. Such a method and its application is indispensable for reliable collaborative linguistic research activities.

Standards for the Formal Representation of Linguistic Data: An Exchange Format for Feature Structures

Rainer Osswald
SFB 991
Heinrich-Heine-Universität Düsseldorf
osswald@phil.hhu.de

Abstract

The International Standard ISO 24610 defines a schema of how to encode feature structures and their declarations in XML. The main goal of this standard is to provide a format for the exchange of feature structures and feature system declarations between applications. This paper gives an overview of the elements of the standard and sketches its development and some of the design decisions involved. We also discuss the role of this standard in relation to other standardization proposals for language resources and we briefly address its relevance for application programming.

1 Feature structures

1.1 Feature structures in linguistics

In modern linguistics, the characterization of linguistic entities as bundles of *distinctive features* has been employed most prominently in phonology.¹ The phoneme /p/, for instance, can be seen as the result of combining the phonetic features *voiceless*, *plosive*, *bilabial*, etc. Phonetic features occur usually in opposition pairs such as *voiced* vs. *voiceless* and *plosive* vs. *non-plosive*. This dichotomy can be taken into account by introducing *binary* features like VOICE, PLOSIVE, etc., with possible values + and -. The phoneme /p/ is then described by a set of feature-value pairs, for which the following matrix notation is in use:

$$\begin{bmatrix} \text{VOICE} & - \\ \text{PLOSIVE} & + \\ \vdots & \end{bmatrix}$$

¹Cf. Chomsky and Halle (1968).

CATEGORY	<i>noun</i>						
WORDFORM	‘Junge’						
AGREEMENT	<table border="0"><tr><td>GENDER</td><td><i>masculine</i></td></tr><tr><td>NUMBER</td><td><i>singular</i></td></tr><tr><td>CASE</td><td><i>nominative</i></td></tr></table>	GENDER	<i>masculine</i>	NUMBER	<i>singular</i>	CASE	<i>nominative</i>
GENDER	<i>masculine</i>						
NUMBER	<i>singular</i>						
CASE	<i>nominative</i>						

Figure 1: Example of an untyped feature structure.

With the advent of unification-based grammars in (computational) linguistics such as Generalized Phrase Structure Grammar (GPSG) (Gazdar et al., 1985) and Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), more complex, nested feature structures arose. Figure 1 shows a simple example of a nested feature structure, in which the non-atomic value of the feature AGREEMENT is again a feature structure. The example describes part of the grammatical information associated with the German noun ‘Junge’ (‘boy’). The feature structure is *untyped* in that neither the main matrix nor the embedded matrix carries any sortal information. In this respect, this example differs from the *typed* feature structure shown in Figure 2, in which the main structure and every substructure carry a type (*phrase*, *word*, etc.). The latter example also illustrates the notion of *structure sharing*. The values of the AGREEMENT feature are (token) identical, which indicates that in a noun phrase, the agreement features of the determiner and those of the head noun must coincide.

The International Standard ISO 24610 provides a schema for the representation of feature structures in XML. The main purpose of such a representation is the standardized exchange of data be-

<i>phrase</i>	CATEGORY <i>np</i>												
SPEC	<table border="0"> <tr> <td><i>word</i></td> <td>CATEGORY <i>determiner</i></td> </tr> <tr> <td>WORDFORM ‘der’</td><td></td> </tr> <tr> <td>AGREEMENT ①</td><td> <table border="0"> <tr> <td><i>agreement</i></td> <td>GENDER <i>masculine</i></td> </tr> <tr> <td>NUMBER <i>singular</i></td><td></td> </tr> <tr> <td>CASE <i>nominative</i></td><td></td> </tr> </table> </td></tr> </table>	<i>word</i>	CATEGORY <i>determiner</i>	WORDFORM ‘der’		AGREEMENT ①	<table border="0"> <tr> <td><i>agreement</i></td> <td>GENDER <i>masculine</i></td> </tr> <tr> <td>NUMBER <i>singular</i></td><td></td> </tr> <tr> <td>CASE <i>nominative</i></td><td></td> </tr> </table>	<i>agreement</i>	GENDER <i>masculine</i>	NUMBER <i>singular</i>		CASE <i>nominative</i>	
<i>word</i>	CATEGORY <i>determiner</i>												
WORDFORM ‘der’													
AGREEMENT ①	<table border="0"> <tr> <td><i>agreement</i></td> <td>GENDER <i>masculine</i></td> </tr> <tr> <td>NUMBER <i>singular</i></td><td></td> </tr> <tr> <td>CASE <i>nominative</i></td><td></td> </tr> </table>	<i>agreement</i>	GENDER <i>masculine</i>	NUMBER <i>singular</i>		CASE <i>nominative</i>							
<i>agreement</i>	GENDER <i>masculine</i>												
NUMBER <i>singular</i>													
CASE <i>nominative</i>													
HEAD	<table border="0"> <tr> <td><i>word</i></td> <td>CATEGORY <i>noun</i></td> </tr> <tr> <td>WORDFORM ‘Junge’</td><td></td> </tr> <tr> <td>AGREEMENT ①</td><td></td> </tr> </table>	<i>word</i>	CATEGORY <i>noun</i>	WORDFORM ‘Junge’		AGREEMENT ①							
<i>word</i>	CATEGORY <i>noun</i>												
WORDFORM ‘Junge’													
AGREEMENT ①													

Figure 2: Example of a typed feature structure with structure sharing.

tween different applications. In what follows, we first give a formal definition of feature structures. Section 2 is concerned with the rationale behind the XML representation defined by the standard, while Section 3 gives an overview of the elements of the standard itself. In the concluding Section 4, we will briefly discuss how this standard is related to other standardization proposals for language resources and what role the standard can play for application programming.

1.2 Formal definition of feature structures

In Part 2 of the International Standard ISO 24610, feature structures are formally defined as follows. The definition presumes a given finite set \mathcal{F} of features, a finite type hierarchy \mathcal{T} with subtyping relation $<$ and a set \mathcal{X} of “built-in” elements (see below).²

Definition A *feature structure* over \mathcal{F} , \mathcal{T} and \mathcal{X} is a quadruple $\langle Q, \rho, \theta, \delta \rangle$, in which

- Q is a set of *nodes*,
- ρ is an element of Q , called the *root*,
- θ is a partial *typing function* from Q to \mathcal{T} ,
- δ is a partial *feature value function* from $\mathcal{F} \times Q$ to $Q \cup \mathcal{X}$,

such that, for every node $q \neq \rho$, there exists a sequence f_1, \dots, f_n of features and a sequence q_1, \dots, q_n of nodes with $q_1 = \rho$, $q_n = q$ and $q_{i+1} = \delta(f_i, q_i)$ for $i < n$.

²A type hierarchy is a finite ordered set $\langle \mathcal{T}, < \rangle$ such that every two elements of \mathcal{T} have a least upper bound in \mathcal{T} . In particular, every type hierarchy has a greatest element.

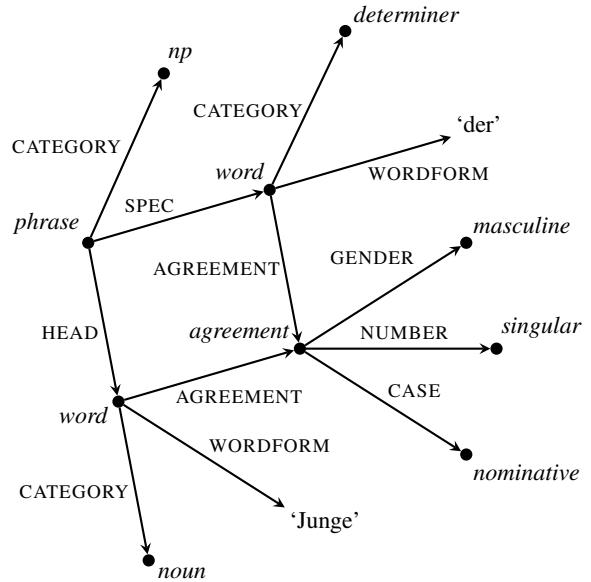


Figure 3: Labeled directed graph representation of the feature structure of Figure 2.

The last condition of the definition basically says that every node of a feature structure can be reached from the root by a feature sequence, or *path*. The given definition slightly generalizes the definitions typically found in the literature (e.g., Carpenter (1992)) in that, first, the typing function is only partial and thus allows for untyped feature structures and, second, the value of a feature can belong to a set \mathcal{X} of elements defined elsewhere. The members of \mathcal{X} can be thought of as built-ins such as binary values, string values, symbolic values and numeric values.

The relation between the above definition of feature structures and the feature-value matrices shown before is fairly straightforward. Roughly speaking, every opening square bracket and every atomic feature value (which is not a built-in) corresponds to a node. The typing function is determined by the types attached to the matrix brackets and by the atomic values (except, again, for built-ins). The feature value function is defined in accordance with how the features in the matrix connect one bracket to another or to an atomic value.³

³There has been some discussion in the literature (e.g. Carpenter (1992), Pollard and Sag (1994)) about the question as to whether feature-value matrices should better be seen as *descriptions* of feature structures. This is not the place to take up this discussion, but it seems that if this dis-

Figure 3 depicts the feature structure of Figure 2 as a labeled directed graph, where the labeled edges represent the feature value function and the labels of the nodes represent the typing function. Note that the two string values are treated as built-ins in this example.

2 XML representation of feature structures

2.1 Data-oriented XML

The Extensible Markup Language (XML) is the most popular format for exchanging structured data and thus a natural candidate for providing an interoperable representation format for feature structures.⁴ Since XML documents have an associated tree model (and even provide means for expressing co-reference), they seem to be fairly closely related to feature structures.

There are, however, crucial differences between the two structures, which requires various design decisions concerning the precise representation of feature structures by means of XML. First, XML document trees consist of categorically different types of nodes, namely element nodes, attribute nodes, text nodes and some others, of which only element nodes can have descendants (keeping aside the case of ‘mixed content’).⁵ A second difference is that XML trees are *ordered* trees (like any representation that is based on serialization), in contrast to features structures. The inherent ordering of XML provides an easy way to encode list values (see Section 3.1) but, on the other hand, gives rise to an identity problem since the order of the features matters for the XML representations but not for represented feature structures. This issue could be resolved, if necessary, by a canonical, say, alphabetic ordering of the features.

Another important aspect of an XML representation is that it is well-suited for declaring and

tinction is relevant at all then an XML-based interchange format is more on the description side of the divide.

⁴Some basic knowledge of XML is assumed in the following; cf. Ray (2003) for an introduction.

⁵An XML document is said to have *mixed content* if it allows for elements which may contain character data that are interspersed with child elements. This is typically the case in text-oriented applications such as DocBook. Mixed content is generally regarded as inappropriate for data-oriented formats like the ones discussed in this article.

checking its structure by available XML schema languages and tools.

2.2 A rationale for feature structure representation in XML

Many of the design decisions that led to the ISO 24610 standard can already be found in the TEI Guidelines P3 dating back to the mid-1990’s, with SGML instead of XML in use at that time. Langendoen and Simons (1995) give a concise exposition of the decision process, which we briefly summarize in the following. First, it is clear that representing features by XML attributes is no option since attributes cannot have descendants and thus would not allow to represent nested feature structures. The following option is more promising: It seems fairly natural to represent features by XML elements. The feature structure shown in Figure 1 could then be represented in XML as follows:

```
<fs>
  <category>noun</category>
  <wordform>Junge</wordform>
  <agreement>
    <fs>
      <gender>masculine</gender>
      <number>singular</number>
      <case>nominative</case>
    </fs>
  </agreement>
</fs>
```

As Langendoen and Simons (1995) point out, the main disadvantage of this representation is the unrestricted proliferation of elements. As a consequence, there would be no way to define a general schema for feature structure representations since the content model would depend fully on application-specific elements. The proposed solution is to employ a general-purpose element *f* for features and to represent the names of the features as attributes of these elements:

```
<fs>
  <f name="category">noun</f>
  <f name="wordform">Junge</f>
  <f name="agreement">
    <fs>
      <f name="gender">masculine</f>
      <f name="number">singular</f>
      <f name="case">nominative</f>
    </fs>
  </f>
</fs>
```

The representation of the feature values poses a second issue. The previous proposal does not distinguish between arbitrary strings such as ‘Junge’ and symbols such as *noun* and *singular*, which belong to a given set of symbols. Moreover, the content model for *f* would allow text content as well as structured content. These problems can be resolved by embedding the feature values in separate elements such as *symbol* and *string*. String values are then encoded as text content of *string* and type symbols are represented as values of the attribute *value* of the element *symbol*, in line with the general practice to use text content in data-oriented XML for unrestricted text only. The resulting XML representation of the above feature structure now looks as follows:

```
<fs>
  <f name="category">
    <symbol value="noun"/>
  </f>
  <f name="wordform">
    <string>Junge</string>
  </f>
  <f name="agreement">
    <fs>
      <f name="gender">
        <symbol value="masculine"/>
      </f>
      <f name="number">
        <symbol value="singular"/>
      </f>
      <f name="case">
        <symbol value="nominative"/>
      </f>
    </fs>
  </f>
</fs>
```

Note that the typing of feature structures is not part of the above schema but could easily be implemented by an appropriate attribute of the *fs* element; cf. Section 3.1 below.

The structural schema of the feature structure representation developed so far is summarized by the following RELAX NG schema:⁶

⁶‘RELAX NG’ stands for ‘REgular LAnguage for XML, New Generation’. The presentations given in this article use the so-called compact syntax of RELAX NG. There is also an XML-based syntax and there are tools to convert RELAX NG schemas into other schema languages such as W3C XML Schema. Cf. van der Vlist (2003) for an excellent introduction to RELAX NG.

```
fs =
  element fs {
    element f {
      attribute name { text },
      ( symbol | string | fs )
    } *
  }

symbol =
  element symbol {
    attribute name { text }
  }

string = element string { text }
```

The proposed schema does not impose any constraints on the features or the values encoded by a feature structure representation. It characterizes the generic structure of feature structure representations in XML. The schema does not even exclude a feature name to occur more than once in the *f* children of an *fs* element. That is, the fundamental property of feature structures that features are functional is not part of the well-formedness of the representations licensed by the above schema. There is also no way to customize the schema in such a way that it would capture more specific constraints about the feature architecture and the type system. Grammar-based schema languages like RELAX NG simply do not support the formulation of conditions between attributes of different elements.⁷

For this reason, Langendoen and Simons (1995, p. 202) argue for devising a separate markup format which is capable of representing all kinds of constraints on the type system and the feature architecture associated with a particular application domain. The resulting XML document is called a *feature system declaration*. It specifies, among others, which features are admissible for a feature structure of a certain type and which values are admissible for a given feature. A further kind of constraint that can be expressed by an FSD is the feature co-occurrence restriction used with untyped feature structures.

3 The International Standard ISO 24610

The International Standard ISO 24610 grew out of a joint initiative of the Text Encoding Initiative

⁷The *rule-based* schema language Schematron, by comparison, is able to assert such constraints since it has full XPath support.

```

fs =
element fs {
    attribute type { xsd:Name }?,
    element f {
        attribute name { text },
        model.featureVal*
    }*
}

model.featureVal.complex =
model.featureVal.complex |
model.featureVal.single

model.featureVal.complex =
fs | vColl | vNot | vMerge

model.featureVal.single =
binary | symbol | numeric |
string | vLabel | default | vAlt

vColl =
element vColl {
    attribute org { "set" | "bag" |
                    "list" }?,
    ( fs | model.featureVal.single )*
}

vLabel =
element vLabel {
    attribute name { data.word },
    model.featureVal?
}

```

Figure 4: Slightly simplified excerpt of the ISO/TEI XML schema for feature structure representations (FSD).

(TEI) Consortium and the ISO Sub-Committee TC 37/SC 4 (Language Resources Management); cf. Lee et al. (2004). The goal was to develop an international standard for the representation of feature structures which can serve as an exchange format between applications.

The standard consists of two parts. The first part, ISO 24610-1 on features structure representation (FSR), was published 2006 while the second part, ISO 24610-2 on features system declaration (FSD), was published more than five years later. Due to the fairly long time span between the two publication dates and the strong interdependency of the two parts, there are plans for revising the ISO 24610-1 standard in order to make it fully compliant with the more recent additions of Part 2.

3.1 Feature structure representation (FSR)

The ISO standard builds on the schema motivated in Section 2.2. Figure 4 shows a slightly simplified excerpt of the XML schema listed as an normative appendix of the second part of the ISO 24610 document. The main differences compared to the schema developed above is an optional `fs` attribute `type`, which allows for the representation of typed feature structures, and an extended list of elements for representing feature values.

Structure sharing. Let us first consider the element `vLabel`. According to its specification shown in Figure 4, it can serve as a kind of “value wrapper”. Its purpose is to allow the representation of structure sharing, with its `name` attribute acting as the co-reference label. For example, the structure sharing between the two *agreement* substructures of the feature structure shown in Figure 2 could be represented as follows:

```

<fs type="np">
    ...
    <f name="spec">
        <fs type="word">
            <f name="agreement">
                <vLabel name="L1">
                    <fs type="agreement">
                        ...
                    </fs>
                </vLabel>
            </f>
        ...
        </fs>
    </f>
    <f name="head">
        <fs type="word">
            <f name="agreement">
                <vLabel name="L1"/>
            </f>
        ...
        </fs>
    </f>
</fs>

```

Built-in value elements. In addition to the already mentioned built-in value elements `symbol` and `string`, the ISO standard defines two further elements: `binary` and `numeric`, which have their obvious intended interpretation.

The `vColl` element allows the encoding of lists, sets and bags (i.e., multisets) of atomic and complex values. The corresponding schema in Figure 4 shows that the members of the collection are represented as children of the `vColl` el-

ement, while the interpretation as a set, list, or bag is simply indicated by the value of the attribute `org`. Note that lists can be straightforwardly represented this way because XML document trees are ordered trees by definition. There is of course no need to make use of the `<vColl org="list">` construct since lists can be represented directly as recursively nested feature structures by means of the two features FIRST and REST.⁸

The `vNot` and `vAlt` elements do not act as constructors but characterize a single value. In the case of `vAlt`, the value is specified as being one of the values listed as children of the element; in the case of `vNot`, the value is specified as being not the value given by the single child of that element.

Symbolic values vs. types. An issue addressed in Part 2 of the ISO 24510 document but not in the TEI P5 Guidelines is the relation between symbolic feature values and types. Starting with untyped feature structures, as we did in Section 2.2, it seems quite natural to introduce an element like `symbol` for encoding symbolic values. In the presence of types, on the other hand, such elements are redundant at best. A symbolic value can be treated as a type, which in turn can be identified with an atomic typed feature structure, that is, a feature structure without any features. The resulting difference in representation is illustrated by the following two examples:

```

<fs type="word">
  <f name="category">
    <symbol value="noun"/>
  </f>
  ...
<fs type="word">
  <f name="category">
    <fs type="noun"/>
  </f>
  ...

```

The advantage of the second representation is that symbolic values become on a par with types and are hence part of the type hierarchy, which is not the case with built-ins (cf. Section 1.2). We will return to this issue below in Section 3.2.

⁸Cf. Witt et al. (2009) for an application of the FSR schema that makes use of the latter option.

```

element fsdDecl { fsDecl+ }

fsDecl =
  element fsDecl {
    attribute type { xsd:Name },
    attribute baseTypes {
      list { xsd:Name+ }
    }?,
    fDecl+, fsConstraints?
  }

fDecl =
  element fDecl {
    attribute name { xsd:Name },
    vRange, vDefault?
  }

fsConstraints =
  element fsConstraints {
    (cond | bicond)*
  }

```

Figure 5: Simplified excerpt of the ISO/TEI XML schema for feature system declarations.

Feature libraries and feature-value libraries. The FSR standard provides means for defining libraries of feature-value combinations and of feature values, collected under the elements `fLib` and `fvlLib`, respectively. The idea is that referencing the items of these collections by unique identifiers may result in a more compact representation of a feature structure. Notice that type systems provide an alternative way of defining complex feature-value combinations that can be re-used at different places in other feature structures.

3.2 Feature system declaration (FSD)

Well-formedness vs. validity. As explained in Section 2.2, well-formedness with respect to the FSR schema is not concerned with any specific constraints on the feature architecture or the type system of the represented feature structure. The solution proposed in the TEI Guidelines starting with version P3 was to introduce a separate XML format for specifying such declarations. This idea and its current implementation in the TEI P5 Guidelines have been integrated into Part 2 of the ISO 24610 standard.

Document structure. The overall document structure of a feature system declaration (FSD) is

roughly described by the schema shown in Figure 5. A feature system declaration (`fsDecl`) consists of one or more feature structure declarations (`fsDecl`). A feature structure declaration specifies the type of the structure and, optionally, one or more “base” types of which the type is a sub-type. In addition, a feature structure declaration consists of one or more feature declarations (`fDecl`), which specify the name of the features, the range of possible values and, optionally, a default value.

Feature structure and type declarations. The following example sketches part of an FSD for the typed feature structure shown in Figure 2. The attribute `baseTypes` is used to declare that the type `word` is a subtype of the type `sign`.

```
<fsDecl type="word" baseTypes="sign">
  <fDecl name="category">
    <vRange>
      <vAlt>
        <symbol value="determiner"/>
        <symbol value="noun"/>
        <symbol value="verb"/>
        ...
      </vAlt>
    </vRange>
  </fDecl>
  <fDecl name="wordform">
    <vRange>
      <string/>
    </vRange>
  </fDecl>
  ...
</fsDecl>
```

The example illustrates how to declare the range of possible feature values by means of the `vRange` construct. In fact, this is the only option if atomic values are represented by built-in elements such as `symbol`. In order to represent *determiner*, *noun*, *verb* etc. as members of the type hierarchy, they have to be treated as atomic feature structures along the following lines:

```
<fsDecl type="word" baseTypes="sign">
  <fDecl name="category">
    <vRange>
      <fs type="category">
    </vRange>
  </fDecl>
  ...
</fsDecl>
```

```
<fsDecl type="determiner"
  baseTypes="category"/>
<fsDecl type="noun"
  baseTypes="category"/>
<fsDecl type="verb"
  baseTypes="category"/>
...
```

Hence, the type hierarchy can be represented by an FSD document by declaring types as atomic feature structures together with the more general types from which the type in question inherits.⁹

Feature structure constraints. One of the goals of the ISO standard is to allow for both, the feature structure declarations of typed feature structures in the style of HPSG (Pollard and Sag, 1994) and the representation of feature co-occurrence restrictions and defaults as employed in GPSG (Gazdar et al., 1985) and other untyped frameworks. While implicational constraints are also relevant for typed frameworks (cf., e.g., the principles of HPSG), they are fundamental to the untyped case. We will not go into detail here because the representation of such constraints is basically a question of how to represent logical expressions in general (which in turn is the topic of other initiatives such as RuleML). Examples adapted from Gazdar et al. (1985) can be found in Burnard and Bauman (2012, Sect. 18.11.4).

4 Discussion

Is FSR/FSD more than an exchange format?

The primary use case of the ISO 24610 standard is the exchange of feature structure data between applications. The question arises whether the XML representation of feature structures compliant with FSR is not only useful for exchanging data but also as a data format for application programming. In other words, can the XML representation serve as the native data format of an application? The appropriate answer is a qualified yes. On the one hand, existing XML technology (XML query languages, native XML databases) provides a powerful environment for working with FSR compliant data. On the other

⁹There is a caveat here. The schema for `fsDecl` in ISO 24610-2 and TEI P5 apparently poses a problem for this strategy since it requires at least one `fDecl` child. Hence, strictly speaking, atomic feature structures cannot be declared.

hand, checking the validity of FSR data directly on the basis of an FSD document is probably a tedious task, for which logic programming of one kind or another seems to be much better suited.

Is there a need for software tools? As an application-oriented exchange format, the ISO standard does not aim at human readability. The question is, whether there is a need for viewing and editing or even validity checking of FSR compliant data. Given that an application system typically uses its own internal representation, which is part of an elaborate editing and programming environment, the answer could be negative. However, if the ISO standard is established as an export format, then a sufficiently powerful, freely available (open source) tool for browsing, editing and probably checking FSR/FSD data would be extremely helpful for accessing the data. Moreover, the existence of such a tool could boost the dissemination of the standard. Potential users would need to transform their data into the ISO format in order to use the tool.

The ISO 24610 standard in context. How does the ISO 24610 standard relate to other standards for language resources developed by the ISO Sub-Committee TC 37/SC 4? According to the International Standard ISO 24612 ‘Linguistic annotation framework’ (LAF), the FSR standard plays a key role for any sort of annotation. LAF draws a clear-cut distinction between *referential structure* and *annotation content structure* and proposes that the latter be represented as feature structures compliant with FSR. It follows that the ISO 24610 standard is tied in with all standards related to linguistic content, be it morphosyntax, syntactic, or pragmatic content. Hence the standard contributes to the interoperability of the ISO standards for language resource management (Lee and Romary, 2010).

Acknowledgments

The writing of this paper has been supported by the Collaborative Research Center 991 ‘The Structure of Representations in Language, Cognition and Science’ funded by the German Research Foundation (DFG).

References

- Lou Burnard and Syd Bauman, editors. 2012. *TEI P5: Guidelines for Electronic Text Encoding and Interchange (Version 2.1.0)*. Text Encoding Initiative Consortium, Charlottesville, Virginia. [<http://www.tei-c.org/Guidelines/P5/>].
- Bob Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge.
- Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.
- Gerald Gazdar, Ewan Klein, Geoffrey K. Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Blackwell, Oxford.
- ISO 24610-1. Language resource management - Feature structures - Part 1: Feature structure representation. Publication date: 2006-04.
- ISO 24610-2. Language resource management - Feature structures - Part 1: Feature system declaration. Publication date: 2011-10.
- ISO 24612. Language resource management - Linguistic annotation framework (LAF). Publication date: 2012-06.
- D. Terence Langendoen and Gery F. Simons. 1995. A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities*, 29:191–209.
- Kiyong Lee and Laurent Romary. 2010. Towards interoperability of ISO standards for language resource management. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 95–103, Hong Kong. City University of Hong Kong.
- Kiyong Lee, Lou Burnard, Laurent Romary, Eric de la Clergerie Thierry Declerck, Syd Bauman, Harry Bunt, Lionel Clément Tomaz Erjavec, Azim Roussanaly, and Claude Roux. 2004. Towards an international standard on feature structures representation. In *Proceedings of LREC 2004*, pages 373–376, Lisbon, Portugal. Universidade Nova de Lisboa.
- Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.
- Erik T. Ray. 2003. *Learning XML*. O'Reilly, Sebastopol, CA.
- Eric van der Vlist. 2003. *RELAX NG*. O'Reilly, Sebastopol, CA.
- Andreas Witt, Georg Rehm, Erhard Hinrichs, Timm Lehmberg, and Jens Stegmann. 2009. SusTEInability of linguistic resources through feature structures. *Literary and Linguistic Computing*, 24(3):363–372.

Towards standardized descriptions of linguistic features: ISOcat and procedures for using common data categories

Menzo Windhouwer

Max Planck Institute for Psycholinguistics, Nijmegen

Abstract

Since 2009 the Max Planck Institute for Psycholinguistics in Nijmegen offers a web-based open source reference implementation of the ISO DCR (Data Category Registry, ISO 12620:2009), which is called ISOcat (“Data Category Registry for ISO TC 37”). ISOcat describes the data model and procedures for DCR. The talk presents the currently stage of the development and the status of ISOcat, and demonstrates - how ISOcat can be used for creating specifications of data category and management of a DRS. The talk also presents the plan for future development finally it give a view, what is planned to develop further..

Standardizing metadata descriptions of language resources: The Common Metadata Initiative, CMDI

Thorsten Trippel
University of Tübingen

Andreas Witt
IDS Mannheim

Abstract

The CMDI stands for Component Metadata Infrastructure. It is a framework for description of basic metadata components, which was developed particularly in the scope of the CLARIN project. This talk presents the framework, its components and provides a glimpse into the development of the ISO/NP 24622-1, which will be the ISO standardization of the basics components of CDMI.

Standardizing Lexical-Semantic Resources – Fleshing out the abstract standard LMF

Judith Eckle-Kohler[‡]

‡ Ubiquitous Knowledge Processing
Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Iryna Gurevych^{†‡}

† Ubiquitous Knowledge Processing
Lab (UKP-DIPF)
German Institute for Educational
Research and Educational Information
www.ukp.tu-darmstadt.de

Abstract

This paper describes the application of the Lexical Markup Framework (LMF) for standardizing lexical-semantic resources in the context of NLP. More specifically, we highlight the question how lexical-semantic resources can be made semantically interoperable by means of LMF and ISO Cat. The LMF model UBY-LMF, an instantiation of LMF specifically for NLP, serves as an example to illustrate the path towards semantic interoperability of lexical resources.

1 Introduction

Lexical-semantic resources (LSR) are used in major NLP tasks, such as word sense disambiguation, semantic role labeling and information extraction. In recent years, the aspects of reusing and merging LSRs have gained significance, mainly due to the fact that LSRs are expensive to build. Standardization of LSRs plays an important role in this context, because it facilitates integration and merging of LSRs and makes reuse of LSRs easy. NLP systems that are built according to standards can simply plug in standardized LSRs and are thus able to easily switch between different standardized LSRs. In other words, standardizing LSRs makes them interoperable.

Two aspects of interoperability are to be distinguished in NLP: syntactic interoperability and semantic interoperability (Ide and Pustejovsky, 2011). While NLP systems can perform the same kind of processing with *syntactically interoperable* LSRs, there is no guarantee, that the results

can be interpreted the same way. Two syntactically interoperable LSRs might use the same term to denote different meanings. *Semantically interoperable* LSRs, on the other hand, use terms that share a common definition of their meaning. Consequently, NLP systems that switch between semantically interoperable LSRs can perform the same kind of processing and the results produced can still be interpreted the same way.

In this paper, we focus on the question how to achieve semantic interoperability by means of the ISO 24613:2008 LMF (Francopoulo et al., 2006) and ISO Cat.¹ The comprehensive LMF lexicon model UBY-LMF (Eckle-Kohler et al., 2012) serves as an example to show how the abstract LMF standard is to be fleshed out and instantiated in order to make LSRs semantically interoperable for NLP purposes.

UBY-LMF covers very heterogeneous LSRs in two languages, English and German, and has been used to standardize a range of LSRs resulting in the large-scale LSR UBY (Gurevych et al., 2012), see <http://www.ukp.tu-darmstadt.de/uby/>. UBY currently contains ten resources in two languages: English WordNet (Fellbaum, 1998), Wiktionary², Wikipedia³, FrameNet (Baker et al., 1998), and VerbNet (Kipper et al., 2008); German Wiktionary, Wikipedia, GermaNet (Kunze and Lemnitzer, 2002) and IMSlex-Subcat (Eckle-Kohler, 1999) and the English and German entries of OmegaWiki⁴.

¹<http://www.isocat.org/>

²<http://www.wiktionary.org/>

³<http://www.wikipedia.org/>

⁴<http://www.omegawiki.org/>

2 LMF and semantic interoperability

First, we give an overview of the LMF standard and briefly describe how to use it. We put a special focus on the question how LSRs can be made semantically interoperable by means of LMF.

LMF – an abstract standard LMF defines a *meta-model* of lexical resources, covering both NLP lexicons and machine readable dictionaries. The standard specifies this meta-model in the Unified Modeling Language (UML) by providing a set of UML diagrams. UML packages are used to organize the meta-model and each diagram given in the standard corresponds to an UML package. LMF defines a mandatory core package and a number of extension packages for different types of resources, e.g., morphological resources or wordnets. The core package models a lexicon in the traditional headword-based fashion, i.e., organized by lexical entries. Each lexical entry is defined as the pairing of one to many forms and zero to many senses.

Instantiating LMF The abstract meta-model given by the LMF standard is not immediately usable as a format for encoding (i.e., converting) an existing LSR (Tokunaga et al., 2009). It has to be instantiated first, i.e., a full-fledged lexicon model has to be developed by choosing LMF classes and by specifying suitable attributes for these LMF classes.

According to the standard, developing a lexicon model involves

1. selecting classes from the UML packages,
2. defining attributes for these classes and
3. linking the attributes and other linguistic terms introduced (e.g., attribute values) to standardized descriptions of their meaning.

Selecting a combination of LMF classes from the LMF core package and from the extension packages establishes the structure of a lexicon model. While the LMF core package models a lexicon in terms of lexical entries, the LMF extensions provide classes for different types of lexicon organization, e.g., covering the synset-based organization of wordnets or the semantic frame-based organization of FrameNet.

Fixing the structure of a lexicon model by choosing a set of classes contributes to syntactic interoperability of LSRs, as it fixes the high-level organization of lexical knowledge in an LSR, e.g., whether synonymy is encoded by grouping senses into synsets (using the `Synset` class) or by specifying sense relations (using the `SenseRelation` class), which connect synonymous senses.

Defining attributes for the LMF classes and specifying the attribute values is far more challenging than choosing from a given set of classes, because the standard gives only a few examples of attributes and leaves the specification of attributes to the user in order to allow maximum flexibility.

Finally, the attributes and values have to be linked to a description of their meaning in an ISO 12620:2009 compliant Data Category Registry (DCR), see (Broeder et al., 2010). ISOCat is the implementation of the ISO 12620:2009 DCR providing descriptions of terms used in language resources.

These descriptions in ISOCat are standardized, i.e., they comply with a predefined format and provide some mandatory information types, including a unique administrative identifier (e.g., `partOfSpeech`) and a unique and persistent identifier (PID, e.g., <http://www.isocat.org/datcat/DC-396>) which can be used to link to the descriptions. The standardized descriptions of terms are called *Data Categories* (DCs).

Semantic Interoperability Connecting the linguistic terms used for attributes and their values in a lexicon model with their meaning defined externally in ISOCat contributes to *semantic interoperability* of LSRs (see also Windhouwer and Wright (2012)). The definitions of DCs in ISOCat constitute an interlingua that can be used to map idiosyncratically used linguistic terms to a set of reference definitions (Chiarcos, 2010). Different LSRs that share a common definition of their linguistic vocabulary are said to be *semantically interoperable* (Ide and Pustejovsky, 2010).

Consider as an example the `LexicalEntry` class of two different lexicon models A and B. Lexicon model A could have an attribute `partOfSpeech` (POS), while lexicon model B could have an attribute `pos`. Linking both attributes to the meaning “A category as-

signed to a word based on its grammatical and semantic properties.” given in ISOCat (<http://www.isocat.org/datcat/DC-396>) makes the two lexicon models semantically interoperable with respect to the POS attribute.

A human can look up the meaning of a term occurring in a lexicon model by following the link to the ISOCat DC and consulting its description in ISOCat. Linking the attributes and their values to ISOCat DCs results in a so-called Data Category Selection.

It is important to stress that the notion of “semantic interoperability” in the context of LMF has a limited scope: it only refers to the meaning of the linguistic vocabulary used in an LMF lexicon model – not to the meaning of the lexemes listed in a LSR.

3 UBY-LMF – Instantiating LMF

Considering the fact that only a fleshed-out LMF lexicon model, i.e., an instantiation of the LMF standard, can be used for actually standardizing LSRs, it is obvious that LMF-compliant LSRs are not necessarily interoperable, neither syntactically nor semantically.

Therefore it is important to develop a single, comprehensive instantiation of LMF, which can immediately be used for standardizing LSRs. The LMF lexicon model UBY-LMF strives to be such a comprehensive instantiation of LMF to be used in NLP.

3.1 UBY-LMF characteristics

UBY-LMF covers a wide range of lexical information types, since it has been designed as a uniform format for standardizing heterogeneous types of LSRs, including both expert-constructed resources – wordnets, FrameNet, VerbNet – and collaboratively constructed resources – Wikipedia, Wiktionary, OmegaWiki.

In UBY-LMF, there is one Lexicon per integrated resource, i.e., one Lexicon for FrameNet, for WordNet and so on. This way, the Lexicon instances can be aligned at the sense level by linking pairs of senses or synsets using instances of the SenseAxis class.

The full model consists of 39 classes and 129 attributes. Please refer to (Eckle-Kohler et al.,

2012) and (Gurevych et al., 2012) for detailed information on UBY-LMF and the corresponding large-scale LSR UBY.

UBY-LMF is represented by a DTD which can be used to automatically convert any given resource into the corresponding XML format. Converters for ten LSRs to UBY-LMF format are publicly available on Google Code, see <http://code.google.com/p/uby/>.

3.2 UBY-LMF attributes

In UBY-LMF, the definition of attributes for the LMF classes was guided by two requirements that we identified as important in the context of NLP: (i), comprehensiveness, and (ii) extensibility (Eckle-Kohler et al., 2012).

Comprehensiveness implies that the model should be able to represent all the lexical information present in a wide range of LSRs, because NLP applications usually require different types of lexical knowledge and it is difficult to decide in advance which type of lexical information will be useful for a particular NLP application.

Extensibility is also crucial in the NLP domain, because UBY-LMF should be applicable across languages (Gurevych et al., 2012), (Eckle-Kohler and Gurevych, 2012) and as well be able to adopt automatically extracted lexical-semantic knowledge.

4 UBY-LMF and semantic interoperability

In section 2, we have pointed out that linking attributes and values used in an LMF lexicon model to DCs in ISOCat is crucial for semantic interoperability.

Now we will take a closer look at the semantic interoperability of UBY-LMF compliant resources by describing in detail the grounding of UBY-LMF attributes and values in the ISOCat repository. First, we introduce ISOCat, then, we describe the process of selecting and creating an ISOCat Data Category Selection for UBY-LMF, and finally, we look at some limitations of the current version of UBY-LMF.

4.1 Overview of ISOCat

The Data Category Registry ISOCat is a collaboratively constructed repository where everybody

can register and create DCs. Users can assign their DCs to so-called Thematic Domains, such as Morphosyntax, Syntax or Lexical Resources. The ISOCat Web interface provides a form that guides the user through the process of creating a DC, also indicating which kind of information has to be provided mandatorily. The well-formedness of a newly created DC is automatically checked and displayed by a flag – a green marking indicating well-formedness.

Users can also group DCs, including self-created ones, into a Data Category Selection. Typically, Data Category Selections are created for projects or resources, e.g., there are Data Category Selections for RELISH⁵ or CLARIN⁶ or for resources, such as the STTS tagset⁷ or UBY.⁸ Data Category Selections can be made publicly available, in order to allow for linking to particular DCs.

It is possible to submit subsets of well-formed DCs to standardization. While the standardization of ISOCat DCs has not yet started, standardized DCs might be important for resources where sustainability is an issue, because standardized DCs can be considered as stable. Non-standardized DCs, on the other hand, could in principle be changed at any time by their owners which might also involve changes in their meaning.

Two types of DCs distinguished in ISOCat are relevant for LMF lexicon models (Windhouwer and Wright, 2012): first, complex DCs which have a typed value domain and typically correspond to attributes in an LMF lexicon model. According to the size of the value domain, DCs are classified further into open DCs (they can take an arbitrary number of values), closed DCs (their values can be enumerated) and constrained DCs (the number of values is too big in order to be enumerated, but yet constrained). Second, there are simple DCs which describe values of a closed DC.

The attributes and values defined in UBY-LMF refer to 175 ISOCat DCs; most of them are simple DCs. The corresponding Data Category Selection is publicly available in ISOCat, see

⁵<http://tla.mpi.nl/relish/>

⁶<http://www.clarin.eu>

⁷<http://www.isocat.org/rest/dcs/376>

⁸<http://www.isocat.org/rest/dcs/484>

<http://www.isocat.org/rest/dcs/484>.

Figure 1 shows a screenshot of the public Data Category Selection for UBY as of 2012.

4.2 UBY-LMF: Selecting DCs from ISOCat

Selecting DCs from ISOCat which are suitable descriptions of terms used in a lexicon model such as UBY-LMF is a task that requires the identification of fine-grained and subtle differences in meaning and hence, is a task where humans are superior to machines in terms of quality. For UBY-LMF, the intended meanings of the lexicon model terms and the textual descriptions of the DCs were manually compared in order to first identify candidate DCs with equivalent or similar meaning and then to select one of the candidate DCs as reference for a specific term.

Searching for DCs in ISOCat Identifying candidate DCs in ISOCat means accessing the ISOCat Web interface (<http://www.isocat.org/interface/index.html>) and searching for the lexicon model term or variants thereof. Searching for candidate DCs in ISOCat is time-consuming and not particularly user-friendly, because there are many DCs with similar names and equivalent or near-equivalent meaning. Currently ISOCat does not display relations between such terms (e.g., the equivalence relation).

This is offered to a limited extent by RelCat, a companion registry to ISOCat (Windhouwer, 2012). However, RelCat is still at an early stage of development and currently provides only relations between selected Data Category Selections, such as GOLD⁹ and RELISH. Therefore, we did not use it for the selection of DCs for UBY-LMF.

There are basically two ways of looking for DCs in ISOCat: first, by entering a search term and second by browsing Thematic Domains, such as Syntax, or by browsing Data Category Selections published by particular groups or projects, such as CLARIN or RELISH.

Choosing among several candidate DCs Typically, an ISOCat search query for a term (using the option *exact match*) yields a list of DCs that have slightly different names, but are very similar in meaning. Consider as an example the linguistic

⁹<http://linguistics-ontology.org/>

#	Name	Version	Administration	Registration status	Owner	Type	Owned by	Scope
4620	subcategorization Frame Set	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4624	subject Complement	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4187	subject Control	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4188	subject Raising	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4613	synset	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4623	that Type	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4162	toInfinitive	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public
4390	transparent Meaning	1:0	private	private	Eckle-Kohler, Judith	simple	Eckle-Kohler, Judith	public

1.1.1 Creation
Creation Date: 2011-11-09
Change Description: Created Data Category.

2. Description Section

Profile	Private
Profile	Lexical Resources
[-] 2.1 English Language Section	
Language	English (en)
2.1.1 Name Section	
Name	toInfinitive
Name Status	admitted name
2.1.2 Definition Section	
Definition	The non-finite verb form infinitive used with "to", as opposed to an infinitive used without "to". The German equivalent of "to" is "zu"; depending on the verb, "zu" can either precede the infinitive as in English or "zu" can be incorporated into the infinitive.
Source	Randolph Quirk et al., A grammar of Contemporary English (Longman)
2.1.3 Example Section	
Example	English: He likes to talk.
Source	Randolph Quirk et al., A grammar of Contemporary English (Longman)
2.1.4 Example Section	
Example	German: Wir freuen uns, ihn zu sehen.
Source	Helbig, Buscha, Deutsche Grammatik (Langenscheidt)
2.1.5 Example Section	
Example	German: Wir freuen uns, ihn abzuholen. (incorporated "zu")

Figure 1: Screenshot of the public Data Category Selection “Uby 2012“ in ISOCat, see <http://www.isocat.org/interface/index.html>.

term *determiner* which is a possible value of the attribute `partOfSpeech` in UBY-LMF. There are 25 DCs in ISOCat that exactly match the term *determiner*.

Another example is the term *direct object* that is sometimes used for specifying the accusative noun phrase argument of transitive verbs in German. In ISOCat, there are two different specifications of this term, one explicitly stating that this accusative noun phrase argument can become the clause subject in passivization (<http://www.isocat.org/datcat/DC-1274>), the other not mentioning passivization at all (<http://www.isocat.org/datcat/DC-2263>).

We tried to select a DC with a meaning as close as possible to the intended meaning in UBY-LMF. When selecting a particular DC from a list of candidates, we followed two simple strategies:

- We sorted the search results by Owner, i.e., by the person who created the DC, and preferred DCs which are owned by experts, either in the standardization community or in the linguistics community.
- We preferred those DCs that had passed the ISOCat test for well-formedness.

Sometimes, selecting an existing DC from ISOCat involved making compromises. For instance, the attribute value `taxonomic` of the attribute `relType` of the `SenseRelation` class links to the ISOCat DC taxonomy (<http://www.isocat.org/rest/dc/4039>) which has a narrower meaning than the meaning intended in UBY-LMF: While `taxonomic` in UBY-LMF denotes a type of sense relation that defines a taxonomy in a general sense, covering

all kinds of lexemes, (see Cruse (1986)), the DC 4039, taxonomy, is restricted to controlled vocabulary terms organized in a taxonomy.

Note that many attributes or attribute values in UBY-LMF refer to ISOCat DCs with a different (so-called admitted) name. This is explicitly supported by ISOCat, because each DC definition may optionally contain Data Element Name Sections in order to record other names for the DC as used in different sources, such as a given database, format or application. In this manner, the number of parallel DCs in ISOCat with equivalent meaning can be limited. We left the specification of Data Element Name Sections in the newly created DCs (holding a deviating UBY name) to future work.

Semantic Drift As ISOCat is a collaboratively constructed and thus dynamically evolving repository, DCs that are not standardized yet could eventually change their meaning over time, e.g., they might be further fleshed out and specified in more detail.

Since the public release of UBY in March 2012, we have already encountered such a case of semantic drift for the DC verbFormMood (see <http://www.isocat.org/rest/dc/1427>) which we selected as a description of the attribute verbForm attached to the SyntacticCategory class in the Syntax part of UBY-LMF.

The DC verbFormMood has been changed after March 2012 (according to the change log, this DC was changed on 2012-06-09) and is no longer similar enough to the intended meaning. Therefore, we will create a new DC which specifies verbForm as a property of verb phrase complements in the context of subcategorization.

4.3 UBY-LMF: Creating new ISOCat DCs

The majority of attributes and values in UBY-LMF refer to already existing ISOCat DCs. Yet, we had to create 38 new DCs in ISOCat, as particular definitions were missing. We decided to create new DCs in those domains where much standardization work has already been done or large sources of linguistic expert knowledge are available. In particular, these are the domains lexical syntax (related to subcategorization), derivational morphology, and frame semantic information.

We used the following expert-based sources as references for newly created DCs:

- the EAGLES synopsis on morphosyntactic phenomena¹⁰ (Calzolari and Monachini, 1996), as well as the EAGLES recommendations on subcategorization¹¹ have been used to identify DCs relevant for lexical syntax
- traditional grammars such as the English grammar by Quirk (1980)
- a Web-based encyclopedia of linguistic terms, collaboratively built by linguists, see <http://www.glottopedia.de>
- the detailed documentation of the FrameNet resource (Ruppenhofer et al., 2010)

Fifteen of the newly created DCs are required for representing subcategorization frames in UBY-LMF. Most of these DCs are simple DCs that are linked to attribute values in UBY-LMF. The corresponding DCs were either missing in ISOCat or there were only DCs with a related, but not sufficiently similar meaning.

For instance, we created a DC for to-infinitive complements as in *He tried to address all questions..* While both *infinitive* and *infinitiveParticle* are present in ISOCat, all available definitions of *infinitive* do not explicitly exclude the presence of a particle, neither <http://www.isocat.org/datcat/DC-1312> nor <http://www.isocat.org/datcat/DC-2753>. What is required, however, for the specification of verbal complements are individual DCs for bare infinitives used without particle and for infinitives used with particle (i.e., *to* in English).

Ten newly created DCs are necessary to represent FrameNet in LMF and primarily describe specific properties of frame-semantic frames, e.g., coreness, incorporated semantic argument or transparent meaning.

4.4 UBY-LMF: Limitations of semantic interoperability

The semantic interoperability of a UBY-LMF compliant resource is naturally limited to those domains within UBY-LMF where attributes and

¹⁰<http://www.ilc.cnr.it/EAGLES96/morphsyn/>

¹¹<http://www.ilc.cnr.it/EAGLES96/synlex/>

their values are fleshed out at a fine-grained level and refer to ISOCat DCs. Consequently, attributes of UBY-LMF classes that are string-valued currently limit the semantic interoperability of UBY-LMF. There are three main areas in UBY-LMF where attributes or their values are currently string-valued due to a lack of standardization: first, the names of semantic relations, second, the names of semantic roles, and third, the types of arbitrary semantic classification schemes, as well as the labels used in these schemes.

Semantic Relations Names of semantic relations are values of the attribute `relName` of the `SenseRelation` and `SynsetRelation` class. It is not clear, if the set of possible names for semantic relations is restricted, considering the names of semantic relations found in Wiktionary, OmegaWiki, WordNet and FrameNet. For instance, in OmegaWiki, relations such as “works in a”, “partners with”, “is practiced by a”, “orbits around” (for moons moving around a planet) are listed. This does not fit in the set of classical lexical-semantic relations described by Cruse (1986).

Semantic Roles Names of semantic roles are values of the attribute `semanticRole` of the `SemanticArgument` class. Although DCs for semantic roles have been proposed by Schiffрин and Bunt (2007), we have not entered them to ISOCat, because there is still ongoing work in ISO committees on the standardization of semantic roles.

Semantic classification schemes Finally, names of semantic classification schemes and the labels used in these schemes are values of the attributes `type` and `label` of the `SemanticLabel` class. On the one hand, the `type` attribute covers such divergent classification schemes as Wikipedia categories, WordNet semantic fields (i.e., the lexicographer file names), VerbNet classes, selectional preference schemes or register classification which is present in Wiktionary.

On the other hand, the string-values of the `label` attribute are even more diverse. For instance, the selectional preference schemes in

FrameNet and VerbNet employ different label sets for selectional preferences. While VerbNet makes use of a combination of high-level concepts from the EuroWordNet hierarchy (Kipper-Schuler, 2005), e.g., `+animate` & `+organization`, FrameNet selectional preferences (called ontological types in FrameNet) correspond to synset nodes of WordNet (Ruppenhofer et al., 2010), such as *Message*, *Container*, *Speed*.

5 Discussion

In this section, we first compare UBY-LMF as an LMF instantiation with the TEI guidelines for Dictionaries. For a detailed discussion of related work, please refer to (Eckle-Kohler et al., 2012). Then, we give an outlook of future work on providing alternative formats for UBY-LMF which make the linking to ISOCat accessible to larger communities.

5.1 The TEI guidelines for Dictionaries

The Text Encoding Initiative (TEI) provides guidelines which define standardized representation formats for various kinds of digitized texts in the Digital Humanities. In particular, the TEI provides guidelines for Dictionaries which are in principle applicable to all kinds of lexical resources, but primarily to those that are intended for human use.

Romary (2010a) has suggested to develop the TEI guidelines for Dictionaries into a full LMF instantiation. In their current form, the TEI guidelines cover already core classes of LMF which are required for representing dictionaries. Turning the TEI guidelines for Dictionaries into an LMF instantiation would require first to select the subset of these guidelines that can be mapped to LMF and then to extend the guidelines, e.g., in order to also cover lexical syntax in more detail (Romary, 2010b).

We did not follow this proposal for UBY-LMF, however, because we aimed at representing very heterogeneous resources, which are important for NLP systems, by a single, uniform lexicon model. This particular goal could be achieved mainly due to the high degree of flexibility offered by the LMF standard. Using the TEI guidelines as a starting point seemed to bring about too many

constraints, because many aspects of the lexicon model would have been fixed already.

5.2 Linking to ISOCat – Outlook

The actual representation format for the linking of UBY-LMF terms and ISOCat DCs should be useful both for humans and for machines. For humans, it is helpful to be able to look up the meaning of the terms used in UBY by reading the descriptions of these terms in ISOCat. Machines, on the other hand, can determine the degree of semantic interoperability of two lexical resources by comparing the ISOCat PIDs each lexical resource links to.

At the time of writing this paper, UBY-LMF is the only LMF lexicon model with a publicly accessible Data Category Selection in ISOCat. We hope that DataCategory Selections used in other LMF models will be added to ISOCat soon, because referring to ISOCat DCs is the only way to automatically determine the degree of semantic interoperability of any two LMF-compliant LSRs.

Currently, UBY-LMF provides human-readable links to ISOCat DCs at the schema level (as suggested by Windhouwer and Wright (2012)), i.e., at the level of the UBY-LMF DTD. Pairs of attributes or attribute values and ISOCat DCs are listed in XML comments. As part of future work, we plan to provide additional versions of the UBY-LMF model with machine-readable links to ISOCat, in particular an XSD version as well as an RDFS version for the Semantic Web community.

In addition, we plan to create a metadata instance for UBY-LMF based on the component metadata infrastructure which has been adopted by large communities developing infrastructures of language resources, such as CLARIN and META-SHARE (Broeder et al. (2012), Labropoulou and Desipri (2012), Gavrilidou et al. (2011)). Such a metadata instance specifies the characteristics of UBY-LMF as a language resource, also including the linking of UBY-LMF terms to ISOCat.

6 Conclusion

We have shown how LMF contributes to semantic interoperability of LSRs which is crucial for NLP systems using LSRs. The two key aspects

are first, using a single instantiation of LMF for multiple resources, such as UBY-LMF, and second, establishing a linking between an LMF lexicon model and ISOCat DCs.

Many DCs required for LSRs are still missing in ISOCat, mainly due to ongoing standardization in various areas. Yet, for UBY-LMF, we have fleshed out a Data Category Selection of considerable size and detail in such areas as morphosyntax and subcategorization. We made the UBY-LMF Data Category Selection available in ISOCat as a starting point for further work on standardizing lexical resources according to LMF.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Silvana Hartmann, Michael Matuschek and Christian M. Meyer for their contributions to this work.

References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montreal, Canada.
- Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta.
- Daan Broeder, Dieter van Uytvanck, Maria Gavrilidou, Thorsten Trippel, and Menzo Windhouwer. 2012. Standardizing a component metadata infrastructure. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1387 – 1390, Istanbul, Turkey.
- Nicoletta Calzolari and Monica Monachini. 1996. EAGLES Proposal for Morphosyntactic Standards: in view of a ready-to-use package. In G. Perissinotto, editor, *Research in Humanities Computing*, volume 5, pages 48–64. Oxford University Press, Oxford, UK.
- Christian Chiarcos. 2010. Towards robust multi-tool tagging. an OWL/DL-based approach. In *Proceed-*

- ings of the 48th Annual Meeting of the Association for Computational Linguistic*, pages 659–670. Association for Computational Linguistic.
- D.A. Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 550–560, Avignon, France.
- Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF - A Uniform Format for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 275–282, Istanbul, Turkey.
- Judith Eckle-Kohler. 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*. Logos-Verlag, Berlin, Germany. PhD Thesis.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.
- Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.
- Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Monica Monachini, Francesca Frontini, Gil Francopoulo, Victoria Arranz, and Valérie Mapelli. 2011. A Metadata Schema for the Description of Language Resources (LRs). In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - A Large-Scale Unified Lexical-Semantic Resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, Avignon, France.
- Nancy Ide and James Pustejovsky. 2010. What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong.
- Nancy Ide and James Pustejovsky. 2011. Issues and Strategies for Interoperability among NLP Software Systems and Resources, Tutorial at the 5th International Joint Conference on Natural Language Processing (IJCNLP).
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A Large-scale Classification of English Verbs. *Language Resources and Evaluation*, 42:21–40.
- Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.
- Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1485–1491, Las Palmas, Canary Islands, Spain.
- Penny Labropoulou and Elina Desipri. 2012. Documentation and User Manual of the META-SHARE Metadata Model, <http://www.meta-net.eu>.
- Randolph Quirk. 1980. *A Grammar of contemporary English*. Longman.
- Laurent Romary. 2010a. Standardization of the formal representation of lexical information for NLP. In *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Mouton de Gruyter.
- Laurent Romary. 2010b. Using the TEI framework as a possible serialization for LMF, presentation at the RELISH workshop, August.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petrucci, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice.
- Amanda Schiffrin and Harry Bunt. 2007. Documented compilation of semantic data categories, LIRICS Deliverable D4.3, <http://lirics.loria.fr/documents.html>.
- Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh, and Kiyoaki Shirai. 2009. Query Expansion using LMF-Compliant Lexical Resources. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 145–152, Suntec, Singapore.
- Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. In *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources*, Hong Kong.

Menzo Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3661 – 3664, Istanbul, Turkey.

A standardized general framework for encoding and exchange of corpus annotations: the Linguistic Annotation Framework, LAF

Kerstin Eckart

Institut für maschinelle Sprachverarbeitung

Universität Stuttgart

Pfaffenwaldring 5b, 70569 Stuttgart, Germany

eckartkn@ims.uni-stuttgart.de

Abstract

The Linguistic Annotation Framework, LAF, proposes a generic data model for exchange of linguistic annotations and has recently become an ISO standard (ISO 24612:2012). This paper describes some aspects of LAF, its XML-serialization GrAF and some use-cases related to the framework. While GrAF has already been used as exchange format for corpora with several annotation layers, such as MASC and OANC¹ the generic LAF data model also proved useful as the basis for the design of data structures for a relational database management system.

1 Introduction

Language data and their annotation have been approached from various perspectives. Some approaches are characterized by a layer-wise architecture, where different layers may build on each other, such as morphological, prosodic, syntactic or semantic layers of information. Other approaches concentrate on one specific layer but from various linguistic view points, such as different syntactic frameworks, e.g. dependency and constituency structures. To make reference to these approaches, we speak of vertical vs. horizontal approaches.

In practical work with these approaches, it is often important to pass on, combine or compare annotations for further processing or conjoint querying of different resources. Mostly this includes a lot of work on converting annotation

encodings, as processing tools and query engines expect a specific input format and representation. Each of these proprietary formats is useful and adapted for its respective environment and should therefore be used when dealing with this environment. However adding a generic intermediate format – or: exchange format – reduces the overhead in conversion work, and draws the attention to tasks regarding content-related differences.

The *Linguistic Annotation Framework*, LAF, is an ISO standard (ISO 24612:2012) designed for such purposes. It has been developed by ISO's Technical Committee 37, TC 37, *Terminology and other language and content resources*, Sub Committee 4, *Language resource management*. LAF specifies an abstract data model for an exchange format to represent different kinds of linguistic annotations, and provides an XML pivot format for this model, the *Graph Annotation Format*, GrAF.

Section 2 discusses some objectives and examples for exchange formats. Sections 3 and 4 introduce the LAF data model and some aspects of its XML-serialization GrAF. Sections 5 and 6 give an overview of some existing examples for the application of the LAF/GrAF framework from recent publications and experience.

Throughout this article we make use of the example sentence (1) from the DIRNDL Corpus of German radio news, cf. Section 6 on DIRNDL.

- (1) Die EU-Kommission bezeichnete das Treffen in Rom als Auftakt für einen neuen Dialog zwischen den europäischen Institutionen und der Jugend.

The European Commission characterized

¹<http://www.anc.org>

the meeting in Rome as a starting point for a new dialogue between the European institutions and the young people.

2 Exchange formats

Some important objectives of an exchange format for linguistic annotations are to be

- *generic*, such that different types of annotations can be mapped onto it;
- *theory-independent*, such that no preference for a linguistic theory is reflected in the data model;
- *human-readable*, i.e. not only machine-readable or in binary code, such that inspection and surface error tracking is facilitated;
- *stand-off*, i.e. annotation layers are represented separately from each other and from the primary data, such that also single annotation layers can be exchanged and the primary data can consist of different media.

Moreover, a generic exchange format also constitutes some kind of *interlingua* or intermediate representation as known from machine translation (Hutchins and Somers, 1992) and compiler design (Aho et al., 2007). By providing a mapping from a source format $S_1 = A_1$ into the exchange format X , there is automatically a conversion from A_1 into every target format T_i for which a mapping from X to T_i exists. Figure 1² visualizes this for the representation formats A_i (i in $1..n$).

Let a converter be a routine that takes input of one format and produces output of another. Then by making use of an interlingua, $2n$ converters are needed to provide mappings between n representation formats. Without the interlingua $n^2 - n$ converters would be needed for the direct mappings between the representation formats.

In fields adjacent to linguistic annotation, standards for general data models or formats for information encoding and exchange have been proposed, such as the encoding guidelines for digital text from the *Text Encoding Initiative*, TEI³, also used by libraries and museums. Another example

²Just another realization of the pictures by Ide and Suderman (2007) and Zipser and Romary (2010).

³<http://www.tei-c.org/>

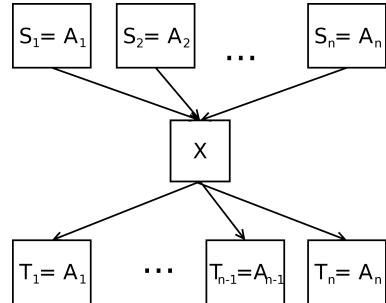


Figure 1: Format conversion with interlingua

is the W3C standard of the *Resource Description Framework*, RDF⁴, to represent linked data in the (Semantic) Web. In the fast growing field of Semantic Web and Linked Open Data⁵, an RDF realization of LAF has been proposed by Cassidy (2010) and recently, POWLA has been presented by Chiarcos (2012) which links a generic representation of linguistic annotations with the standard query utilities of RDF⁶. Thereby a connection with other linguistics-related resources from the Linked Open Data cloud is enabled.

Models for the generic representation of linguistic annotations are for example NXT/NITE (Carletta et al., 2003), PAULA (Dipper, 2005), of which the above mentioned POWLA format is a serialization, and Salt, the data model of the converter framework Pepper, cf. (Zipser and Romary, 2010). Also exchange formats for annotations of specific layers have been proposed as ISO standards, such as the SynAF standard ISO 24615:2010, for a *Syntactic Annotation Framework*, the upcoming standard for a *Morpho-syntactic Annotation Framework*, MAF, and the proposals for components of a *Semantic Annotation Framework*, SemAF.

3 The LAF data model

The Linguistic Annotation Framework proposes a generic graph-based data model to represent linguistic annotations. Figure 2 shows the LAF data model and is cited here, with permission of the authors, from (Ide and Suderman, 2012).

The model contains three parts: The annota-

⁴<http://www.w3.org/RDF/>

⁵cf. (Chiarcos et al., 2012) on Linked Data in Linguistics

⁶POWLA also makes use of OWL/DL to introduce controlled vocabulary and constraints for the data model.

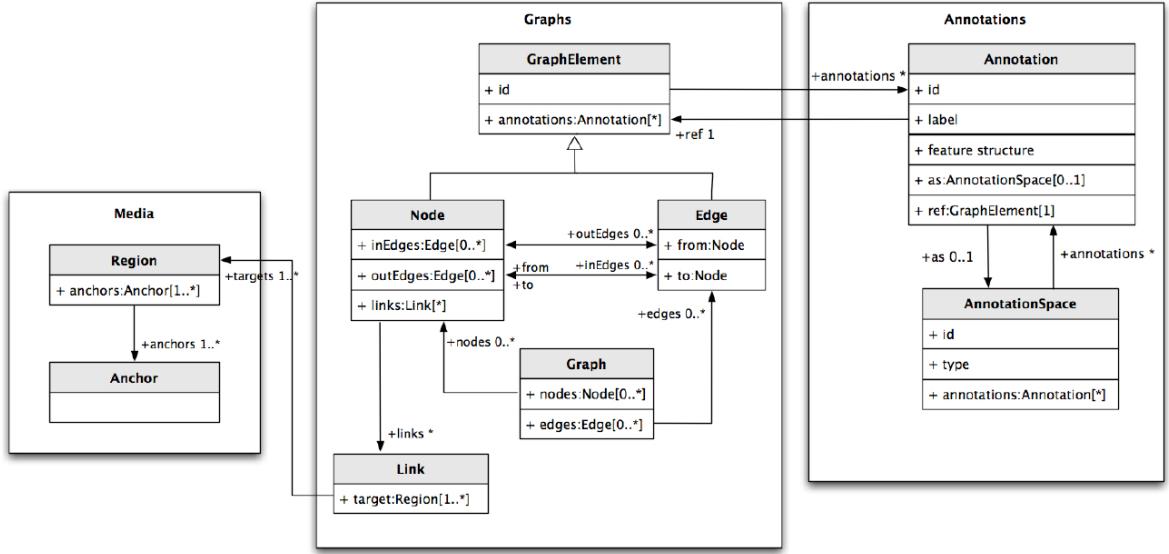


Figure 2: LAF data model by Ide and Suderman (2012), cited with permission

tion part, shown in the rightmost box in Figure 2, which contains data structures to represent annotation labels and complex annotations making use of feature structures⁷. A graphs part, shown in the middle box in Figure 2, allows to represent all annotation structures in a graph based fashion. The graphs consist of nodes and directed edges, where both, nodes and edges can equally be associated with annotations. From the graph structure, links represent the connection to primary data references in the third part which is the media part, shown in the leftmost box in Figure 2.

LAF supports stand-off annotation, which has two effects regarding the data model. Firstly, the primary data is not changed by any annotation. An anchoring mechanism is used to define parts of the primary data to be annotated, i.e. a segmentation which leaves the primary data untouched. The type and encoding of the primary data have to be specified, such that anchors can be interpreted to define virtual positions in between the base units of the encoding.

Let plain text be the medium of sentence (1), UTF-8 its encoding and the sentence the beginning of a primary data set⁸, then Figure 3 visualises the anchors in between the characters for the first part of the sentence.

⁷cf. (Ide and Suderman, 2012) on AnnotationSpace

⁸So the first visualised anchor is 0.

|D|i|e| |E|U|-|K|o|m|m|i|s|s|i|o|n|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
1

Figure 3: LAF anchoring mechanism

Regions of different granularity can be defined by one or more anchors, e.g. anchors 0 and 3 denote the region *Die*, while anchors 15 and 16 denote the last *o* in *Kommission*. Regions in other media make use of other types of anchors, such as timestamps for audio data, coordinates for image data, etc. The defined regions can be referenced by the links from the annotation graphs, and therefore be related to an annotation.

The second aspect of stand-off annotation is that annotations are layered one on top of the other, making reference to the layers beneath them. In LAF this is represented by edges between nodes belonging to different annotation layers. A node from a syntactic annotation does probably not link the related regions directly, but has edges to one or more nodes from a morpho-syntactic annotation of the respective regions.

LAF also proposes an architecture for corpus resources containing primary data, annotations and metadata, by dividing the data conceptually into different types of documents and headers.

Primary data documents can contain any type of primary data, e.g. text, audio or video data.

As these documents provide the reference point for the anchors, they must not be changed. Annotation documents contain the annotation graphs and the segmentations related to the primary data. Ide and Suderman (2012) state that creating independent segmentation documents, i.e. annotation documents, that only contain segmentation information, is recommended, when the same segmentation is directly referenced from different annotation documents.

Metadata is represented via headers, i.e. a resource header for the combined resource, a header for each primary data document⁹, and a header for each annotation document.

A central concept of LAF is the separation of representation and content. A given annotation content, e.g. a syntactic constituent tree according to a specific linguistic theory or annotation scheme¹⁰, can be represented in an XML format, like TIGER-XML, cf. (König et al., 2003) or in a bracketing format like that of the Penn Treebank, cf. (Bies et al., 1995), while still containing the same annotation content. Therefore LAF separates representation and content, and provides an exchange format with respect to representation. This is also often called syntactic interoperability with respect to different annotation formats. Its complement, semantic interoperability, i.e. the (linguistic) specification and the exchange of data categories and concepts is not in the scope of LAF, but is tackled by other frameworks such as the ISO standard on a Data Category Registry for language resources, ISO 12620:2009 and its implementation ISOcat¹¹. Nevertheless LAF provides structures in the resource header and the annotation documents, where reference can be made to the utilized annotation schemes or to single data categories (Ide and Suderman, 2012).

4 The GrAF XML format

The Graph Annotation Format is an XML-serialization of the LAF data model and architecture,

⁹Represented as a stand-alone header because the primary data document should not be changed and should only contain primary data.

¹⁰For an example tree see the unfilled nodes and continuous lines in Figure 8.

¹¹<http://www.isocat.org/>

and has been proposed as a part of the standard.¹².

GrAF provides XML structures to represent annotations, as well as for headers containing metadata and data on corpus organisation, as presented by Ide and Suderman (2012).

An example of the corpus organisation information is the `fileStruct` element in the resource header, which mirrors the directory structure and indicates the root directory of the resource. Other metadata such as a source description containing the publisher, and text classification elements referring to the domain of the primary data are included in the primary data document header. Relevant for processing is the information on dependencies and anchor types which can be found in the resource header and in the annotation document headers. It states, on which files an annotation document is dependent, i.e. which files have to be present, so that the respective annotation can be correctly processed. The utilized anchor type depends on the medium and encoding of the respective primary data document. The media used in the resource are listed in the resource header and the anchor types are related to the respective medium they apply to. With the LAF anchor mechanism, which abstracts over the different anchor types, anchors in GrAF can be consistently represented, the information on the respective anchor type being accessible via the resource header, cf. Ide and Suderman (2012).

The graph structure specified in LAF is instantiated in GrAF by `node`, `edge` and `link` elements and the annotation structure by `a` elements containing a label and/or a feature structure element `fs`. An annotation element `a` makes reference to an `edge` or `node` element and therefore attaches annotation features to the respective graph element. A contained feature structure can be complex, as defined in ISO 24610-1:2006, the ISO standard on *Feature Structure Representation* (FSR). Edges exist between a start and an end node¹³, and nodes which reference regions in the primary data contain `link` elements, stating which regions of the primary data are referenced by the respective node. Primary data regions as defined in the LAF data model, are specified by

¹²The schema and an API can be found at <http://www.xces.org/ns/GrAF/1.0/>

¹³Therefore GrAF implements directed graphs.

a particular number of anchors and a specific anchor type, both depending on the medium type and encoding of the primary data. In GrAF they are represented by `region` elements. Figures 4 to 7 show parts of a GrAF standoff annotation for sentence (1) and anchors as specified in Figure 3. The same annotation is displayed in Figure 8, by the unfilled and dark gray nodes and the respective edges.

```
<!-- Die -->
<region xml:id="r1" anchors="0 3"/>
<!-- EU -->
<region xml:id="r2" anchors="4 6"/>
<!-- - -->
<region xml:id="r3" anchors="6 7"/>
<!-- Kommission -->
<region xml:id="r4" anchors="7 17"/>
```

Figure 4: Token segmentation

Figure 4 shows a part of a segmentation into regions. Note, that the segmentation does not cover every base unit, as the first whitespace is not part of any region, and that this segmentation is not the most granular one, but fits the annotation.

```
<node xml:id="n1"> <!-- Die -->
<link targets="r1"/>
</node>
<node xml:id="n2"> <!-- EU-Kommission -->
<link targets="r2 r3 r4"/>
</node>
<a label="tok" ref="n1">
<fs>
<f name="word" value="die"/>
<f name="pos" value="D[std]"/>
<f name="seq" value="1"/>
</fs>
</a>
<a label="tok" ref="n2">
<fs>
<f name="word" value="EU-Kommission"/>
<f name="pos" value="N[comm]"/>
<f name="seq" value="2"/>
</fs>
</a>
<a>
</a>
```

Figure 5: Nodes with part-of-speech annotation

The first annotation layer which builds on the segmentation from Figure 4 is displayed in Figure 5. Two nodes are shown, where node `n2` references more than one region to represent the token *EU-Kommission*. The annotations refer to these nodes and contain feature structures denoting the part-of-speech of the token (`pos`), the token string (`word`) and the position of the token in the token sequence composing a sentence (`seq`).

```
<node xml:id="n3"/> <!-- NP -->
<edge xml:id="e1" from="n3" to="n2"/>
<node xml:id="n4"/> <!-- DPx[std] -->
<edge xml:id="e2" from="n4" to="n1"/>
<edge xml:id="e3" from="n4" to="n3"/>
<node xml:id="n5"/> <!-- DP[std] -->
<edge xml:id="e4" from="n5" to="n4"/>
<a label="syn" ref="n3">
<fs>
<f name="pos" value="NP"/>
</fs>
</a>
<a label="syn" ref="n4">
<fs>
<f name="pos" value="DPx[std]"/>
</fs>
</a>
<a label="syn" ref="n5">
<fs>
<f name="pos" value="DP[std]"/>
</fs>
</a>
```

Figure 6: Nodes denoting syntactic constituents

The syntactic annotation layer in Figure 6 builds on the token layer in Figure 5, by making reference to the nodes `n2` and `n1` with the edges `e1` and `e2`. Nodes `n3`, `n4` and `n5` denote syntactic constituents according to the German LFG grammar by Rohrer and Forst (2006).

```
<node xml:id="n6"/>
<edge xml:id="e5" from="n6" to="n5"/>
<a label="is" ref="n6">
<fs>
<f name="is" value="UNUSED-KNOWN"/>
</fs>
</a>
```

Figure 7: Node denoting information status of a syntactic phrase

The last layer, exemplified here in Figure 7, is an information status annotation according to Rister et al. (2010). Information status is annotated to a phrase and therefore related to the nodes from Figure 6. The node labeled `UNUSED-KNOWN` is directly related to node `n5`, which stands for the determiner phrase *Die EU-Kommission*.

5 Use-cases for LAF/GrAF representation and exchange

In this section we give a short overview on examples of use-cases where GrAF has been applied as a representation and exchange format. While some of the examples reference an earlier version of the GrAF schema related to a pre-standard LAF draft, the concepts still apply.

Two multi-layer annotated corpora related to the *American National Corpus*, ANC, have been distributed in the GrAF XML format. The *Open American National Corpus*, OANC¹⁴, which is annotated for part-of-speech as well as noun and verb chunks, and the *Manually Annotated Sub-Corpus*, MASC, cf. (Ide et al., 2010), which is annotated for various annotation layers and also with different annotations for some layers. Both corpora exemplify the corpus architecture supported by GrAF and can be seen as a reference example for GrAF format encoding. Interfaces for ANC data in GrAF exist for frameworks like UIMA and GATE, and the GrAF representation of MASC has been input to a conversion into POWLA (Chiarcos, 2012).

Distiawan and Manurung (2012) proposed a speech recognition web service within the Language Grid framework¹⁵ where the output is encoded in a pre-standard GrAF format aiming at the combination of process interoperability provided by the Language Grid with annotation interoperability provided by the LAF/GrAF framework. Here the primary data is audio or video data and the anchors defining the segments to be annotated therefore are timestamps, providing for flexible annotation/transcription of the primary data. An additional advantage described by Distiawan and Manurung (2012), is that multiple segmentation results can be conjointly represented, which allows that (the n-best) alternative recognition results can be provided to the user. An objective of the described web service approach was also to get information back from the user for retraining the actual recognizer: this setup might encourage the user to provide feedback, as he or she does not have to propose a correction, but just has to choose from alternatives.

Hayashi et al. (2010) present a wrapper architecture for GrAF-based encoding of dependency structures produced by different dependency parsers. The input to a wrapper is therefore a proprietary format from a dependency parser, such as the XML formats of the Stanford English parser (de Marneffe and Manning, 2008) or the Japanese dependency parser CaboCha (Kudo and Matsumoto, 2002). The implemented wrappers

make use of a sub-schema of GrAF, proposing a general representation for dependency structures, where word-like tokens and phrase-like chunks can be equally parts of a dependency relation. As this proposes a subtyping of GrAF for a specific type of annotation, it is close to annotation-layer-based instantiations such as SynAF; obviously, this work also provided feedback to the respective ISO groups.

6 Use-cases for LAF/GrAF-based data structures

The B3-Database, B3DB, cf. (Eckart et al., 2010) is a database to collect and relate data which occur in the workflow of a linguistic project. This comprises primary data, annotations and information about how the annotations were produced, e.g. taking tool versions or annotation schemes into account. The annotations are stored in the database as objects of manually created information or tool output. Then each of these objects can be decomposed and mapped upon one or more graph structures inside the database. Some important design decisions on annotation handling in the database are the following:

All information is annotation All information from the original annotation file is kept as annotation, e.g. identifiers which are unique within the file or technical data such as processing time.

All annotations are graphs Each annotation is mapped to a set of nodes and edges, or to one of them; sequential annotations, as in prosodic transcriptions, are mapped onto trivial graphs.

Graph representations are static Whenever an error is found in a graph representation, or when it should be changed in any way, the complete graph has to be rebuilt.

The B3DB is implemented as a PostgreSQL¹⁶ relational database system. Its data structures for the graph representations are based on the LAF data model and the respective GrAF structures. In the database, tables are implemented for nodes, edges, links, regions, annotations, feature structures, features and different kinds of name/value pairs. On top of that, nodes, edges and annotations are typed, which helps to distinguish the different types of annotations that exist due to the

¹⁴<http://www.anc.org/OANC/>

¹⁵<http://langgrid.org>

¹⁶<http://www.postgresql.org/>

first design decision stated above. A table containing the transitive closure of the graphs helps to facilitate queries on graph structures where basic SQL does not provide for recursion. Here the third design decision comes into play. As the graphs may not be changed dynamically, the closure remains static and avoids overhead from closure updates. That way efficient queries can be processed on the database structures, and a GrAF representation of the graph structures can be easily extracted from the database system.

The database has proven useful for different resources, such as DIRNDL, cf. (Eckart et al., 2012), a German radio news corpus based on two primary data sets. For this corpus there had been two parallel workflows on two closely related primary data sets: a textual version of the radio news, probably the manuscript read by the speakers, and the audio version of the actually spoken news text. The two primary data sets deviated slightly, due to slips of the tongue or minimally modified statements. However, the phonetic processing could only be applied to the primary data based on the audio files, while the syntactic and semantic processing could only be applied to the textual version, cf. (Eckart et al., 2012).

Automatic transcription using forced alignment (Rapp, 1995) was utilized with the audio version, and it was manually annotated with prosodic phrase boundaries and pitch accents according to GToBIS (Mayer, 1995). The textual version was parsed by an LFG-parser¹⁷ with a German grammar by Rohrer and Forst (2006), and the constituency trees were manually annotated for information status according to the annotation scheme by Riester et al. (2010).

To relate and conjointly query the two annotation sets, a linking of the two token layers was semi-automatically created and represented by edges between the two graph representations in the database. Here the LAF edges, which serve as a connection between annotation layers could also bridge the gap between the slightly deviating data sets.

Figure 8 shows an excerpt from the DIRNDL annotations of sentence (1). The information status graph is denoted by the dark gray nodes, the prosody graph by the light gray nodes, and the

¹⁷XLE parser with Lexical Functional Grammar.

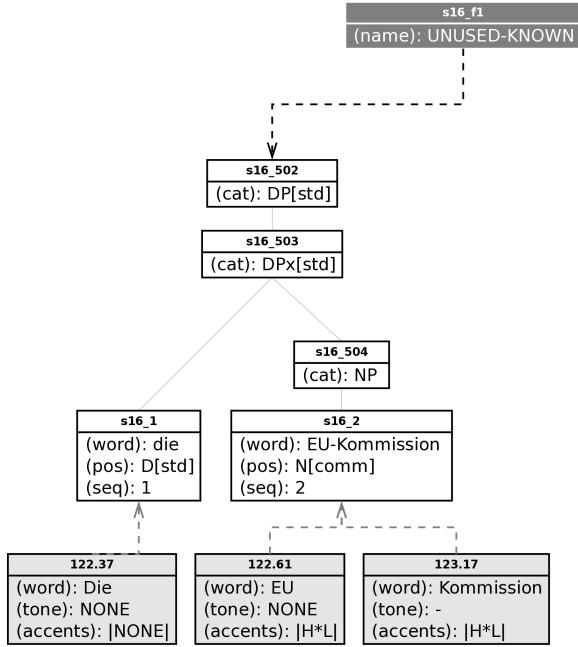


Figure 8: Annotation subgraph for sentence (1) in DIRNDL

syntactic constituent tree by the unfilled nodes and the continuous lines¹⁸. The dashed lines denote edges between the different annotation graphs. While *EU-Kommission* is one token in the syntactic graph, it was divided into two tokens for the prosodic processing¹⁹. Therefore the two tokens from the prosodic graph are related to the single token of the syntactic graph. While in this case a fine-grained segmentation of the transcription as a common document would have done the same, the cases of slips of the tongue or statement modifications found in the DIRNDL sample make clear that there are actually two primary data sets in the corpus²⁰.

The edges in the database are also directed. Nevertheless this has no impact on the query, as the information extraction can follow an edge from source to target or from target to source. Therefore all directions between information status, syntax and prosody can be equally processed.

¹⁸DP: determiner phrase, NP: noun phrase; D[std]: determiner, N[comm]: noun; |H*L|: falling accent; -: intermediate phrase boundary

¹⁹The hyphen is omitted as it is not pronounced.

²⁰While one layer could have been marked as the main primary data set and the deviations as a separate annotation layer of 'corrections', we did not want to introduce a weighting wrt the different primary data sets.

As audio data is not directly stored in the database, we decided to represent the timestamps of the prosodic annotation as annotations. As can be seen in Figure 8, each prosody node is annotated with a timestamp denoting the end of the utterance of the respective token in the audio files. Figure 9 however shows both encodings in GrAF, the timestamps as annotations to the phonetic transcription, and as anchors to the database-external audio files.

```
<!-- timestamp annotation -->
<region xml:id="r2" anchors="4 6"/>
<node xml:id="n12">
  <link targets="r2">
</node>
<a label="prosody" ref="n12">
  <fs>
    <f name="timestamp" value="122.61">
    ...
  </fs>
</a>

<!-- timestamp anchors -->
<region xml:id="r52" anchors="122.37 122.61"/>
```

Figure 9: Timestamps as annotations, and as anchors

Other use-cases related to the B3DB have been described by Haselbach et al. (2012) and Eckart et al. (2010). They take among other things different syntactic dependency structures into account. Figure 10 shows abstract dependency relations²¹ for the part *Die EU-Kommission bezeichnete das Treffen* of sentence (1).



Figure 10: Dependency relations

Note that in the LAF concept of nodes, a node which has a direct link to a region, cannot be the starting point of an edge. So in case no intermediate layer is given, this has to be handled e.g. by introducing additional nodes as start and end points of a dependency relation, cf. Figure 11²².

Hayashi et al. (2010) propose another representation of dependency structures in GrAF, where an additional dependency node is introduced. This additional node is the start node of

²¹SPEC: specifier, SUBJ: subject, OA: object, accusative case

²²This example applies an alternative segmentation to the one given in the examples in Chapter 4.

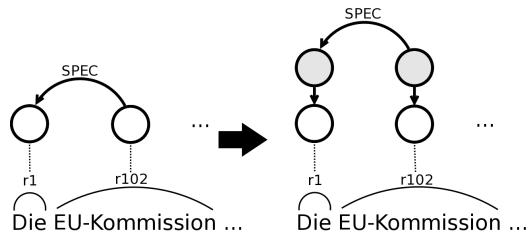


Figure 11: Introducing additional dependency nodes

two edges, one to the node representing the head and one to the node representing the dependent. The annotation elements of the dependency nodes contain information regarding the dependency relation. Hayashi et al. (2010) argue that a single edge between head and dependent nodes introduces additional semantics by combining the from and to attributes of an edge with the head and dependent notions from dependency structure. In our understanding, this does not deviate from the usual disposition of GrAF, as edges in a constituency tree also carry semantics, and also some dependency representations tend to introduce edges from dependent to head. However, all these cases can directly be represented in GrAF. Nevertheless the approach of Hayashi et al. (2010) goes beyond the idea of simply representing an existent annotation format in GrAF and aims at a common layer representation for dependency structures. Therefore it is important to choose a representation which makes the common semantics explicit.

7 Conclusion

In this paper we reported on the Linguistic Annotation Framework, LAF, and some use-cases of its serialization and data model. Since the early versions of the LAF proposal, the model itself and related use-cases have been discussed and implemented, extensions (Kountz et al., 2008) and derivations (Hayashi et al., 2010) have been proposed and upcoming fields have been included (Cassidy, 2010; Chiarcos, 2012). In 2012 LAF has become an ISO standard. As Chiarcos (2012) concludes, the different but related approaches should be utilized in a complementary fashion, making use of the respective properties and advantages coming with the respective data models and serializations.

An important concept of LAF is that it embodies clear distinctions. Distiawan and Manurung (2012) sum this up very nicely as separating 'user format from exchange format', 'primary data from annotation' and 'representation from content'. In conclusion, one could also add the separation of the 'data model from its serialization', and be prepared for more LAF/GrAF related use-cases to come.

References

- Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. 2007. *Compilers: Principles, Techniques, and Tools (2nd Edition)*. Addison Wesley, August.
- Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*.
- Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voermann. 2003. The NITE XML toolkit: Flexible annotation for multimodal language data. *Behavior Research Methods*, 35:353–363.
- Steve Cassidy. 2010. An RDF realisation of LAF in the DADA annotation server. In *Proceedings of ISA-5*, Hong Kong, January.
- Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.
- Christian Chiarcos. 2012. A generic formalism to represent linguistic corpora in RDF and OWL/DL. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.
- Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.
- Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Berliner XML Tage*, pages 39 – 50.
- Bayu Distiawan and Ruli Manurung. 2012. A GrAF-compliant indonesian speech recognition web service on the language grid for transcription crowdsourcing. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 67–74, Jeju, Republic of Korea, July. Association for Computational Linguistics.
- Kerstin Eckart, Kurt Eberle, and Ulrich Heid. 2010. An infrastructure for more reliable corpus analysis. In *Proceedings of the Workshop on Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation (LREC'10)*, pages 8–14, Valletta, Malta, May.
- Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg.
- Boris Haselbach, Wolfgang Seeker, and Kerstin Eckart. 2012. German "nach"-particle verbs in semantic theory and corpus data. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.
- Yoshihiko Hayashi, Thierry Declerck, and Chiharu Narawa. 2010. LAF/GrAF-grounded representation of dependency structures. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May.
- William J. Hutchins and Harold L. Somers. 1992. *An introduction to machine translation*. Academic Press.
- Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.
- Nancy Ide and Keith Suderman. 2012. A model for linguistic resource description. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 57–66, Jeju, Republic of Korea, July. Association for Computational Linguistics.
- Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Uppsala, Sweden, July. Association for Computational Linguistics.
- Esther König, Wolfgang Lezius, and Holger Voermann, 2003. *TIGERSearch 2.1 User's Manual. Chapter V - The TIGER-XML treebank encoding format*. IMS, Universität Stuttgart.
- Manuel Kountz, Ulrich Heid, and Kerstin Eckart. 2008. A LAF/GrAF-based encoding scheme for underspecified representations of dependency structures. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.
- Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of CoNLL-2002*, pages 63–69. Taipei, Taiwan.

- Jörg Mayer. 1995. Transcription of German Intonation. The Stuttgart System. ms.
- Stefan Rapp. 1995. Automatic phonemic transcription and linguistic annotation from known text with hidden markov models – an aligner for German. In *Proceedings of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry" (Russia)*.
- Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information structures. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722, Valletta, Malta.
- Christian Rohrer and Martin Forst. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.
- Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, Malta.

Standard for morphosyntactic and syntactic corpus annotation: The Morphosyntactic and the Syntactic Annotation Framework, MAF and SynAF

Laurent Romary
Humboldt University of Berlin

Abstract

This talk is about the standards for morpho-syntactic and syntactic corpus annotation: MAF (Morpho-syntactic Annotation Framework, ISO/FDIS 24611) and SynAF (Syntactic Annotation Framework, ISO 24615:2010). Both standards complement each other and are closely related. In contrast to MAF, which describes features such as part of speech, morphological and grammatical features, SynAF describes relations between single words, and how words are arranged with each other and connected to build phrases and sentences. The talk presents an overview of the current state of both standards.

Towards standardized lexical semantic corpus annotation: Components of the Semantic Annotation Framework, SemAF

Tibor Kiss

Ruhr University Bochum

Abstract

Spatial annotations form part of the Semantic Annotation Framework (SemAF). The current development SemAF-Space (ISO 24617-7) provides a formal specification but does not provide annotation guidelines. In my talk I will compare this approach with the approach developed for the annotation of preposition senses in Müller et al. (2011), where the annotation guidelines form the annotation specification.

Towards standards for corpus query: Work on a Lingua Franca for corpus query

Elena Frick
IDS Mannheim

Piotr Bański
University of Warsaw

Andreas Witt
IDS Mannheim

Abstract

In this presentation, we report about the ongoing work on the development of a standard for corpus query languages. This work takes place in the context of the ISO TC37/SC4 WG6 activity on the suggested work item proposal „Corpus Query Lingua Franca“ (Bański and Witt, 2011). We have collected a set of requirements on a corpus query language motivated by the needs of linguists and we will present the initial ideas for the First Committee Draft.

Getting involved into language resource standardization: The map of standards and ways to contribute to ongoing work

Gottfried Herzog
Deutsches Institut für Normung e.V. (DIN)

Abstract

This contribution addresses the questions:

- Why International Standards?
- What is ISO?
- How does ISO develop standards?

The focus are the Key principles in standard development:

1. ISO standards respond to a need in the market
2. ISO standards are based on global expert opinion
3. ISO standards are developed through a multi-stakeholder process
4. ISO standards are based on a consensus

Finally, there will be some practical hints how to get involved in developing ISO standards.