# CONTENTS

# 1 | ABSTRACTS


9

# "What happened to . . . ?" Entity-based Timeline Extraction

Tommaso Caselli
VU Amsterdam
t.caselli@gmail.com

Antske Fokkens
VU Amsterdam
antske.fokkens@vu.nl

Roser Morante
VU Amsterdam
r.morantevallejo@vu.nl

Piek Vossen
VU Amsterdam
p.t.j.m.vossen@vu.nl

We present the VUA-Timeline module for extracting cross-document timelines. The system has been developed in the framework of the SemEval-2015 Task 4 TimeLine: Cross-document event ordering ([1]). Innovative aspects of this task are cross-document information extraction and centering of timelines around entities which requires a careful handling of nominal and event coreference information. Timeline Extraction is a complex task composed by a set of subtasks: named entity recognition, event detection and classification, coreference resolution of entities and events, event factuality, temporal expression recognition and normalization, and extraction of temporal relations. The VUA-Timeline system has been developed as an additional module of the NewsReader NLP pipeline, which consists of tools which provide state-of-the-art results carrying out the subtasks mentioned above. The output of this pipeline is a rich representation including events, their participants, coreference relations, temporal and causal relations and links to external references such as DBpedia. We extract cross-document timelines concerning specific entities in two steps. First, we identify timelines within documents, selecting events that involve the entity in question and filtering out events that are speculative. These events are placed on a timeline based on normalized temporal indications (if present) and explicit temporal relations. In the second step, we merge document specific timelines to one cross-document timeline by applying cross-document event coreference and comparing normalized times. We present the performance of our system on the SemEval data which will be released on December 10th.

[1] http://alt.qcri.org/semeval2015/task4/

# A Corpus of Machine Translation Errors for English-into-Dutch Language Pair

Arda Tezcan
Ghent University
arda.tezcan@ugent.be

Lieve Macken
Ghent University
lieve.macken@ugent.be

Veronique Hoste
Ghent University
veronique.hoste@ugent.be

We present a taxonomy of typical machine translation errors within the scope of the Smart Computer-Aided Translation Environment (SCATE) project and the efforts for building a corpus of typical machine translation errors for English-Dutch. We discuss the annotation scheme, the annotation guidelines, the annotation tool and the inter-annotator results. The corpus consists of source sentences in English, MT output in which errors are annotated, post-edited MT output and reference translations in Dutch. Adequacy and fluency errors, being the two main error categories in the taxonomy, refer to the errors in the target text only and the errors regarding the transfer of content from source text to target text, respectively. Based on this structure, fluency errors are labelled on the MT output only, while adequacy errors are labelled both on the source sentences and the MT output and are linked to each other. This error taxonomy and the corpus aims to serve as a basic resource for analysing errors made by MT systems, for assessing the impact of different types of errors on post-editing effort and for building quality estimation tools for MT, especially for the English-into-Dutch language pair.

# A Multi-Strategy Approach for Lexicalizing Linked Open Data

Rivindu Perera
Auckland University of Technology
rivindu.perera@aut.ac.nz

Parma Nand
AUT
parma.nand@aut.ac.nz

In recent years, there has been an increasing interest in lexicalizing Linked Data. Linked Data represent the structured content of text harvested from unstructured text in web. Lexicalization is used to convert this structured form into respective natural surface text form. We present a multi-strategy approach that can generate patterns to lexicalize Linked Data stored in DBpedia. Our approach is based on three key components; Wikipedia based pattern mining, verb frame based pattern mining and a rule based module to assign patterns. The Wikipedia based pattern miner uses large collection of Wikipedia texts and extract patterns thorough a relation extraction process. The verb frame based module uses WordNet and VerbNet to extract verb frames which can be converted to lexicalization patterns. In addition to this we have also utilized a rule based module to assign a predetermined pattern to left over Linked Data triples. Each extracted pattern is ranked using a scoring system and enriched with additional properties which are useful later when searching for a matching lexicalization pattern for a triple. The identified patterns are categorized according to the DBpedia ontology class hierarchy and stored in an indexed database. The framework achieved 70.36% accuracy and a Mean Reciprocal Rank value of 0.72 for five DBpedia ontology classes generating 101 accurate lexicalization patterns.

# A Parsed Corpus of Historical Low German

Mariya Koleva
Ghent University
MariyaStoyanova.Koleva@UGent.be

Melissa Farasyn
Ghent University
melissa.farasyn@ugent.be

Anne Breitbarth
Ghent University
Anne.Breitbarth@UGent.be

Veronique Hoste
Ghent University
veronique.hoste@ugent.be

We report on a study on the construction of a part-of-speech tagger for Middle Low German (MLG) - a group of West Germanic dialects spoken and written in northern Germany c.1250 - 1600. The work is within the Corpus of Historical Low German project, which aims to create a resource for the diachronic study of MLG syntax across five written dialects in the form of a POS-tagged and syntactically annotated corpus. The source data come from manually transcribed manuscripts and prints. Tool creation for historical language varieties is often restricted by limited training data. For our initial experiments, we use legal texts from 1330-1340 from Oldenburg (15,464 tokens) and Soest (8,224) tokens. The two dialects - Lower North Saxon and Westphalian (respectively) - are linguistically similar and our texts share a 25% token overlap. Although there existed a certain standardization in MLG varieties, the corpus texts show a remarkable amount of orthographic variation which increases the data sparsity. Since the accuracy of the tagger is sensitive to the orthographic uniformity, we worked out a set of preprocessing rules to define and normalize certain characters and character clusters. The changes are based on relative character or cluster frequency, taking the environment of the phenomenon into account. After normalization, we train several language models using different combinations of features on the texts of each locality and their combination. Here we present the preliminary results of the POS-tagging experiments and we elaborate on the effect normalization has on the tagging accuracy.

# A general purpose spelling correction tool in the post-OCR error correction task: comparative evaluation and feasibility study

Kalliopi Zervanou
Utrecht University
k.a.zervanou@uu.nl

Job Tiel Groenestege
Gridline
job@gridline.nl

Dennis de Vries
Gridline
dennis@gridline.nl

Tigran Spaan
Gridline
tigran@gridline.nl

Wouter Klein
Utrecht University
W.Klein@uu.nl

Jelle van der Ster
Utrecht University
jellevdster@gmail.com

Peter van den Hooff
Utrecht University
P.C.vandenHooff@uu.nl

Frans Wiering
Utrecht University
F.Wiering@uu.nl

Toine Pieters
Utrecht University
t.pieters@uu.nl

The digitisation process for text documents that were not born digital typically entails document image scanning, often followed by optical character recognition (OCR), so as to make the text machine readable for subsequent information processing. Despite the progress in OCR systems software, OCR text output still often contains so much error, that both human readability and computer processing is impaired. That is especially the case for old documents, where page quality is degraded and font style often follows obsolete typographic conventions. A solution to this problem is provided by post-OCR error correction methods which often combine corpus statistics with lexical resources (Reynaert, 2014a) and/or additional linguistic information (Baron et al. 2009) with human rule pattern input (Vobl et al. 2014). In this work, we investigate the feasibility of using a general purpose spelling error correction application, the Gridline Taalserver (TM) toolsuite, the knowledge resources of which have been optimised for post-OCR error correction. For this purpose, we comparatively evaluate the Gridline Taalserver toolsuite to the latest version of TICCLops (Reynaert, 2014a), a purpose built post-OCR error correction tool, using two evaluation corpora, the 1800s EDBO DPO35 OCR gold standard corpus (Reynaert, 2014b) and a subcorpus of 1950s newspapers from the VU-DNC corpus (VU DNC). The results of our comparative evaluation show that both approaches present advantages in reducing error and could be applied in combination.

# A new automatic spelling correction model to improve parsability of noisy content

Rob van der Goot
RuG
r.van.der.goot@rug.nl

Gertjan van Noord
RuG
g.j.m.van.noord@rug.nl

To improve the parsing of noisy content (here: Tweets), a new automatic spelling correction model is proposed that normalizes input before handing it to the parser. Spelling correction is done in three steps: finding errors, generating solutions and ranking the generated solutions. Finding errors Normally, finding errors is done by comparing tokens against a dictionary. A more elaborate approach uses n-grams to estimate word probabilities. If the probabilities of n-grams around a certain word are very low, this probably indicates an error. Generating solutions Most spelling correctors use a modified version of the edit distance to find the correct words/phrases. Aspell includes an efficient and effective algorithm for this, which combines the normal edit distance with the double metaphone algorithm (L. Philips, 2000). The generation process should be adapted to the domain of application. Because our test data originates from Twitter, we assume that people tend to use shorter variants of words. This justifies a modification of the costs of insertion and deletion in the Aspell code. Ranking solutions In previous work ranking is based on edit distance and context probability(ie. M Schierle, 2008). To utilize more estimators for the ranking, a logistic regression model is trained. The features used are: - Edit distance: as calculated by Aspell, using the modified costs. - Different n-gram probabilities: uni- bi- and tri-gram probabilities are included. - Parse probability: The parse probability of the best parse of the Stanford parser. In the presentation, we will present the model, as well as experimental results.

# A semi-supervised method for cleanup of a User-Generated Content corpus

Enrique Manjavacas
Freie Universität Berlin
enrique.manjavacas@gmail.com

Ben Verhoeven
CLiPS, University of Antwerp
ben.verhoeven@uantwerpen.be

Walter Daelemans
CLiPS, University of Antwerp
walter.daelemans@uantwerpen.be

The construction of corpora specifically tailored to tackle certain tasks is an inevitable step in the workflow of many NLP projects. These corpora are often harvested from sources with user-generated content, such as wikis, tweets, blogs, etc. However, this online content, though useful, typically includes a large part of undesired text that influences the quality of the resulting corpus and may need to be filtered out carefully. In an ongoing research involving a large corpus of blog texts (200 million words) we explore and evaluate a principle of pre-filtering such undesired content. We apply a kernel density estimation method in which a likelihood is assigned to each blog. The estimation is based on general features of blogs, such as the type/token ratio, average post length or post frequency. Blogs with lower estimates can be interpreted as outliers in the total feature space and prioritized in a subsequent sampling for detection of noisy data. The computed estimates are thus used as a guideline in the filtering process under the working assumption that undesired blogs are represented by extreme datapoints in the dataset. Finally a manual inspection of the extracted outliers is conducted in order to test both the assumption of the rareness of noisy data and the reliability of kernel density estimates as a method for outlier detection in such a setup.

# A sentiment journey from the 1950s until 2010 - Extracting negative opinions from political articles in De Telegraaf

Carlotta Casamassima
Network Institute, Vrije Universiteit Amsterdam
c.casamassima@student.vu.nl

Antske Fokkens
Network Institute, Vrije Universiteit Amsterdam
antske.fokkens@vu.nl

Laura Hollink
Network Institute, Vrije Universiteit Amsterdam
l.hollink@vu.nl

Wouter van Atteveldt
Network Institute, Vrije Universiteit Amsterdam
w.h.van.atteveldt@vu.nl

Annick van der Peet
Network Institute, Vrije Universiteit Amsterdam
a.e.h.vander.peet@student.vu.nl

News coverage of Dutch politics has greatly changed since the 1950s to the present day. Theory on the mediatization of politics describes a growing influence of the "media logic" of journalistic news values and market demands. This results in the news becoming more personalized, more negative, and focusing more on conflict than on political substance. In previous work, we have analysed all 1.7 million political articles that appeared in De Telegraaf, showing that individualization of political news has substantially increased. In this talk, we present the next step where we investigate whether the use of negative opinions has also changed in the Netherlands. We use existing sentiment analysis and opinion mining techniques as a baseline and apply this to our dataset. The performances of the tools are analyzed on a selected set of the data for intrinsic evaluation. We furthermore present a crowd-sourcing task that can be applied to a larger set of our data. These annotations can directly answer questions regarding changes of negativity in the news and be used for extrinsic evaluation of our tools. In addition to answering this specific question, we aim to address more general questions on the reliability and impact of sentiment analysis tools to be used in research on communication science.

# Adpostion stranding in Dutch. Omdat we er dol op zijn!

Liesbeth Augustinus
CCL, KU Leuven
liesbeth@ccl.kuleuven.be

Frank van Eynde
CCL, KU Leuven
frank@ccl.kuleuven.be

Dutch adpositions may be stranded, as shown in the following examples: (1) Daar had ik nog niet aan gedacht. (2) Dat ik daar nog niet aan gedacht had! In (1), the pronominal complement 'daar' (there) is realised in the Vorfeld, while in (2) it occurs in the Mittelfeld. Stranding is not (always) obligatory, however. In (3) the R-pronoun is realised in situ: (3) Dat ik nog steeds daar aan moet denken. For a descriptive and/or theoretical account of adposition stranding in Dutch, see amongst others Van Riemsdijk (1978), Beeken (1991), Haeseryn et al. (1997), and Broekhuis (2013). Since the majority of previous studies is based on introspection and elicitation, it is interesting to compare those findings to corpus data. The research presented here is a corpus-based account of adposition stranding in Dutch, relying on on data from two Dutch treebanks (CGN and LASSY). On the one hand, the corpus study allows us to investigate which adpositions can be stranded. On the other hand, the corpus data show variation regarding the type of complement (e.g. nominal versus pronominal) and the position in which the complement appears (Vorfeld, Mittelfeld, in situ). Besides the linguistic analysis, methodological issues regarding the extraction of the relevant data from the treebanks will be addressed as well.

# An Exploration of Automatic Poetry Generation in Dutch

Tim van de Cruys
IRIT & CNRS
tim.vandecruys@irit.fr

The automatic generation of poems is a challenging task for a computational system. For a poem to be meaningful, both linguistic and literary aspects need to be taken into account. First of all, a poetry generation system needs to properly model language phenomena, such as syntactic well-formedness and topical coherence. Furthermore, the system needs to incorporate various constraints (such as form and rhyme) that are related to a particular poetic genre. And finally, the system needs to exhibit a certain amount of literary creativity, making the poem interesting and worthwhile to read. In recent years, a number of fruitful NLP approaches have emerged that are able to adequately model various aspects of natural language. In particular, neural network language models have improved the state of the art in language modeling, while topic models are able to capture a certain form of topical coherence. In this talk, we will explore how these approaches can be adapted and combined in order to model both the linguistic and literary aspects that are required for poetry generation. The presented framework will be applied to the generation of poems in Dutch.

# An agent-based model of Germanic verbal cluster word order change

Jelke Bloem
University of Amsterdam
j.bloem@uva.nl

Arjen P. Versloot
University of Amsterdam
a.p.versloot@uva.nl

Fred Weerman
University of Amsterdam
F.P.Weerman@uva.nl

Our study models the historical development of verbal cluster word order in Germanic languages using an agent-based model. We show that construction usage frequencies can help in explaining how current orders of English, Dutch and German verbal groups might have developed and diverged from the proto-Germanic cluster orders. Our basic model consists of surface order patterns in which we view both possible cluster orders ("that we have understood" vs "understood have") as separate outcomes, with a probability distribution over the outcomes. We initialize the model with (relative) frequency figures as reconstructed for 6th century Germanic, based on a comparison of Old English, Old High German and Old Frisian, and incorporated some known and relevant historical changes into the model. Language change is induced in the model by having agents learn from each other's output realizations. Our results show that the current order in German and Frisian verbal clusters may have developed partly due to the grammaticalization of embedding — an increased use of subordinate clauses over time. Conversely, the English order is explained by the model through faster grammaticalization of 'have'-clusters. The situation for Dutch is more complicated. Like many other phenomena, Dutch verbal cluster orders follow the Van Haeringen distribution, sharing features with both English and German. Historical data show that Dutch verbal clusters originally developed similar to German verbal clusters, but then diverged. Our model agrees with this, supporting the current state of Dutch only as an intermediate state in a process of language change.

# And what became of the (ISO)cat ?

Ineke Schuurman
K.U.leuven
ineke@ccl.kuleuven.be

Menzo Windhouwer
The Language Archive, DANS
Menzo.Windhouwer@dans.knaw.nl

Marc Kemps-Snijders
Meerstens Instituut
marc.kemps.snijders@meertens.knaw.nl

Daan Broeder
The Language Archive, MPI for Psycholinguistics
Daan.Broeder@mpi.nl

Since the end of last year, the CLARIN community no longer uses ISOcat as data category registry. The reason? ISOcaty was considered to be too complex, asking for more data than necessary for our purposes, and it was an open registry, one of the consequences being a proliferation of data. ISOcat does still exist, and ISO TC37 is planning the next DCR generation. But at the Meertens Institute (Amsterdam), the CLARIN Concept Registry (CCR) has been built, suited to our (CLARIN) needs. The ISOcat entries we need as the CLARIN community are exported into the CCR, where they show up in a modified, i.e., simplified, way. Exported are at least all DCs related to CMDI, plus those the national CLARIN groups wanted to be included as they are relevenant for their work. New entries can be added, but in a more controlled way: everybody has read-access, but only the national CCR coordinators can insert new entries, meaning that (proposals for) new entries should be passed on to them. This way we want CCR to become a registry with high-quality content. The cat's used up one of its lives, we're looking forward to its next CLARIN life.

# Applying terminology extraction to aspect-based sentiment analysis

Orphée de Clercq
T3, Language and Translation Technology Team, Ghent University
Orphee.DeClercq@UGent.be

Marjan van de Kauter
LT3, Language and Translation Technology Team, Ghent Universit
Marjan.VandeKauter@UGent.be

Els Lefever
T3, Language and Translation Technology Team, Ghent University
Els.Lefever@UGent.be

Véronique Hoste
LT3, Language and Translation Technology Team, Ghent Universit
veronique.hoste@ugent.be

There exists a large interest in sentiment analysis of user-generated content from both a societal and business perspective. Until recently, the main research focus was on discovering the overall polarity of a certain text or phrase. A noticeable shift has occurred to consider a more fine-grained approach, known as aspect-based sentiment analysis. For this task the goal is to automatically identify the aspects of given target entities and the sentiment expressed towards each of them. In this presentation, we will present the system that was developed to participate in the SemEval2015 shared task on aspect-based sentiment analysis. We focus on restaurant reviews and will reveal how we apply techniques originally developed for terminology extraction to the difficult task of aspect extraction. In a next step we will discuss our classification approach for finding both the categories to which certain aspects belong and how we discern between positive, negative and neutral sentiment for each of these individual aspects.

# Approach to non-standardised languages in Asian and African markets

Dorota Iskra
Appen
diskra@appen.com

As a global provider of language resources and linguistic services Appen often has to work with languages where no established standard is present for orthography, vocabulary or grammar. This problem, although it may concern some minorities' languages in Europe, becomes only apparent in its full dimensions when entering Asian and African markets. Examples are Arabic, Pashto, Urdu and various Indian languages. This presentation describes the problems we have encountered when moving to Asian and African languages and proposes a methodology for establishing internal standards for non-standardised languages. In the initial phase of the project when creating orthographic transcription from speech we end up with multiple spellings for the same word. These are all added to the dictionary since there is no spell checker. Our linguistic expert checks the dictionary and makes decisions about which words should actually have the same spelling. As a result a rough spell checker is created which can be used in the next phase of the project. But because it is far from comprehensive, an exception list is kept. In multiple iterations the linguistic expert goes through the list and adds words to the dictionary. Language technology expects consistency in the data which can only be achieved through a systematic approach in the face of lacking standards. The approach we propose has been developed and tested in a number of languages and with extensive volumes of data, not only by ourselves, but also by our clients who are developers and users of language technology.

## Architectures and representations for string transduction

Grzegorz Chrupała
Tilburg University
g.chrupala@uvt.nl

String transduction problems are ubiquitous in natural language processing: they include transliteration, grapheme-to-phoneme conversion, text normalization and translation. String transduction can be reduced to the simpler problems of sequence labeling by expressing the target string as a sequence of edit operations applied to the source string. Due to this reduction all sequence labeling models become applicable in typical transduction settings. Sequence models range from simple linear models such as sequence perceptron which require external feature extractors to recurrent neural networks with long short-term memory (LSTM) units which can do feature extraction internally. Versions of recurrent neural networks are also capable of solving string transduction natively, without reformulating it in terms of edit operations. In this talk I analyze the effect of these variations in model architecture and input representation on performance and engineering effort for string transduction, focusing especially on the text normalization task.

# Assessing the impact of frequency and local adaptation mechanisms on child-caregiver language: a recurrence-quantificational approach

Robert Grimm
Universiteit Antwerpen
robert.grimm@uantwerpen.be

The present work utilizes the relatively novel technique of recurrence-quantificational analysis (RQA) in order to investigate local adaptation processes in child-caregiver dialogue. RQA involves the construction of recurrence plots—structures which plot two data series against one another, and which allow for the extraction of further quantitative measures. Here, such plots are used to obtain information about: (1) the general use of a linguistic element in the other interlocutor's speech, and (2) the reuse of a linguistic element in temporally close child-adult turns Given this, it is possible to track the degree to which the prevalence of words and syntactic structures in child-caregiver language is influenced by (1) their frequency in the other interlocutor's speech and (2) their involvement in local adaptation processes (e.g. priming, repetition). Conducting one analysis each for child and caregiver speech, respectively, we concentrate on the usage of content words, function words, and part-of-speech tag bigrams (POS bigrams). That way, we aim to measure a more meaning-driven usage (content words) and a more syntactically oriented usage (function words) of lexical items. POS bigrams are assumed to correspond roughly to syntactic structures. Three corpora from the CHILDES database reveal that the frequency of most words and syntactic structures is generally more strongly determined by their frequency in the other interlocutor's speech. However, this changes when considering only high-frequency elements, where local adaptation mechanisms appear to exert a stronger influence on usage. Across corpora and interlocutors, there is both variance and remarkable homogeneity in how usage is affected.

# Automatic Conversion of an Episodes Script to a 3D Movie

Kim Hens
KU Leuven
kim.hens@student.kuleuven.be

Quynh Ngoc Thi Do
KU Leuven
quynhngocthi.do@cs.kuleuven.be

Marie-Francine Moens
KU Leuven
sien.moens@cs.kuleuven.be

In the project Machine Understanding for interactive StorytElling (EU ICT FP7 FET) at KU Leuven, we have introduced a new way of exploring and understanding textual information by "bringing text to life" through 3D interactive storytelling. As one step in the conversion of any text to a 3D movie, in this paper, we present an approach to convert a script of a television series written in natural language to a knowledge representation which can be used to generate a 3D movie automatically. We analyse all the challenges and difficulties of the automatic conversion from the text format/quality to the performance of state-of-the-art natural language processing softwares. The knowledge representation includes action templates, which are used to generate the 3D movie. An action is represented with an action name and a list of parameters. For example, an action MoveTo should contain a Character who is the mover, an Object to which the character moves, and a MovementType such as "run", or "walk". We detect entity mentions and semantic frames in the scripts by using coreference resolution and semantic role labeling. The entity mentions are referred to as the main characters, and the semantic frames are mapped to the action templates. We evaluated the technology with scripts of the television series of Grey's anatomy. One of our main results regard the adaptation of the existing Hobbs algorithm for coreference resolution to our problem domain while obtaining an F1 of 0.872. Another contribution is to extract and classify nouns as classes of characters, props and locations with an F1 of 0.750. Finally, we obtained an F1 of 0.737 for our conversion of textual scripts to action templates.

# Automatic Limerick Generator

Eric Sanders
CLS/CLST, Radboud University Nijmegen
e.sanders@let.ru.nl

In this presentation a demo is presented that automatically generates Dutch limericks. A limerick is a well known type of poem with a fairly fixed meter and a strictly fixed rhyme scheme. The generator works by filling in slots in a template. In the current version a job or animal and a city or country are introduced in the first line. Other words are randomly chosen from preprocessed parts of E-lex in such a way that meter and rhyme follow the rules for a limerick. This is done with the syntactic, syllable, stress and phonemic information in E-lex. The big challenge is to make meaningful poems (even within the broad poetic freedom boundaries). To achieve this Cornetto and word2vec will be used to find strongly related words and n-gram frequencies to find often co-occurring words. The demo is implemented as a web service. Users can input a job/animal or city/country with which they like to have a limerick generated. The user can score the generated limerick on a scale from 1 to 10. This allows us to optimise the different parameters in the system, like the strictness of the meter and rhyme, the allowed semantic distance between words and the used words themselves.

# Automatic extraction of disease features from Wikipedia

Eva D'Hondt
LIMSI
eva.dhondt@limsi.fr

Brigitte Grau
LIMSI
bg@limsi.fr

Pierre Zweigenbaum
LIMSI
pz@limsi.fr

Diagnostic clinical decision support systems (DCDSS) are interactive expert systems which are designed to assist physicians in determining the diagnosis of a patient's illness or condition. Such systems typically take a set of symptoms, signs and/or test results (a.k.a. disease features) as input and return a ranked list of hypothetical diagnoses. DCDSS have a long history and are used in multiple settings, such as general practitioner offices (Farmer et al. 2012), hospital wards (Elkin et al. 2011) as well as a self-diagnostic tool online (WebMD). Earlier systems only used hand-written rules but recently there has been a shift to systems that incorporate machine learning techniques, such as decision trees or neural networks. However, these systems need a database of signs/symptoms associated with disease to train on. Manually constructing and updating such a database is time-consuming work which requires medical specialists. Consequently, the existing DCDSS are almost all proprietary software with high subscription costs. Moreover the coverage of such databases is often fairly limited. In this paper we present an approach to mine Wikipedia for diseases and related features. Not only is Wikipedia a freely accessible resource, it has a wide coverage and contains enough high-quality medical information to be a common resource for doctors (Boyer et al., 2012). We employ MetaMap to annotate pertinent sections of the Wikipedia articles and train a Conditional Random Fields tagger to annotate the disease features. Afterwards we use the UMLS to normalize the extracted symptoms. This way we acquire a database that is freely distributable and has a much larger coverage than similar proprietary databases.

# Automatic word sense disambiguation for Dutch using dependency knowledge

Hessel Haagsma
Rijksuniversiteit Groningen
hesselhaagsma@gmail.com

An automatic word sense disambiguation system utilizing dependency knowledge is implemented using existing language resources for Dutch (Lassy, Alpino, Cornetto) and tested on a subset of DutchSemCor. The disambiguation method used largely follows the method first proposed by Lin (1997). It defines words by their local context, represented as dependency triples. The notion that words occurring in the same local contexts are semantically close to the ambiguous word is used to create a list of similar words. The correct sense is then found by selecting the sense that is semantically closest to the words in this list. Performance on a set of nouns, verbs and adjectives is tested, and overall performance is comparable or slightly higher than that reported by Lin: almost 9% over baseline for fine-grained sense distinctions and over 3% over baseline for coarse-grained sense distinctions. In absolute terms, disambiguation accuracy was highest for nouns, slightly lower for verbs and lowest for adjectives. The effect of using different local contexts and semantic databases was tested, which indicated that a reliably sense-annotated corpus is still required and that quality and types of dependency relations in the local context database matters more than quantity. Overall, performance is as expected, showing that dependency contexts are a useful feature for word sense disambiguation for Dutch.

# Between Hawks and Doves: Measuring Central Bank Communication

Ellen Tobback
University of Antwerp, Belgium
ellen_tobback@hotmail.com

David Martens
University of Antwerp, Belgium
david.martens@uantwerpen.be

Stefano Nardelli
ECB
stefano.nardelli@ecb.int

During the financial crisis, the implementation of monetary policy has changed significantly and communication has become a crucial instrument to conduct an effective monetary policy. As key interest rates approach the zero bound and the ability of steering interest rates at the short end of the money market curve has been partially hindered, the central banks introduced communication in the form of forward guidance as a powerful instrument to guide expectations and to provide additional monetary stimulus. While central bank communication has been a widely analyzed topic in the economic literature for several years, its perception by its main stakeholders (i.e. financial markets and public opinion) has hardly been addressed, in particular research using quantitative methods is lagging behind. In this research project, we present the development of an index that measures the interpretation of ECB official communication. The index represents the perceived degree of "hawkishness" or "dovishness" by classifying a set of news articles that report on ECB monetary policy decisions in these two categories, using both semantic orientation and a supervised SVM classification model. The resulting Hawkish-Dovish indicator is supported by a plot that shows the most popular topics around each press conference. These topics are extracted from the dataset of news articles using Latent Dirichlet Allocation. The index is currently used by the European Central Bank and will soon be reported in their Monthly Bulletin.

# Beyond training data: Sequence labeling using continuous vector representation of words

Chao Li
Textkernel BV
chaoli@textkernel.nl

Carsten Hansen
Textkernel BV
hansen@textkernel.nl

Gerard Goossen
Textkernel BV
goossen@textkernel.nl

Lena Bayeva
Textkernel BV
bayeva@textkernel.nl

Florence Berbain
Textkernel BV
berbain@textkernel.nl

Mihai Rotaru
Textkernel BV
rotaru@textkernel.nl

Information extraction from CVs (resumes) is one of the success stories of applying NLP in industry. Word type (i.e. word itself) based sequence labeling models (e.g. CRF) are typically used for this task. However, one disadvantage of this approach is its poor generalization to CVs from new sectors. For example, a typical training set contains about 3000 annotated CVs, which can struggle in terms of coverage for fields like job titles or company. When parsing CVs from other sectors (e.g. off-shore oil industry), many job titles or company words will be unknown. To solve this, one approach is to annotate more CVs from these sectors but this is an expensive solution. In this paper, we propose to use the recent developments in continuous vector representation of words to solve the unknown word problem. We use the word2vec package which was shown in previous work to produce representations that capture semantic similarities between words in an unsupervised fashion (e.g. "director" and "CEO" will be close in the vector space). When constructing the input features for the sequence labeling model, instead of using the word type as a sparse binary vector, we use directly the vector representation of the word. For example, assume 'director' is in our 3000 CVs training data but 'CEO' is not. At testing time, 'CEO' will have input features similar to 'director' for our model even if it is not present in the training data. We train the word representations on a large collection of CVs, including unlabeled CVs from the target sector. Next, we train our CRF model using these vectors. Our results show that compared to the word type based model, the new approach results in a significant improvement for CVs from new sectors without affecting performance on regular blind test data. Another practical advantage of this approach is that it allows us to tune the extraction model on the customer domain.

# Bilingual Markov Reordering Labels for Hierarchical SMT

Gideon Maillette de Buy Wenniger
Institute for Logic, Language and Computation (ILLC)
gemdbw@gmail.com

Khalil Sima'An
Institute for Logic, Language and Computation (ILLC)
k.simaan@uva.nl

Earlier work on labeling Hiero grammars with monolingual syntax reports improved performance, suggesting that such labeling may impact phrase reordering as well as lexical selection. In this paper we explore the idea of inducing bilingual labels for Hiero grammars without using any additional resources other than original Hiero itself does. Our bilingual labels aim at capturing salient patterns of phrase reordering in the training parallel corpus. These bilingual labels originate from hierarchical factorizations of the word alignments in Hiero's own training data. In this paper we take a Markovian view on synchronous top-down derivations over these factorizations which allows us to extract 0th- and 1st-order bilingual reordering labels. Using exactly the same training data as Hiero we show that the Markovian interpretation of word alignment factorization offers major benefits over the unlabeled version. We report extensive experiments with strict and soft bilingual labeled Hiero showing improved performance up to 1 BLEU points for Chinese-English and about 0.1 BLEU points for German-English.

# Combining rules with data: a synergistic approach to automatized transcription

Merijn Beeksma
Radboud University Nijmegen
merijnbeeksma@gmail.com

Johan Zuidema
Radboud University
joh.zuidema@online.nl

Anneke Neijt
Radboud University Nijmegen
a.neijt@let.ru.nl

Transcription between the orthographic spelling of words and their phonetic representations is an exercise in pattern recognition. Some systems convert between graphemes and phonemes based on linguistic rules. Though theoretically well-founded, exceptions to the rules make such systems difficult to apply in practice. Other systems are based on extensive databases, in which linguistic rules do not figure prominently. Zuidema developed the BasisSpellingBank (Zuidema & Neijt 2012), an extensive database covering orthography and phonology, and all the relations linking them together. Encoding linguistic knowledge not in algorithms but as data reduces the task of transcription to an exercise in data retrieval, allowing for simpler, more flexible systems. Our main transcription algorithm converts graphemes into phonemes and vice versa by combining fragments of words from the database optimally. The program not only provides solutions, but also explains how it arrives at them. The direct mapping between letters and sounds in the BasisSpellingBank allows the same software to transcribe bidirectionally. This presentation will show how the algorithms team up with the database. The effectiveness of the algorithm is demonstrated against existing corpora such as Celex.

# Comparing NLP techniques to represent unstructured data sources for the prediction of clinical codes

Elyne Scheurwegs
UAntwerpen
elyne.scheurwegs@uantwerpen.be

Tim van den Bulcke
UZA
tim.vandenbulcke@uza.be

Walter Daelemans
UAntwerpen
walter.daelemans@uantwerpen.be

Clinical coding is the process of assigning diagnostic and procedural codes to a patient stay in a hospital. In this project, we automatically predict a prevalent type of clinical codes (i.e. ICD-9-CM) using a range of unstructured data sources (e.g. medical test reports, discharge letters). We represent these data sources using a bag of words (BoW) of the word form, lemmas, nouns and verbs, combinations of these representations and derivative meta-features. These representations will be converted to features to predict clinical codes. This is evaluated on a large set of anonimised patient files.

# Computational Construction Grammar: A Survey of the State-of-the-Art

Remi van Trijp
Sony Computer Science Laboratory Paris
remi@csl.sony.fr

Construction Grammar (Goldberg, 2006, henceforth CxG) is a family of linguistic theories that propose a wholly different architecture for language processing than found in most NLP systems. Rather than slicing language into horizontal, largely independent modules, CxG hypothesizes that the central units of linguistic knowledge are constructions that vertically cut across different sources of information. Schneider and Tsarfaty (2013) have therefore argued that a scalable implementation of "the CxG perspective presents a formidable challenge to the computational linguistics/natural language processing community" that "would [...] make for a more comprehensive account of language processing than our field is able to offer today" (p. 451). This presentation will present the state-of-the-art in the young but booming discipline of computational construction grammar. I will first explain the fundamental assumptions of CxG and show how a "construction" can become a new and powerful data structure for language processing. I will then survey the most important developments (here in alphabetical order): Embodied Construction Grammar (Bergen & Chang 2005), Fluid Construction Grammar (Steels 2011), Sign-Based Construction Grammar (Boas and Sag 2012) and Template Construction Grammar (Arbib 2012).

# Coordinating on the semantics of referring expressions: miscommunication drives abstraction

Gregory Mills
University of Groningen, Netherlands
g.j.mills@rug.nl

One of the central findings in dialogue research is that interlocutors rapidly converge on a shared set of contracted referring expressions (Krauss and Weinheimer, 1966; Clark, 1996) which become progressively systematized and abstract. This occurs for a wide range of referents, e.g. when referring to spatial locations (Garrod and Doherty, 1994; Galantucci, 2005), music (Healey, 2002), concepts (Schwartz, 1995), confidence (Fusaroli et al., 2012), and temporal sequences (Mills, 2011). Cumulatively, these findings suggest that interaction in dialogue places important constraints on the semantics of referring expressions. However, there is currently no consensus about how best to account for how coordination develops. To investigate in closer detail the development of referential coordination, we report a variant of the "maze task" (Pickering and Garrod, 2004). Participants communicate with each other via an experimental chat tool (Mills and Healey, 2006), which interferes with the unfolding dialogue by inserting artificial probe clarification requests that appear, to participants as if they originate from each other. The clarification requests signal apparent miscommunication of participants' referring expressions. Participants who received clarification requests converged more rapidly on more abstract and more systematized referring expressions. We demonstrate how this beneficial effect is due to the artificial clarification requests amplifying naturally occurring miscommunication (Healey et al, 2013), yielding enhanced problem detection and recovery from error.

# Crowdsourcing Temporal Relations in Italian and English

Tommaso Caselli
VU Amsterdam
t.caselli@gmail.com

Rachele Sprugnoli
Fondazione Bruno Kessler
sprugnoli@fbk.eu

This paper reports on two crowdsourcing experiments on Temporal Relation Annotation in Italian and English. The aim of these experiments is three-fold: first, to evaluate average Italian and English native speakers on their ability to identify and classify a temporal relation between two verbal events; second, to assess the feasibility of crowdsourcing for this kind of complex semantic task; third to perform a preliminary analysis of the role of syntax within such task. Two categories of temporal relations were investigated: relations between the main event and its subordinated event (e.g. So che hai visto Giovanni / I know you've seen John) and relations between two main events (e.g. Giovanni bussò ed entrò / John knocked and got in). Fifty aligned parallel sentences in the two languages from the MultiSemCor corpus were extracted. In each sentence, the source and the target verbs of the relations were highlighted and contributors were asked to select the temporal relation from 7 values (AFTER, BEFORE, INCLUDES, IS INCLUDED, SIMULTANEOUS, NO RELATION, and DON'T KNOW) inspired by the TimeML Annotation Guidelines. For each sentence, 5 judgments were collected. The results of the annotator agreement is 0.41 for Italian and 0.32 for English. Analysis of the data has shown that annotating temporal relations is not a trivial task and that dependency relations between events have a major role in facilitating the annotation. Future work aims at conducting new experiments with an additional parameter, namely factivity, and with texts in a different domain, i.e. History.

# Detecting Implicit Opinions with a Target-specific Opinion Thesaurus

Sergei Kulikov
I-Teco and Institute of Linguistics of Russian Academy of Sciences
sukulikov@gmail.com

Implicit opinions are usually regarded as comparative or user-oriented opinions [Liu, 2012]. These approaches rely heavily on extra-linguistic resources, such as a user's likes or dislikes. Furthermore, focusing on syntactic relations they overlook a certain lexical class which can be named target-specific opinion lexica. in this paper we present an approach of automatic identification of target-specific opinion words. The said class usually denotes negative attitude towards an entity the meaning of which is incorporated into the word itself. The most frequently used type are words denoting xenophobia, e.g. islamophobia, latinophobia. These words are often countered by 'philia-words', e.g. obamamania, judeophilia. The structure of these words includes their referent, thus making it possible to link them semi-automatically via a thesaurus. The structure of a target-specific opinion thesaurus makes it possible to improve the coverage by automatically generating POS-derivatives (e.g. adjectives and verbs). Apart from 'phobia' and 'philia' words some other types of words can be included into thus a thesaurus. In our opinion among these types should be language-specific opinion words, like Russain 'nedobank' (meaning 'a bad bank') which is neutral in all of the contexts unrelated to banking. The main use of a target-specific opinion thesaurus is automatic detection of hatred-related comments alongside with an increase in coverage of implicit opinions for political entities.

# Distributional semantics for child-directed speech: a multimodal approach

Giovanni Cassani
University of Antwerp
cassani.giovanni@gmail.com

Marco Baroni
University of Trento
marco.baroni@unitn.it

In this work, linguistic and multimodal Distributional Semantic Models (DSMs) are applied to the study of child-directed speech gathered from the CHILDES dataset with the aim of starting to explore their potential utility and effectiveness in approximating patterns and phenomena that takes place during language acquisition. Most of the ideas developed are based onto a recent line of research that encodes features extracted from images into distributional vectors that are then combined with distributional vectors encoding linguistic information, with the scope of overcoming one major flaw of traditional DSMs, i.e. the absence of whatever link to the external world and the objects to which words refer. We discuss three different experiments: the first two aims to test to which extent semantic information contained in child-directed speech is informative and meaningful. We first address this issue by comparing semantic relatedness scores for several words, computed using distributional vectors derived from child-directed utterances, to human-generated gold standard scores. Secondly, we investigate how well semantic information extracted from the same data can reproduce human-generated features for different concepts in a feature listing task. Results are encouraging, showing that information extracted from child-directed speech is semantically and perceptually richer than information contained in one of the most commonly used corpus for standard English, viz. UkWac. Finally, we exploit the multimodal (text- and image-based) vectorial representations to perform zero-shot learning, a task in which a mapping function from visual to linguistic vectors is learned for a set of known concepts and then applied to a set of unseen concepts. The learned function then takes the visual vectors for a set of unseen images and project them onto the linguistic space, mapping new images to the right word labels. This third experiment aims at testing whether such mapping can account for, or at least approximate, the way children acquire certain aspects of language. The outcome of this last experiment is far from being satisfactory, but possible ways of improving the design and hopefully the outcome are discussed.

# Dutch Terminology Service Centre (Steunpunt Nederlandstalige Terminologie)

Anneleen Schoen
VU Amsterdam
a.m.schoen@vu.nl

Hennie van der Vliet
VU Amsterdam
h.d.vander.vliet@vu.nl

The Nederlandse Taalunie (Dutch Language Union, NTU), a Dutch-Flemish governmental organization for the interests of the Dutch language, founded the Dutch Terminology Service Centre (TSC) in 2007. The aim of the Dutch TSC is to inform and provide support to end-users of Dutch terminology. There is a need for this support. People involved in practical terminology work are often unaware of the existence of theories and tools, let alone they know how to make the best use of them. Furthermore, terminological theories and tools are often much too complicated and time-consuming to be of any practical use. The NTU and the Dutch TSC want to bridge the gap between theory and practice in Dutch Terminology. That's why the NTU is developing simple, yet powerful tools for terminologists. One of our goals for next year is to introduce a term extraction tool for Dutch. This tool, called 'TermTreffer', will be the first language specific term extractor for Dutch. With TermTreffer people can extract domain specific words from a given corpus. In its default modus the tool offers term extraction in a straightforward way. However, TermTreffer also offers the possibility of tuning the extraction pipelines. As a result, the extractor will be a useful tool in various research scenarios. We would like to give a demonstration of the tool and/or present the tool in a poster.

# ESO: A Frame-based Ontology for Events and Implied Situations

Roxane Segers
Vrije Universiteit Amsterdam
roxane.segers@gmail.com

Piek Vossen
VU University Amsterdam
piek.vossen@vu.nl

We present the Event and Situation Ontology (ESO), a resource which formalizes the pre and post conditions of events and the roles of the entities affected by an event. As such, the ontology allows for extracting information from text that otherwise would have been implicit. For example, from the expression 'Apple fired Steve Jobs', ESO can infer that an Employee (Steve Jobs) worked for an Employer (Apple) before he was fired and that he isn't working for Apple after he was fired. The ontology reuses and maps across existing resources such as FrameNet, Wordnet and SUMO. Through these mappings, ESO serves as a hub to other vocabularies as well, such as Princeton Wordnet (PWN) and the Wordnets in the Global Wordnet grid. As such, ESO provides a hierarchy of events and their implications across languages. Currently, the ontology consists of 59 event classes with 94 mappings to FrameNet and 49 mappings to SUMO on class level. In total, 24 properties were defined to model the pre and post conditions of events. For the roles of the entities affected by an event, 60 mappings to FrameNet Frame Entities were created. We show how ESO is designed and employed to assign additional annotations on millions of processed articles on both predicate and role level thus allowing for inferencing over various events and implications. The ontology itself is available in OWL at: https://github.com/newsreader/eso

# Error analysis of Word Sense Disambiguation results

Rubén Izquierdo
VU University Amsterdam
ruben.izquierdobevia@vu.nl

Marten Postma
VU University Amsterdam
martenp@gmail.com

Word Sense Disambiguation is still an unsolved problem in Natural Language Processing. We claim that most approaches do not model the context correctly, by relying too much on the local context (the words surrounding the word in question), or on the most frequent sense of a word. In order to provide evidence for this claim, we conducted an in-depth analysis of all-words tasks of the competitions that have been organized (Senseval 2&3, Semeval-2007, Semeval-2010, Semeval 2013). We focused on the average error rate per competition and across competitions per part of speech, lemma, relative frequency class, and polysemy class. In addition, we inspected the "difficulty" of a token(word) by calculating the average polysemy of the words in the sentence of a token. Finally, we inspected to what extent systems always chose the most frequent sense. The results from Senseval 2, which are representative of other competitions, showed that the average error rate for monosemous words was 33.3% due to part of speech errors. This number was 71% for multiword and phrasal verbs. In addition, we observe that higher polysemy yields a higher error rate. Moreover, we do not observe a drop in the error rate if there are multiple occurrences of the same lemma, which might indicate that systems rely mostly on the sentence itself. Finally, out of the 799 tokens for which the correct sense was not the most frequent sense, system still assigned the most frequent sense in 84% of the cases. For future work, we plan to develop a strategy in order to determine in which context the predominant sense should be assigned, and more importantly when it should not be assigned. One of the most important parts of this strategy would be to not only determine the meaning of a specific word, but to also know it's referential meaning. For example, in the case of the lemma 'winner', we do not only want to know what 'winner' means, but we also want to know what this 'winner' won and who this 'winner' was.

# Evaluation of context-free language learning systems across languages

Menno van Zaanen
Tilburg University
mvzaanen@uvt.nl

Nanne van Noord
Tilburg University
N.J.E.vanNoord@uvt.nl

Grammatical Inference (GI) deals with finding a compact, finite representation, i.e. a grammar, given information from a possibly infinite language. Research in formal GI leads to mathematical proofs of efficient learnability of families of languages. A language family is a set of languages sharing a common property, such as context-freeness. Such properties are used in mathematical proofs to show that in specific learning environments, all languages in the family can be learned within polynomial time. In contrast, empirical GI is based on mathematical principles and is about imitating what nature (or humans) do. The aim is to try to learn natural languages, which are known to be efficiently learnable (as we humans do this). In this setting, however, a formal description of the underlying language family is currently unknown. Because of this, exact language identification in empirical GI is often impossible and we allow for approximate identification. Several empirical GI systems have been developed. In particular, context-free grammar learning systems have been applied to natural language sentences to try to learn syntax. Here, we will compare and evaluate several systems. Firstly, we will describe a general framework, which consists of two distinct phases. Instantiations of these two phases can be identified in the existing systems and alternative systems based on new combinations of the instantiations are evaluated as well. Secondly, we will provide results of the systems on several languages, which shows that these systems can indeed identify and generalize over syntactically different languages.

# Extending n-gram language models based on equivalent syntactic patterns

Lyan Verwimp
KU Leuven
lyan.verwimp@esat.kuleuven.be

Joris Pelemans
KU Leuven
joris.pelemans@esat.kuleuven.be

Hugo van Hamme
KU Leuven
hugo.vanhamme@esat.kuleuven.be

Patrick Wambacq
KU Leuven
patrick.wambacq@esat.kuleuven.be

In automatic speech recognition, the language model helps to disambiguate between words with a similar pronunciation. A standard language model is typically based on n-grams (sequences of n consecutive words) and their probabilities of occurrence. These n-gram models however suffer from data sparsity and cannot model long-span dependencies. The purpose of this research is to alleviate the former problem by automatically generating more n-grams than the ones based on surface structure. Like many other languages, Dutch often can have the same syntactic pattern in different word orders (e.g. subject – verb inversion, switching between verb – object in head clause and object – verb in subordinate clause, conjunctions). In this work, we investigate whether we can generate new, meaningful n-grams based on these word order switches in an attempt to increase n-gram coverage. We do this in the following way: first our training data is parsed by Alpino, a dependency parser for Dutch. Based on these parses, we then extract those patterns for which the word order can be reversed and add the corresponding reversed n-grams to the language model. Some probability needs to be assigned to these extra n-grams and to that end we investigate several existing ways of redistributing probability mass. Finally, we compare the performance of the extended language model with the original n-gram model by evaluating their predictive power on a test set of Southern Dutch newspaper material.

# Factored and hierarchical models for Dutch SMT

Joachim van den Bogaert
CCL Leuven
joachim@crosslang.com

Factored and hierarchical models for Statistical Machine Translation have been around for some time, but not many results with respect to Dutch have been reported yet. An important cause is probably the fact that simple phrase-based models require less effort to train while performing on par with more involved approaches. In this presentation, we discuss a couple of experiments we conducted using some of Moses SMT's more advanced features (multiple translation and generation steps, and multiple decoding paths), while leveraging monolingual data for inflexion prediction using lemmata and part-of-speech information. In a follow-up discussion, we focus on the asymmetric nature of Machine Translation from English into Dutch and Dutch into English, and present some strategies we discovered for data sparsity reduction using the factored model framework. Reordering will be discussed by comparing hierarchical models and pre-ordering strategies. We conclude the presentation with an overview of how different strategies can be combined and to which extent such strategies remain effective when larger data sets are used.

# Finding and Analyzing tweets from Limburg and Friesland

Dolf Trieschnigg
University of Twente
d.trieschnigg@utwente.nl

Dong Nguyen
University of Twente
d.nguyen@utwente.nl

Lysbeth Jongbloed
Fryske Akademy
ljongbloed@fryske-akademy.nl

Jolie van Loo
Meertens Instituut
j.j.e.vanloo@gmail.com

Leonie Cornips
Meertens Instituut / Maastricht University
leonie.cornips@meertens.knaw.nl

Theo Meder
Meertens Instituut
theo.meder@meertens.knaw.nl

Many Dutch people are multilingual and also their language use on Twitter is not restricted to Dutch exclusively. In the provinces Friesland and Limburg Twitter people may also tweet in their minority, regional languages. How frequent do these users switch to a different language or use linguistic forms associated with different languages and why? In this paper we will address the first question and present how we collected Twitter users from these provinces by georeferencing the location provided in the user's profile. We analyzed a sample of over 4000 tweets from both provinces. 5% to 8% of these tweets display the use of (elements of) regional languages. Less than 2% of the analyzed tweets are expressed in a language other than Dutch, English, or a regional language. When using automatic dialect identification we observe similar percentages of regional language use in a larger collection of tweets. An 'error' analysis of the automatic language identification shows that Limburgish and Dutch are frequently mixed,; Frisian seems less frequently mixed. Of course, languages are no discrete objects and, hence, the mixing behavior of Twitter users does not constitute an error. Therefore, we will problematize the automatic language identification tool: is it suitable to analyze and understand recent daily language practices on Twitter?

# From Certainty to Doubt: A Corpus-Based Analysis of Epistemic Expressions in Pre-Lockean and Contemporary Scientific Discourse

Marcelina Florczak
University of Warsaw
florczak.marcela@gmail.com

The presentation is going to depict a semantic change 'from certainty to doubt' in English scientific discourse. Epistemic expressions, such as 'verily', 'probably', 'evidently', 'I doubt', 'I think/ methinks', 'I know/ I wot', will be analyzed in two corpora compiled for this study. The underlying assumption is that the publication of 'An Essay Concerning Human Understanding' by John Locke gave a significant rise in the use of epistemic expressions in the language of the academia, reflecting the acknowledgement of the limitations of one's knowledge. The linguistic analysis suggests scientific discourse of the 16th and 17th centuries was strongly influenced by the ethos of truth, faith and certainty. It is claimed that contemporary discourse, however, mirrors suspicion, fragmentation, as well as scepticism towards scientific progress.

## From paper dictionary to electronic knowledge base

Hans Paulussen
K.U.Leuven
hans.paulussen@kuleuven-kulak.be

Martin Vanbrabant
KU Leuven kulak
Martin.Vanbrabant@kuleuven-kulak.be

Gerald Haesendonck
UGent
Gerald.Haesendonck@UGent.be

This talk describes a work in progress on the conversion of a paper dictionary into an electronic knowledge base. The resulting dictionary will be one of the resources of an enrichment engine developed to enrich electronic texts published on a language learners' platform. Thanks to the converted dictionary, we intend to automatically provide short descriptions, sample sentences and illustrations for words selected in reading texts for language learners of NT2 (Dutch as a foreign language). Although quite a number of dictionaries are available on the internet, they are not always adapted to the needs of language learners. Wiktionary, for example, gives ample information on interesting topics, but the dictionary articles can be overwhelmingly difficult for beginners of Dutch, especially at A1 and A2 CEFR level. Moreover, dictionaries are still often produced for paper publication, so that electronic formating is mainly focused on paper layout. In this project, we started from a learners' dictionary stored in Indesign INDD format. The rich INDD layout format can easily be exported to XML, which could be considered the basic format for further processing. However, IDML (the XML format of Indesign) is mainly used for layout purposes, so that the logical structure of the content is not necessarily maintained. This talk will describe the approach followed to semi-automatically convert the original XML file into an RDF stored knowledge base. Special attention will be given to the main challenge of the automatic conversion task consisting in rebuilding the logically sequences of texts that lay scattered over layout oriented lines and columns.

# Generating Genitive Alternation using Projective Discourse Representation Theory

Noortje Venhuizen
Rijksuniversiteit Groningen
n.j.venhuizen@rug.nl

Valerio Basile
Rijksuniversiteit Groningen
v.basile@rug.nl

For the purpose of generating natural language from logical structures, it is important to have a semantic formalism capable of expressing the fine-grained distinctions that are reflected in the syntax of the natural language. Such a formalism should be able to represent differences in the linguistic surface form while remaining faithful to the model-theoretic properties of a logical formalism (Shieber, 1993). Projective Discourse Representation Theory (PDRT; Venhuizen et al., 2013) presents an attractive framework for this purpose, because the representations of PDRT directly correspond to the surface structure of an utterance. The interpretation of linguistic content is indicated using projection variables, which explicitly reflect interactions between the surface structure and the logical interpretation, and can therefore be used to capture the effects of different (syntactic) constructions. As a case study, we investigate the English genitive alternation, i.e., "John's dog" versus "The dog of John". Various studies have shown that the use of these constructions varies according to the context in which the possessive is used (see Rosenbach, 2014). This suggests that the constructions should also be distinguishable within the semantic formalism applied for language generation. We show how the different genitive constructions can be represented in the PDRT framework, and describe some predictions about the felicity constraints of the different constructions. In particular, we show how PDRT can be employed to improve the quality of generation in the context of statistical NLG that exploits the alignment between logical formulas and surface forms (see Basile and Bos, 2013).

# HLT Agency – no service like more service

Remco van Veenendaal
Taalunie/TST-Centrale
rvanveenendaal@taalunie.org

Ten years ago, the Dutch Language Union took the initiative to set up the HLT Agency as a central repository for Dutch digital language resources (LRs). The rationale behind the HLT Agency was to prevent LRs developed with public money from becoming obsolete and therefore useless and to enhance visibility, accessibility and re-usability by having one organisation managing the lifecycle of Dutch LRs. Today's landscape is different, but the maintenance and reusability of LRs are still topical issues. The HLT Agency is adapting to this changing landscape. Under the aegis of the Dutch Language Union the HLT Agency will continue its established acquisition, management, maintenance, distribution and support services, but it will widen its scope from HLT-specific resources to more general Dutch digital language resources. The current emphasis on data management will certainly contribute to saving many language data, but to make sure that such data also enjoy a second life will require some extra efforts. To this end, the HLT Agency intends to focus on networking activities aimed at bringing together resources, knowledge and potential users in order to stimulate reuse, cooperation and innovation. Tomorrow's landscape will thus include an HLT Agency that is still in charge of LRs, but which has a wider view and offers more services so that the prerequisites for reuse, cooperation and innovation can be met. In turn this will contribute to strengthening the position of the Dutch language in the digital age.

# High-quality Flemish Text-to-Speech Synthesis

Lukas Latacz
Vrije Universiteit Brussel / iMinds
llatacz@etro.vub.ac.be

Wesley Mattheyses
Vrije Universiteit Brussel / iMinds
wmatthey@etro.vub.ac.be

Werner Verhelst
Vrije Universiteit Brussel / iMinds
wverhels@etro.vub.ac.be

Even though speech synthesis is the most frequently needed language technology for people with communicative disabilities (e.g. see [1]), the number of commercially-available synthetic voices is still rather small, especially for medium-sized languages such as Dutch and for specific language-variants such as Flemish (i.e. Southern-Dutch). The lack of high-quality Flemish synthetic voices available for research purposes inspired us to construct a new high-quality speech synthesizer, the DSSP synthesizer, able to synthesize Flemish speech. New voices for the synthesizer are constructed using a speech database containing recordings of a single speaker and the corresponding orthographic transcription. Our aim has been to automate the voice-building process as much as possible by 1) an automatic annotation of these recordings, 2) automatically spotting potential errors, and 3) deriving optimal settings without human intervention. The annotated speech recordings form the basis for constructing a speaker-dependent front-end that captures the speaking style of the original speaker by modelling speaker-specific pronunciations, prosodic phrase breaks, silences, accented words and prominent syllables. The front-end performs the language-dependent processing part of the synthesis, for which it uses multiple linguistic levels (segment, syllable, word and phrase) and shallow linguistic features. The DSSP synthesizer uses the two dominant speech synthesis techniques, unit selection synthesis [2] and hidden Markov model-based (HMM) synthesis [3]. HMM-based synthesis uses a flexible statistical parametric model of speech, but sounds less natural than unit selection synthesis, which selects small units from the speech database and concatenates the corresponding waveforms. Various demonstrations of our Flemish voices will be given during the talk.

# How Synchronous are Adjuncts in Translation Data?

Sophie Arnoult
ILLC
s.arnoult@gmail.com

Khalil Sima'An
University of Amsterdam
k.simaan@uva.nl

The argument-adjunct distinction is central to most syntactic and semantic theories. As optional elements that refine (the meaning of) a phrase, adjuncts are important for recursive, compositional accounts of syntax, semantics and translation. In formal accounts of machine translation, adjuncts are often treated as modifiers applying synchronously in source and target derivations. But how well can the assumption of synchronous adjunction explain translation equivalence in actual parallel data? In this paper we present the first empirical study of translation equivalence of adjuncts on a variety of French-English parallel corpora, while varying word alignments so we can gauge the effect of errors in them. We show that for proper measurement of the types of translation equivalence of adjuncts, we must work with non-contiguous, many-to-many relations, thereby amending the traditional Direct Correspondence Assumption. Our empirical results show that 70% of manually identified adjuncts have adjunct translation equivalents in training data, against roughly 50% for automatically identified adjuncts.

# How does In-domain Terminology Improve Statistical Machine Translation?

Liling Tan
Saarland University
alvations@gmail.com

Francis Bond
Nanyang Technological University
bond@ieee.org

Josef van Genabith
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
josef.van_genabith@dfki.de

Extending parallel data with lexicon/terminology/dictionary prior to training a Statistical Machine Translation model has shown to improve overall phrase-based MT performance. The primary motivation is to use the additional lexical information for domain adaptation (Koehn and Schroeder, 2007; Meng et al. 2014). Theoretically, adding out-of-vocabulary words into the parallel data will always improve SMT performance because the language model do not need to backoff to unknown word/ngram probability when optimizing the log-linear likelihood. Alternatively, adding in-domain lexicon to the parallel data is another popular approach to improve phrase-base MT. The intuition is that adding extra counts of isolated lexical entries overweighs the alignments between dictionary entries and their translations and minimizes bad word/phrasal alignment with the other context words (Tsetkov and Wintner, 2012; Skadins et al. 2013). Experimentally, the second motivation checks out in the BLEU improvements as shown in previous studies. Yet empirically, we are unsure of how and why adding terminological or lexical dictionaries improves SMT performance. Our pilot experiment on Jap-Eng MT using WAT shared task dataset (Nakazawa et al. 2014) has shown that appending an in-domain dictionary has minimal effect on the system (BLEU: 18.57 -> 18.87) and adding the lexicon more than once further improves the overall BLEU scores (18.87 -> 18.91). However, upon appending the dictionary beyond five times, the performance degrades (<18.57). Our current assumption is that the overweighed lexical items cause alignment model and language model to shift their probability mass too drastically. For the rest of this paper, we will (i) investigate this overweighing effect by tracking the alignment and language model probabilities of the entries in the lexicon and (ii) examine the breaking point where too much lexicon is harming the system.

# I had the most wonderful dream: A text analytics exploration of reported dreams

Antal van den Bosch
Radboud University Nijmegen
Antal.vdnBosch@uvt.nl

Maarten van Gompel
Radboud University Nijmegen
proycon@anaproy.com

Iris Hendrickx
Radboud University Nijmegen
iris@i-hx.nl

Ali Hürriyetoğlu
Radboud University Nijmegen
ali.hurriyetoglu@gmail.com

Folgert Karsdorp
Meertens Instituut
folgert.karsdorp@meertens.knaw.nl

Florian Kunneman
Radboud University Nijmegen
f.kunneman@let.ru.nl

Louis Onrust
Radboud University Nijmegen
l.onrust@let.ru.nl

Martin Reynaert
Tilburg University, Radboud University Nijmegen
reynaert@uvt.nl

Wessel Stoop
Radboud University Nijmegen
w.stoop@let.ru.nl

Dreams, the involuntary perceptions that occur in our minds during sleep, have been the topic of studies in many fields of research, including psychiatry, psychology, ethnography, and neurobiology. Dreams have a narrative content and can be considered as a form of storytelling with a bias towards one's personal life and a typical personal perspective. We present a study on dreams aimed at the large-scale analysis of dreams using text analytics. Our first case studies zoom in (1) on automatically tracing real-world entities (such as famous people and locations) in dreams descriptions, (2) on learning to reproduce standardized dream annotations from labeled examples, and (3) on discovering latent topics in dreams descriptions. We apply these supervised and unsupervised methods to a collection of about 22K English dream reports from the benchmark dataset Dreambank (www.dreambank.net) and a sample of 80K Dutch tweets containing dream descriptions.

# Improving Dutch coreference resolution by using noun-clusters

Rik van Noord
University of Groningen
rikvannoord@gmail.com

Coreference resolution is the task of determining whether two noun phrases refer to the same entity in the real world. In this study automatically generated noun clusters are used as a semantic information source in a memory-based learning approach to solve the task of Dutch coreference resolution, as was done in Hendrickx et al. (2008), using the clustering method of van der Cruys (2005). The effect of cluster size (average amount of words per cluster) is investigated and the effect of the clusters is tested on a dataset of only common nouns. We try to find an optimal cluster size and hypothesize that the optimal average size of a cluster will be higher than the average size (10) of the clusters used in Hendrickx et al. (2008). Adding cluster features yielded a 27,9% higher F-score than our baseline, and by testing different cluster sizes we found an optimal cluster size of 10 words per cluster, with slightly lower F-scores the more we move away from 10 words per cluster. In a more detailed analysis the specific effects of the different cluster sizes are shown. The results suggest that an average size of 10 words per cluster is optimal for the task of coreference resolution.

# Inducing Semantic Roles within a Reconstruction-Error Minimization Framework

Ivan Titov
University of Amsterdam
titov@uva.nl

Ehsan Khoddam Mohammadi
University of Amsterdam
ehsan.khoddam@gmail.com

In recent years there has been some interest in inducing semantic role representations from unannotated data. The existing methods have a number of serious shortcomings. First, they make very strong assumptions, for example, assuming that arguments are conditionally independent of each other given the predicate. Second, unlike state-of-the-art supervised semantic role labelers, they rely on a very simplistic set of features of a sentence. These assumptions are especially problematic for languages with freer word order than English, where richer features are necessary to account for interactions between surface realizations, syntax and semantics. In order to address these challenges, we introduce an expressive feature-rich model of semantics and also propose a reconstruction-error minimization method for its effective unsupervised estimation. Unlike much of the previous work, in our model we do not rely on any prior linguistic knowledge about the language. Our preliminary experiments show that, even without this prior knowledge, out method achieves state-of-the-art results on English and significantly outperforms baselines on German.

# Inferring Hypernym/Hyponym Relations in Dutch/English Parallel Texts

Johannes Bjerva
University of Groningen
j.bjerva@rug.nl

Johan Bos
University of Groningen
johan.bos@rug.nl

Natural language inference is the problem of drawing conclusions from language which are not explicitly stated, e.g., by determining whether a natural language hypothesis 'h' can be inferred from a natural language premise 'p'. The problem can be considered in a monolingual or cross-lingual setting, and in various contexts, such as sentences, phrases or semantic concepts. This is an important problem to consider for various applications, such as machine translation and question answering. The task of inferring this relation has been the focus of much research, although previous work has focussed on monolingual approaches. This work focusses on a cross-lingual approach. A first step towards cross-language inference is to infer hypernymy/hyponymy cross-lingually. In this work, we present an approach to make such inference in Dutch-English parallel texts. We solve this problem by creating a semantic hierarchy for English using word embeddings, which is is projected onto a Dutch word space. We show that the approach can infer cross-lingual hypernyms/hyponyms between Dutch and English within a sub-domain of WordNet. In future research, we will bypass the step of mapping between word spaces in different languages, by using bilingual word embeddings.

# LECSIE - Linked Events Collection for Semantic Information Extraction

Juliette Conrath
IRIT
juliette.conrath@irit.fr

Stergos Afantenos
IRIT, CNRS/Université Paul Sabatier
Stergos.Afantenos@irit.fr

Nicholas Asher
CNRS Laboratoire IRIT, Université Paul Sabatier
asher@irit.fr

Philippe Muller
IRIT, Toulouse University
muller@irit.fr

Our research work focuses on the extraction of semantic relations essential to the analysis of discourse, in which links are made between units of text in virtue of semantic information. These relations are frequently implicit and require for their identification inference using lexical and compositional semantic information. Our approach addresses this challenge, using non annotated data with automatically detected features to find adjacent clauses in which various discourse relations occur, to allow the extraction of typical lexical features. Specifically, when these clauses contain certain discourse connectives, we recover automatically a discourse relation that we associate with the main verbs of those clauses. We extract triples consisting of the two verbs and a semantic relation from a large corpus with the aim of inferring that such a pair of verbs can suggest the semantic relation even in the absence of an explicit discourse marker. Altough this analysis allows for noisy results, its application on a large corpus yields significant triples. Thus we compute significance measures to build a resource composed of ranked triples. This resource is evaluated first intrinsically by computing the correlation between our significance measures and a human association value obtained by manual annotation for selected triples, then extrinsically by computing coverage on discourse annotated corpora, and finally by including our triple significance scores as additional features in an automatic discourse relation prediction model. Our positive results show the potential impact of our resource for discourse analysis tasks as well as other semantically oriented tasks.

# Lexical choice in generation from Abstract Dependency Trees

Dieke Oele
Rijksuniversiteit Groningen
d.oele@rug.nl

Gertjan van Noord
Rijksuniversiteit Groningen
g.j.m.van.noord@rug.nl

Lexical choice in Natural Language Generation is the problem of deciding what words to use to adequately express the intended meaning. We address this problem by selecting words for a given synset in abstract dependency representations for generation. We experimented with a number of techniques to select the appropriate lemma for a given synset, in order to see the effect on the generator. In a first experiment, we select the first lemma in a given synset. This did not amount to satisfactory output for two reasons. In many cases, the chosen lemma was inappropriate in the given context. In almost of a third of the cases, the generator had difficulty generating a sentence. The results for both the understandability of the output sentences as the ability of the generator to correctly generate sentences improved if the most frequent lemma for a given synset was chosen. In the last experiment, dependency information is included to verify whether a potential lemma can occur in the given dependency context. With the addition of dependency information, the generator was able to produce more output for much more sentences. Moreover, the output contained more understandable sentences. It appears, therefore, that lexical choice should take into account both context and frequency information. We propose to model lexical choice in abstract dependency structures by Hidden Markov Trees, where we are given the synset of a node, and the task is to recover the hidden lemma, using the probability that a given lemma occurs in a specific dependency relation with another lemma, and the probability that a given synset occurs for a given lemma. A tree-modified Viterbi algorithm can then be used to recover the most probable dependency tree over lemma's.

# Linguistic Research with PaQu (Parse and Query)

Jan Odijk
Utrecht University
j.odijk@uu.nl

In this presentation I illustrate the use of PaQu (Parse and Query), an application being developed at Groningen University in the context of the CLARIN-NL project. PaQu enables a researcher to upload one's own (Dutch) language corpus and have it parsed by Alpino, after which the application makes it possible to search in the corpus for grammatical relations between two words. As an illustration, I use an example that I have used earlier (inter alia [Odijk 2011, 2014]) to investigate what is needed and desired in the CLARIN infrastructure: the distribution of the Dutch modifiers heel, erg and zeer (which are (almost) synonymous and mean 'very'), and the acquisition of this distribution by first language learners. In the presentation I will show how PaQu helps determining the distribution of these words in the adults' speech in the Dutch CHILDES corpora. For this experiment I had to prepare the data myself, but this should be not be necessary in the final version of PaQu. I will end with sketching requirements for PaQU (and similar applications for other grammatical search applications such as OpenSONAR and GrETEL).

# MT evaluation with BEER

Milos Stanojevic
ILLC, UvA
milosh.stanojevic@gmail.com

Khalil Sima'An
ILLC, UvA
k.simaan@uva.nl

Sentence level evaluation in MT has turned out far more difficult than corpus level evaluation. Existing sentence level metrics employ a limited set of features, most of which are rather sparse at the sentence level, and their intricate models are rarely trained for ranking. In this talk we will present our evaluation metric dubbed BEER which uses dense features for which counts are much more reliable and smoother on the sentence level. Examples of these features are precision and recall of character n-grams and PET(permutation tree) nodes. The weights of these features are estimated using learning-to-rank techniques in such a way to optimize correlation with human judgments. BEER was the best performing sentence level metric on the WMT14 evaluation task. We will give some analysis of what is wrong with the current metrics and which properties of BEER made it perform better than other state of the art metrics like METEOR and BLEU. BEER (together with the trained models) is available at https://github.com/stanojevic/beer

# Mapping from Written Stories to Virtual Reality

Oswaldo Ludwig
KU Leuven
oswaldoludwig@gmail.com

Quynh Do
KU Leuven
quynhdtn.hut@gmail.com

Marie-Francine Moens
KU Leuven
sien.moens@cs.kuleuven.be

This work is part of the project Machine Understanding for interactive StorytElling (MUSE), which introduces a new way of text understanding by “bringing text to life” through 3D interactive storytelling. The paper focuses on the mapping from the unstructured information found in knowledge representation of written stories, to a context-bounded and structured knowledge or meaning representation, suitable to be processed by a virtual reality module. That process yields an exponential explosion of instance combinations, since each sentence may contain a set of high-level concepts, each one giving place to a set of low-level instance candidates. The selection of the best combination of instances is a structured classification problem that yields a combinatorial optimization problem, which is computationally highly demanding and which is approached by a special genetic algorithm (GA) formulation, able to exploit the conditional independence among variables, while improving the parallel scalability. On the other hand, a large number of feasible combinations are usually related to a large uncertainty, which means high entropy, in the information theory sense, i.e. a high demand of information. Therefore, the automatic rating of the resulting set of instance combinations, i.e. the suitable text interpretations, demands an exhaustive exploitation of the state-of-the-art resources in natural language processing to feed the system with evidences to be fused by the proposed framework. In this sense, a statistical framework able to reason with uncertainty, integrating supervision, i.e. training with annotated examples, and evidence from external sources was adopted. The effectiveness of the proposed algorithm has been evaluated in detail on the story "The Day Tuk Became a Hunter" and is being currently evaluated on benchmark datasets.

# Methods for Part-of-Speech Tagging 17th-Century Dutch

Dieuwke Hupkes
University of Amsterdam
dieuwkehupkes@gmail.com

Rens Bod
University of Amsterdam
rens.bod@gmail.com

Automatically assigning POStags to languages for which little annotated training data is available is a challenging task in computational linguistics. When developing a POStagger for historical data one is confronted with an additional difficulty: a large variation in spelling. To tag historical Dutch texts, researchers often resort to taggers trained on modern Dutch data, although their adequacy for historical corpora is highly questionable. We present an analysis of this adequacy on 17th century Dutch data, and investigate the effect of modernising/normalising spelling and employing information extracted from a diachronic parallel corpus consisting of Dutch Bible texts from 1637 and 1977. We found that the baseline performance of a tagger trained on modern Dutch data is low (õ60%), but can be easily improved by applying a small set of rules to normalise spelling. Employing a more sophisticated method that makes use of alignments in the 17th- and 20th-century Bible versions results in an even higher within-domain score, but does not easily generalise to other 17th-century text. The best results (õ94% accuracy) on the 17th-century Bible text were achieved with a tagger trained on a corpus created by projecting the tags from the contemporary to the 17th-century version of the Bible via automatically generated word-alignments. Adding more words to the lexicon of the tagger is an important step in generalising this result to other domains. We argue that combining the various methods discussed in the paper allows for the development of a general diachronic tagger for historical Dutch texts.

# Modeling the learning of the English past tense with memory-based learning

Rik van Noord
University of Groningen
rikvannoord@gmail.com

Children's acquisition of the English past tense has been subject to a lot of debate. A number of rule-based and connectionist models have been proposed over the years, but the former approach usually has no explanation of how the rules are learned and connectionist models often use implausible vocabulary growth and feedback. We propose an approach that is able to address these criticisms, a memory-based learning model that is based on analogy. Other analogy models usually focused on modeling individual experiments on adults, hereby ignoring the learning phase in children. We were able to model the learning of the English past tense quite well, since our results were very similar to the rule-based model of Taatgen and Anderson (2002) and fit most of the findings in child language acquisition research. However, a more detailed analysis showed a number of implausibilities of memory-based learning that still need to be addressed. For one, our model is by definition not able to produce double-inflections (e.g. broked, wented). Second, there is an unusual high amount of irregular verbs that get the wrong irregular classification. Third, the degree of variability in the model is questionable, due to the absence of probabilistic decision making.

# Named Entity Disambiguation with two-stage coherence optimization

Filip Ilievski
VU University Amsterdam
filip.dbrsk@gmail.com

Marieke van Erp
VU University Amsterdam
marieke.van.erp@vu.nl

Piek Vossen
VU University Amsterdam
piek.vossen@vu.nl

Wouter Beek
VU University Amsterdam
w.g.j.beek@vu.nl

Stefan Schlobach
Vrije Universiteit Amsterdam
schlobac@few.vu.nl

To automatically populate knowledge bases, an array of text processing tools is currently used to extract events, recognize and link entities, and discover relations between the two. Contemporary Natural Language Processing modules solve Entity Linking, Event Detection, and Semantic Role Labeling as separate problems. From the semantic point of view, each of these processes adds another brush-stroke onto the canvas of meaning: entities and events are constants that occur in relations which correspond to roles. We present an approach where we extend such NLP processes with a semantic process of coherence optimization. In this paper, we present an approach where we use both binary logic and probabilistic models built through manual and automatic techniques. In the binary filtering phase, we use restrictions from VerbNet and a domain-specific ontology to jointly narrow down the possible interpretations of both the predicates and the entities. However, purely logical techniques are only able to exclude inconsistent joint interpretations, generally resulting in multiple logically coherent results. We argue that a second phase is needed in which the coherence between the remaining candidates is optimized in a probabilistic manner based on available background knowledge about the entities. Our optimization method assigns scores to the entity candidates based on factors such as graph distance, the number of shared properties, popularity metrics, and distance in the class hierarchy. The approach is evaluated on a domain-specific gold standard consisting of 120 manually-annotated news articles. Results are compared to existing entity rankers, such as DBpedia Spotlight.

# No longer lost in the forest..

Liesbeth Augustinus
CCL, KU Leuven
liesbeth@ccl.kuleuven.be

Vincent Vandeghinste
CCL, KU Leuven
vincent@ccl.kuleuven.be

Ineke Schuurman
CCL, KU Leuven
ineke@ccl.kuleuven.be

Frank van Eynde
CCL, KU Leuven
frank@ccl.kuleuven.be

We present the updated version of GrETEL (Greedy Extraction of Trees for Empirical Linguistics[2],), a search engine for the online exploitation of treebanks. There are two search modes: the XPath search mode and the example-based search mode. In the XPath search mode, the user can provide an XPath expression to query the treebanks. In the example-based search mode, it is possible to query treebanks with a natural language example instead of a formal search instruction. The input example is automatically parsed, enriched with additional input information from the user, and converted into an XPath expression, which is used to look for similar sentences in the treebank. GrETEL makes it thus possible to query treebanks without knowledge of (complicated) formal query languages or data formats. In the first version of GrETEL, two small treebanks were included. In the updated version, GrETEL was adapted for querying very large treebanks. More specifically, we included the SoNaR reference corpus for written Dutch (ca. 500M tokens, 41M sentences). Besides including more data, the GUI was adapted and extended based on user tests. We will show the results of both the GUI and the treebank adaptations in a demonstration of GrETEL 2.0.

[2] http://gretel.ccl.kuleuven.be

# On the Issue of Mining Fuzzy Text

Helena Blumhardt
University of Kassel
h.blumhardt@uni-kassel.de

Since the 1960s there has been a lot of research on the processing of text with the aid of machines – a field we nowadays refer to as text mining. Text mining was initiated in order to catalogue documents, but with the invention of Natural Language Processing techniques, the focus shifted towards the extraction of important or relevant information. The dissertation project described here will present text mining in a linguistic context. In this research the linguistic preprocessing techniques (stemming, tokenization etc.) that are usually regarded as a preparatory step in a text mining system will be highlighted as independent text mining tasks. Additionally, what will be provided is terminological groundwork in a fuzzy textual domain, a notion inspired by Lotfi Zadeh and his introduction of the concept of fuzzy logic (1965). For demonstration purposes the domain of international trade unions will be used, since the terminology in this domain is influenced by many different fields such as political science and other social sciences. The problem of fuzziness already occupied the minds of philosophers in ancient Greece. Remarkably enough, these early considerations can be applied to current issues in linguistic research – namely textual domains. Compared to domains from natural sciences as biology or medicine, fuzzy textual domains do not offer a clear linguistic structure, which makes it difficult to use statistical text mining methods and which will be one of the main issues of this project.

# Open Source Dutch WordNet

Marten Postma
VU University Amsterdam
martenp@gmail.com

Piek Vossen
VU University Amsterdam
p.t.j.m.vossen@vu.nl

We present Open Source Dutch Wordnet (OWDN1.0): an open source version of Cornetto (Vossen et al., 2013). Cornetto is currently not distributed as open source, because a large portion of the database originates from the commercial publisher Van Dale. We use English WordNet 3.0 (Miller, 1995; Fellbaum, 1998) as our basis. This means that we replace the Van Dale synsets and internal semantic relations by Wordnet synsets and internal semantic relations. We replace the English content in English WordNet in two ways: 1. (a) When there exists an external semantic relation (ESR) between a Cornetto synset and a WordNet synset, all synonyms from the Cornetto synset are inserted into the WordNet synset. Only synonyms that originate from the Referentie Bestand Nederlands (RBN1) are included. Synonyms from Van Dale are ignored. (b) The ESR relations in Cornetto are mostly generated automatically, resulting in many scored alternatives. To get a more optimal performance, they were therefore first filtered before step 1(a) was applied. Four students manually checked 12,966 relations, of which 6,575 were removed. Afterwards, the unchecked semantic relations were filtered using a decision tree algorithm that used the manual inspection as training. This resulted in a removal of 32,258 ESRs (with an F-score of 0.80, as evaluated on the manual set) 2. Using open source resources (Wikipedia (Wikipedia, 2014;Foundation, 2014a), Wiktionary (Foundation, 2014b), Google Translate (Google, 2014), Babelnet (Navigli 2010)), the English synonyms in English WordNet are translated into Dutch. Open Source Dutch WordNet contains 116,992 synsets, of which 95,356 originate from WordNet 3.0 and 21,636 synsets are new synsets. The number of English synsets without Dutch synonyms is 60,743, which means that 34,613 WordNet 3.0 synsets have been filled with at least one Dutch synonym. Dutch synonyms added through step 1) have a mapping to lexical units in RBN. Synonyms added through step 2) and not through step 1) have no equivalent to RBN. The obvious next step will be to find Dutch synonyms for the 60K synsets that have not been filled. We also observed that the ESR relations in Cornetto for adjectives are not very reliable. Currently, the database therefore only includes nouns and verbs. For future work, we plan to add adjectives through method 2.

# Open-domain extraction of future events from Twitter

Florian Kunneman
Radboud University
f.kunneman@let.ru.nl

Antal van den Bosch
Radboud University
a.vandenbosch@let.ru.nl

Many future social events are explicitly mentioned on the social media platform of Twitter. We developed a system that leverages such mentions in Dutch and Flemish tweets in order to generate a calendar of future events. The extracted events have a wide range of types and might take place in the near or distant future. While the overview of future events is certainly not complete, it does represent the bulk of future events that are popular on Twitter. We provide an overview of the different components of our system, giving special attention to the challenges that were involved in extracting events from Twitter. We also report on its performance on a month of Dutch tweets and display subsequent analyses. The system is called 'Lama Events' and can be found on[3].

[3] http://applejack.science.ru.nl/lamaevents/

# PICCL: Philosophical Integrator of Computational and Corpus Libraries

Martin Reynaert
TiCC - Tilburg University and CLTS - Radboud University Nijmegen
reynaert@uvt.nl

PICCL constitutes a complete workflow for corpus building. It is the integrated result of developments in the CLARIN-NL @PhilosTEI and NWO Nederlab projects. Input can be either images or text. Images may be e.g. the scanned pages of a book in DjVu, PDF or TIFF formats. Text may be plain, in Office formats, embedded in PDF or in the OCR engine output formats, i.e. hOCR HTML, Page XML or Alto XML. The open-source workflow is a CLAM web service/application. FoLiA XML is its pivot format. The workflow can handle texts in a broad range of – currently – European languages. Provisions are available for dealing with old print or diachronical language variation. Images are converted into electronic text by Tesseract. Language classification may be performed at the paragraph level. TICCL or Text-Induced Corpus Clean-up performs automatic post-correction of the OCRed text. Output text is in FoLiA XML. Optionally, Dutch texts may be annotated by Frog, i.e. lemmatized and classified for Parts of Speech and Named Entities. The texts are then ready for indexing and integrating in the Nederlab diachronic Dutch corpus or perhaps for extending OpenSoNaR, the online version of the contemporary written Dutch corpus SoNaR. For allowing e.g. philosophers to build critical editions of books, TEI XML is produced. PICCL has a highly intuitive user-friendly interface to allow even the most computer-weary user to obtain texts in a corpus-ready, annotated format. PICCL is available to all researchers in the CLARIN infrastructure.

# Part-of-Speech Tagging of Twitter Microposts only using Distributed Word Representations and a Neural Network

Fréderic Godin
Ghent University - iMinds
frederic.godin@ugent.be

Wesley de Neve
Ghent University - iMinds
wesley.deneve@ugent.be

Rik van de Walle
Ghent University - iMinds
rik.vandewalle@ugent.be

Many algorithms for natural language processing rely on manual feature engineering. However, manually finding effective features is a labor-intensive task. Moreover, whenever these algorithms are applied on new types of content, they do not perform that well anymore and new features need to be engineered. For example, current algorithms developed for Part-of-Speech (PoS) tagging of news articles with Penn Treebank tags perform poorly on microposts posted on social media. As an example, the state-of-the-art Stanford tagger trained on news article data reaches an accuracy of 73% when PoS tagging microposts. When the Stanford tagger is retrained on micropost data and new micropost-specific features are added, an accuracy of 88.7% can be obtained. We show that we can achieve state-of-the-art performance for PoS tagging of Twitter microposts by solely relying on automatically inferred distributed word representations as features and a neural network. To automatically infer the distributed word representations, we make use of 400 million Twitter microposts. Next, we feed a context window of distributed word representations around the word we want to tag to a neural network to predict the corresponding PoS tag. To initialize the weights of the neural network, we pre-train it with large amounts of automatically high-confidence labeled Twitter microposts. Using a data-driven approach, we finally achieve a state-of-the-art accuracy of 88.9% when tagging Twitter microposts with Penn Treebank tags.

# Polemics Visualised: experiments in Syriac text comparison

Hannes Vlaardingerbroek
VU University Amsterdam
hannes@vlaardingerbroek.nl

Marieke van Erp
VU University Amsterdam
marieke.van.erp@vu.nl

Wido van Peursen
VU University Amsterdam
w.t.van.peursen@vu.nl

Many research fields are increasingly relying on automatic text analysis to further their research. For many modern languages, various text analysis tools are available, but such tools can also be a useful research on ancient texts such as analysing >500,000 words of old Syriac Aramaic stemming from a formative period of the early church. This corpus is hardly studied because of a lack of resources. The Polemics Visualised project aims to adapt existing computational analysis tools that have been developed for modern languages to the Syriac language. Although some computational tools exist for Syriac (e.g. SyroMorph [1], developed for the Peshitta New Testament), resources are scarce. With Polemics Visualised, we aim to extend the usability of existing tools to more obscure texts, with as main objective visualising a topical comparison between the writings of the second century theologian Bardaisan and the polemics against him by the fourth century church father Ephrem. We present experiments in which we show the flexibility of our developed tools for Syriac by analysing and comparing the topics in Bardaisan and Ephrem. We perform this analysis by first preprocessing the texts (tokenisation, part-of-speech tagging etc) after which we extract topics from salient paragraphs which are then compared between the two authors.

# Predicting OOV pronunciations for TTS using FSTs

Esther Judd-Klabbers
ReadSpeaker
esther.judd@gmail.com

One essential component in a text-to-speech system is a pronunciation prediction component that produces pronunciations for words not in the lexicon. Pronunciations can be predicted via rewrite rules or machine-learning mechanisms. The pronunciation of a word consists of phonemes and syllable and stress markers. For most languages there is no one-to-one mapping between letters (graphemes) and phonemes, and thus there needs to be an alignment between these two representations. Because there are rules that govern the language at a higher level we need to look at a longer window of symbols. In our machine-learning approach to pronunciation prediction we use n-gram models to train a Finite-State-Transducer (FST) that takes as input a string of graphemes and outputs a string of symbols (phonemes, syllable markers, stress markers). We present an integrated approach that predicts the full symbol set at once. We compare it to the existing CART approach and discuss ways to optimize the resulting FST. Optimization includes use of a constraint Finite-State-Acceptor that dictates that each syllable has one vowel, and that there is only one primary stressed syllable in a word. Using k-fold cross-validation where the lexicon is divided into n blocks and the FST is trained on 90% and tested on 10% of the lexicon shows that CART and FST are quite evenly matched on the grapheme-phoneme conversion portion, but stress prediction is far superior in the FST approach.

# Predicting concreteness and perceivability

Emiel van Miltenburg
VU University Amsterdam
emiel.van.miltenburg@vu.nl

Concreteness ratings are often used as a measure of readability. There have been several attempts to build a model that accurately predicts human judgments of concreteness, but none of them incorporate a measure of sensory perceivability. This is striking because of two reasons: (1) it is a salient aspect of concrete terms that they tend to denote objects that are more directly experienced, (2) recent literature shows that sensory perceivability is a strong predictor of readability. We created a model to predict concreteness as well as sensory perceivability ratings. We looked at factors common in the literature, and in addition our model is enriched with corpus data indicating: * The relative frequency of perception related modifiers (e.g. blue, tasty, rough) occurring with each noun. Perception-relatedness was determined using WordNet mappings from the SUMO ontology. * The relative frequency of kind-level modifiers (e.g. academic, vegetarian, military) occurring with each noun. These modifiers show us that the noun involved is a higher-level noun that can have subclasses. Whether a modifier is a kind-level modifier is determined on the basis of their morphology (ending in –ic, -ary, -ian). We show the performance of our model on human ratings, and discuss some hidden assumptions behind recent studies. For example: Feng et al.'s (2011) model assumes that concreteness ratings are based on all senses of a word. Is this true, or do participants in concreteness rating studies base their judgments only on the predominant sense?

# Proof-of-Concept Experiments for the Fine-Grained Classification of Cyberbullying Events

Cynthia van Hee
LT3, Ghent University
cynthia.vanhee@ugent.be

Ben Verhoeven
CLiPS, University of Antwerp
ben.verhoeven@uantwerpen.be

Els Lefever
LT3, Ghent University
els.lefever@ugent.be

Guy de Pauw
CLiPS, University of Antwerp
guy.depauw@uantwerpen.be

Veronique Hoste
LT3, Ghent University
veronique.hoste@ugent.be

Walter Daelemans
CLiPS, University of Antwerp
walter.daelemans@uantwerpen.be

In the current era of online interactions, both positive and negative experiences are abundant on the web. As in real life, these negative experiences can have a serious impact on youngsters. Recent research reports cybervictimization rates among teenagers between 3% and 24% (Olweus, 2012; Hinduja & Patchin, 2012). In the research project AMiCA, we strive to automatically detect harmful content such as cyberbullying on social networks. We collected data from social networking sites and by simulating cyberbullying events with volunteer youngsters. This dataset was annotated for a number of fine-grained categories related to cyberbullying such as insults and threats. More broadly, the severity of cyberbullying, as well as the author's role (harasser, victim or bystander) were defined. We present the results of our preliminary experiments where we try to determine whether an online message is part of a cyberbullying event. Moreover, we explore the feasibility to classify online posts in five more fine-grained categories. For the binary classification task of cyberbullying detection, we obtain an F-score of 54.71%. F-scores for the more fine-grained classification tasks vary between 20.51% and 55.07%.

# Real Time Classification of Conceptually Related Tweets

Parma Nand
Auckland University of Technology
pnand@aut.ac.nz

Rivindu Perera
Auckland University of Technology
rperera@aut.ac.nz

This paper presents an architecture for downloading conceptually related Tweets in real time. The objective of the architecture is to be able to download Tweets related to a concept which is represented by a noun phrase. We use DBpedia as the knowledge source to enrich the concept corresponding to a given set of words which is then used as a basis for selecting similar Tweets. The architecture consists of 4 modules. The first one, Locality module, parses the noun phrase to establish the relevant locality of the concept, since there might be multiple pages for different localities related to a concept. The second module, concept enrichment, searches for the DBpedia page related to the concept in an appropriate locality. The third module, Similarity calculator, uses a selection of similarity computation algorithms to compute the similarity of newly arriving Tweets to the enriched concept. The resulting similar Tweets can then be used to provide a visual summary of the information contained in the set of Tweets (this is not part of this paper). The architecture was tested with a set of 1500 downloaded Tweets related to concepts around 2014 national elections in New Zealand. The set of 1500 Tweets data was annotated by postgraduate NLP students into 3 nonexclusive concepts with a Cohen's Kappa coefficient of 0.87. Using this data we obtained precision values in the vicinity of 0.85 and recall values in the vicinity of 0.74. We intend to make both the data and the architecture available as open source software for academic research.

# Recognizing Humor with Web's Joint Probability

Abdullah Alotayq
Imam University
asalotayq@imamu.edu.sa

Humor is categorized as the tendency of particular cognitive experiences to provoke laughter and provide amusement. Many theories exist about what humor is and what social function it serves. Psychological theories stand up first with many attempts to extend human knowledge with regard to such human behavior aspect. Humor has been a subject for research in NLP, and other fields. It has been considered one of the most complex human behaviors. With regard to what makes a laugh, I found that readings in philosophy, psychology, and linguistics have all pointed to the fact that humor exists when some incoherency is found within the structure of some statement. This paper presents a method to detecting humor within question-answer data. Web's joint probability information is used to help in the recognition of humorous content. The method was tested on a number of 2k data, and a significant improvement over the baseline was achieved. More information with regard to different thresholds, and extent-span is showed as well. This work tries to prove the fact that humor contains some sort of inconsistency within the utterance's structure, and it's detectable by the use of very large background-knowledge. The web is the biggest corpus ever existed for written text. This makes it a good resource for information to mine for "inconsistency". In this paper, I bring some evidence that computational approaches can be successfully applied to the task of humor recognition.

# Relating language and sound: two distributional models

Alessandro Lopopolo
VU University Amsterdam
a.lopopolo@vu.nl

Emiel van Miltenburg
VU University Amsterdam
emielonline@gmail.com

We present preliminary results in the domain of sound labeling and sound representation. Our work is based on data from the Freesound database, which contains thousands of sounds complete with tags and descriptions, under a Creative Commons license. We want to investigate how people represent and categorize different sounds, and how language reflects this categorization. Moreover, following recent developments in multimodal distributional semantics (Bruni et al. 2012), we want to assess whether acoustic information can improve the semantic representation of lexemes. We have built two different distributional models on the basis of a subset of the Freesound database, containing all sounds that were manually classified as SoundFX (e.g. footsteps, opening and closing doors, animal sounds). The first model is based on tag co-occurrence. On the basis of this model, we created a network of tags that we partitioned using cluster analysis. The clustering intuitively seems to correspond with different types of scenes. We imagine that this partitioning is a first step towards linking particular sounds with relevant frames in FrameNet. The second model is built using a bag-of-auditory-words approach. In order to assess the goodness of the semantic representations, the two models are compared to human judgment scores from the WordSim353 and MEN database.

# Robust Language Processing in Fluid Construction Grammar. A Case Study for the Dutch Verb Phrase

Paul van Eecke
Sony Computer Science Laboratory Paris
vaneecke@csl.sony.fr

The Dutch verb phrase (VP) is notorious for its syntactic intricacies. There are three main reasons why it is so difficult to robustly handle this complexity in a processing model. First of all, a single VP can contain multiple modal auxiliaries ('modal stacking'). Second, the perfect aspect can be expressed on any of the verb forms, so either on the lexical verb or on a modal auxiliary. Finally, there are various word orders in which the verb forms can appear. This presentation consists of two parts. In the first part, I will demonstrate a fully operational processing model of the Dutch VP in Fluid Construction Grammar (FCG) that works for both parsing and production. The implementation shows that the aforementioned challenges can be overcome by carefully managing the hierarchical relations between the elements of the verb phrase. Next, I will illustrate the validity of the approach by showing that the model remains robust when faced with erroneous input. More specifically, I used a modified version of the grammar to intentionally produce, based on correct meaning representations, a corpus of verb phrases containing errors (e.g. a finite form instead of a past participle). I will show that my robust parsing algorithm can recover the intended meaning and correct the verb form using a strategy based on re-entrance. This strategy consists of parsing as much meaning as possible from the input, constructing hypotheses about the intended meaning, producing utterances expressing these hypotheses and calculating which utterance is the closest to the input.

# Strong 'islands of resilience' in the weak flood. Dutch strategies for past tense formation implemented in an agent-based model

Dirk F. Pijpops
University of Leuven
dirk.pijpops@kuleuven.be

Katrien Beuls
Vrije Universiteit Brussel
katrien@ai.vub.ac.be

Since Germanic parted from the Indo-European mother tongue, the past tenses of most Germanic languages have formed the battle ground of the weak and strong inflection systems. This opposition has often been used as a case study of regular-irregular competition in both computational and statistical models of language (e.g. Hare & Elman 1995, Lieberman 2007, Pugliesi et al. 2014, Pijpops et al. 2014). However, weak and strong are not synonymous to regular and irregular (Carrol et al. 2012). In Dutch and English, one may find irregular weak forms, such as 'zocht-sought' and 'dacht-thought', while some strong ablaut patterns form 'islands of resilience' (Mailhammer 2007), exhibiting clear regularity, and even incidental productivity and expansion (Salverda 2006: 170-179, Knooihuizen & Strik 2014). Our agent-based model therefore no longer means to model a regular-irregular distinction, but rather a choice between a generally available dental suffix and several vowel-dependent ablaut patterns. Agents will have to employ either strategies to communicate past events to one another, and adapt their own linguistic behavior to that of their fellows. As a starting position, the present Dutch situation is used, with forms and frequencies extracted from the Corpus of Spoken Dutch (Van Eerten 2007). The model itself is built within the Babel2-framework (Loetzsch et al. 2008), with the agents' grammars operationalized in Fluid Construction Grammar (Steels 2011). Amongst the questions we seek to answer are: What mechanisms do we need to implement to obtain realistic results? How are some ablaut patterns better able to resist the weak expansion?

# Suhuf: Morpho-Syntactically Tagged Islamic Corpora

Mahmoud Shokrollahi-Far
Tilburg University
m.shokrollahifar@uvt.nl

Peyman Passban
Dublin City University
ppassban@computing.dcu.ie

Arabic language has always been a scientific focus for computational linguistics (CL), because of which the dominant data-driven approach in the field requires large amount of Arabic corpora. However, the limited amount of reliably tagged corpora usually annoys the researchers in Arabic CL. More challenging is the lack of relevant tagged corpora for CL research in Islamic texts, a troubling fact that is nowadays more felt. A usual reason for such a lack of tagged corpora is the lack of appropriate and reliable taggers. The present paper is reporting on a new collection of Islamic corpora that is morpho-syntactically tagged employing Mobin, a knowledge-based system which reliably bootstraps grammatical tags in Arabic vowelized texts. The corpora, called Suhuf, at the moment consists of the Quran together with few other Arabic books on the sayings of Prophet Mohammad and his successors, all amounting to 250k words. The corpora are soon going to be freely available for research purposes at /www.quranmining.net/ website. Utilizing Suhuf we have implemented many text mining researches in Islamic texts going to be briefly reported, as well.

# Suicidality detection in social media

Bart Desmet
LT3, Ghent University
bart.desmet@ugent.be

Véronique Hoste
LT3, Ghent University
veronique.hoste@ugent.be

Suicide prevention hinges on adequate and timely risk detection. Online platforms are increasingly used for expressing suicidal thoughts, but prevention workers are faced with an information overload when monitoring for such signals of distress. We present ongoing work on suicidality detection in Dutch user-generated content, using text classification. We developed a new annotation scheme grounded in prevention practice, and annotated a gold standard corpus consisting of message board and blog posts. Annotations indicated a post's relevance to suicide, and the subject and severity of a suicide threat, if present. Two detection tasks were derived from this: of suicide-related posts, and of severe, high-risk content. In a series of experiments, we sought to determine how well these tasks can be carried out automatically, and which information sources and techniques contribute to classification performance. We study a range of features, and test the effects of model optimization through feature selection, hyperparameter optimization or a combination of both, and of cascaded classification for the detection of severe messages.

# Synset embeddings help named entity disambiguation

Minh N. Le
VU University Amsterdam
minhle.r7@gmail.com

Many algorithms of named entity disambiguation (NED) hinge on the notion of semantic relatedness. A measure of semantic relatedness is simply a way to assign a number to each pair of entities and there are many such ways, each leads to a certain performance. So far, most measures have been based on overlapping text or linked entities, sometimes augmented by graph algorithms such as random walk with restart. I will present an ongoing work on a novel semantic relatedness measure based on synset embeddings. A BabelNet synset is a WordNet synset, a Wikipedia title or both. The combination of WordNet lexical relations and ontological relations from DBpedia creates a rich source of knowledge on which I train a feedfoward neural network. For each triple of two synsets and a relation, it learns to predict the likelihood of the triple being a genuine one or one that was generated from a probability distribution. Using this model, I can map both concepts and entities into points in a common multidimensional space. The proximity of those points is then used as a measure of semantic relatedness. I will evaluate the measure on standard datasets as it is plugged in a graph-based collective NED algorithm. Because the measure provides relatedness of entities and concepts alike, it is possible to use disambiguated words to improve NED. I will also present the effectiveness of this method.

# Syntax-based fuzzy matching in translation memories

Tom Vanallemeersch
Centre for Computational Linguistics, KU Leuven
tallem@ccl.kuleuven.be

Vincent Vandeghinste
Centre for Computational Linguistics, KU Leuven
vincent@ccl.kuleuven.be

Computer-aided translation (CAT) tools support translators through various means, such as looking up sentences in a translation memory (TM). When a sentence to translate is equal or similar to a sentence in TM (exact or fuzzy match), the translation of the latter sentence is suggested to the translator, who can accept or edit it. Current CAT tools mostly apply string-based fuzzy matching: they apply edit distance to detect the differences between sentences, and typically use limited or no linguistic knowledge in this process. In the recently started SCATE project (Smart Computer-Aided Translation Environment), which aims at improving translators' efficiency, we apply syntax-based fuzzy matching in order to detect abstract similarities and to select matches which lead to more useful translation suggestions than those produced by the baseline, string-based edit distance. Our syntax-based fuzzy matching procedure uses parses of sentences which identify constituents and/or dependencies. We apply the Translation Error Rate (TER) metric, which originates from the field of machine translation (MT), to compare the translation suggestions produced by the baseline with those resulting from syntax-based fuzzy matching. The TER scores express the editing effort needed to convert a suggestion into the desired translation. We also calculate TER scores for translation suggestions which merely consist of the translation of the parts of the sentence in TM that match with the sentence to translate. The translation of these parts is determined through the word alignment produced by a statistical MT system.

# Taking into account linguistic structure improves information processing tasks

Koos van der Wilt
N/A
kooswilt@gmail.com

Electronic text will proliferate in the future and become a tool with which research is done, journalism is performed, and, overall it will play a large role in communication. Document collections contain documents that need to be summarized, put in clusters, and classified, for instance. The hope is this is easier to do with electronic text than with paper text. There is already quite a bit of software that seeks to perform these tasks, and research is ongoing. Most of the software seeking to perform these tasks, however, is based on statistics. My paper seeks to support the hypothesis that involving the structure of language in these information-processing tasks improves their performance. We investigate a formula from information theory to do document classification, and study if and how document classification improves when we take the syntactic structure of the sentences that make up a text into account.

# Text-to-pictograph translation for six language pairs

Leen Sevens
University of Leuven
sevensleen@hotmail.com

Vincent Vandeghinste
University of Leuven
vincent@ccl.kuleuven.be

Ineke Schuurman
University of Leuven
ineke@ccl.kuleuven.be

Frank van Eynde
University of Leuven
frank@ccl.kuleuven.be

The Able-To-Include project aims to improve the living conditions of people with Intellectual or Developmental Disabilities. As failure to access or use ICT (email, social media) is considered a major form of social exclusion, the goal is to build an "accessibility layer" that is based on three key technologies, automatic translation from pictographs to text being one of them. We present the latest version of the Text2Picto translation system, which works for Dutch, English and Spanish to Beta or Sclera pictographs. We describe how our original Dutch Text2Picto system has been modified in order to be usable for English and Spanish, and we present the results of our evaluation experiments, revealing the progressive improvements of the systems when adding more linguistic background, such as lemmatizers and WordNet synsets.

# The Greek Discourse Relations Corpus: Main Challenges and Prospects

Alexandros Tantos
Aristotle University of Thessaloniki
alextantos@lit.auth.gr

Konstantinos Vlachos
Aristotle University of Thessaloniki
vlackons@lit.auth.gr

Katerina Lykou
Aristotle University of Thessaloniki
kate_lyk@hotmail.com

Meropi Papatheohari
Aristotle University of Thessaloniki
papathem@lit.auth.gr

Antonia Samara
Aristotle University of Thessaloniki
antosama@lit.auth.gr

Georgios Chatziioannidis
Aristotle University of Thessaloniki
georgidc@lit.auth.gr

The aim of the talk is twofold: a) to present a short scale project (funded by the Aristotle University of Thessaloniki) of building and analyzing the first annotated corpus with discourse relations in Greek, the Greek Discourse Relations' Corpus (GDRC), and b) to examine closely the notion of discourse topic and discourse coherence through quantitative analysis based on the annotated corpus. The set of discourse relations included in the inventory of the annotation scheme for GDRC is partially adopted from Segmented Discourse Representation Theory (SDRT) as defined by Asher and Lascarides (2003). Apart from discourse semantic annotation, we include lexically-based semantic annotation, mostly related and the thematic roles of verbal arguments, that supports discourse inference and interpretation. The praxis of annotating our corpus has shown a number of difficulties in the definition and usage of discourse relations related to both the nature of utterance relations and the algorithm of segmenting discourse utterances. Moreover, the notion of discourse topic construction, central for a number of discourse relations such as Narration, is hard to pin down and although it is extensively examined by Asher (2004), it seems a difficult task to achieve inter-annotator agreement and, thus, it reflects the diversity of the term and the difficulty of creating a common sense about the concept of topic; especially in observed cases where the topic shift signals an abrupt change of the topic but still allows for discourse continuity. For this reason, a) we propose a flexible tagset that reflects the textual reality and b) we use standoff annotation as implemented by Linguistic Annotation Framework for achieving balanced discourse-and lexically-based annotation.

# The third way: triplet description of Dutch orthography

Johan Zuidema
Radboud University Nijmegen
joh.zuidema@online.nl

Anneke Neijt
Radboud University Nijmegen
a.neijt@let.ru.nl

The BasisSpellingBank (BSB) follows a lexical approach based on triplets which specify a phoneme or string of phonemes (Zuidema&Neijt 2012): (1) riet'reet' dimensie'dimension' phonemes [r i t] d i= [m E n]= s i [. . . ] main stress graphemes r ie t d i= m e n= s ie dv default value relation dv dv dv dv 3.3 dv dv dv 3.13 4.4 = syll.boundary The first triplet in (1) is {[r,r,dv}, representing {phoneme,grapheme,relation}. Relationships other than the default value are indicated by a numerical reference to spelling categories which are relevant for spelling instruction. All phonemes of riet are spelled according to the default, but some phonemes of dimensie are spelled differently. In the lowest tier, 3.3 refers to /i/ in non-native words (not spelled with default ie), 3.13 refers to /s/ in non-native words (often pronounced as /z/), and 4.4 is for morpheme-final /i/. Triplets are the smallest strings of phonemes and graphemes relevant for the relations one needs to learn. The BSB contains triplet descriptions of 100.000 words (18.000 lemmas plus inflected forms) learned at primary school. These words are described in 600.000 triplet tokens, 4000 triplet types, of which only 1000 are frequently used. In our presentation we will show the advantages of a triplet analysis. For instance, corpus-based forward and backward consistency of the grapheme-phoneme relation and detailed analysis of spelling tests. Zuidema, J. & A. Neijt (2012) Verkennend onderzoek naar de wenselijkheid en de haalbaarheid van een verrijking van de Woordenlijst Nederlandse Taal ten behoeve van spellingonderwijs. Rapport in opdracht van de Nederlandse Taalunie[4].

[4] http://taalunieversum.org/sites/tuv/files/downloads/rapport%20VWS%2015022013.pdf

# Topic Modelling in Online Discussions

Chris Emmery
CLiPS, University of Antwerp
chris.emmery@uantwerpen.be

Menno van Zaanen
TiCC, Tilburg University
mvzaanen@uvt.nl

Methods for extracting information from social media and their application in machine learning and text mining tasks have been well studied over the last few years. As a result, the issues that the large amounts of unlabelled, noisy data still pose for NLP-related tasks are well-known. The current research chooses a different forum for the topical analysis of user-generated content, and shows its potential for further research. By focussing on the discussions present in large Dutch news communities, some of the mentioned issues are circumvented. These discussions consist of so-called comments, which are still unlabelled. However, they are to a large extent topically framed by their relation with a posted news article. The author-provided tags of these articles can therefore be used as topical labels for supervised learning. Obviously, not all comments are on-topic with the associated news article; however, by naively assuming that they predominantly are, L-LDA (Ramage, Hall, Nallapati, & Manning, 2009) performs well in overcoming this label noise. We show that, despite the gold standard handicap, reliable topic distributions are still learnt from these comments when inferring and evaluating topic labels for the articles. As such, we create a correct evaluation of the classifier performance, and can assume reliability of analyses made with these classifications. Most importantly, we demonstrate that the classifications made can be used to quantify these discussions to be effectively employed in trend analysis.

# Towards a Diachronic Semantic Lexicon of Dutch

Katrien Depuydt
Instituut voor Nederlandse Lexicologie
katrien.depuydt@inl.nl

Jesse de Does
Instituut voor Nederlandse Lexicologie
jesse.dedoes@inl.nl

The release of Open Dutch Wordnet is a major event in the development of a truly open semantic lexicon for Dutch. In this contribution, we will discuss how the plethora of historical lexicographical data from the major scientific dictionaries of Dutch (ONW, VMNW, MNW, WNT) could be used to enrich this lexicon with diachronic information. We propose extensions to the core data model in order to incorporate the time dimension, and discuss what would be needed to account for shifts in word meaning in a systematic way. The dictionaries however are not the only starting point. Corpus-based computational techniques for detecting semantic change, and conversely, finding cross-temporal near-synonyms cannot be ignored. We will consider how methods for (semi-)automatical data acquisition can be used in combination with the lexicographical data. In this connection, one of the principal challenges for a diachronic wordnet , built from the scholarly dictionaries of Dutch, is the level of detail of the semantic descriptions. We discuss strategies for clustering closely related word senses.

# Tracking Linguistic Complexity in Second Language Writing: A Sliding-Window Approach

Marcus Ströbel
RWTH Aachen University
marcus.stroebel@rwth-aachen.de

Elma Kerz
RWTH Aachen University
kerz@anglistik.rwth-aachen.de

Daniel Wiechmann
University of Amsterdam
D.Wiechmann@uva.nl

Recent years have seen the emergence of the Complexity-Accuracy-Fluency (CAF) triad as a prominent conceptual framework to assess second (L2) language proficiency. However, there is still much controversy in regard to how the CAF constructs should be operationally defined and measured (cf., e.g. Housen et al., 2012). This particularly applies to the dimension of complexity, often defined along a wide range of syntactic, lexical, and phraseological indicators (cf. Wolfe-Quintero et al., 1998). Guiding current efforts to better understand the relevant dimensions of linguistic complexity in L2 learning, Lu (2010, 2012) presents two computational tools that automatically generate a wide range of syntactic and lexical complexity measures of learner written productions of English. Like other more general applications for the automatic analysis of texts used in contexts of assessment of text readability, Lu's applications produce point estimates of all relevant measures. In this contribution, we present a novel approach to the automatic assessment of complexity, which adopts a sliding-window technique to track the changes in complexity within a text, producing what we term "complexity contours". We demonstrate the relevance of this informational gain on the basis of a classification task, in which indicators of linguistic complexity are used to discriminate the written productions of advanced learners and domain experts and find that the employment of complexity contours strongly improves classification accuracy across indicators. The assessment of the complexity contours of various indicators also permits the discovery of compensatory effects of different constitutive aspects of linguistic complexity (e.g. lexical and syntactic aspects).

# Translation-based Word Clustering for Language Models

Joris Pelemans
KU Leuven
joris.pelemans@esat.kuleuven.be

Hugo van Hamme
KU Leuven
hugo.vanhamme@esat.kuleuven.be

Patrick Wambacq
KU Leuven
patrick.wambacq@esat.kuleuven.be

One of the major challenges in the field of language modelling (and others) is data sparsity. Even with the increasing amount of data, there is simply not enough data to reliably estimate probabilities for short word sequences, let alone full sentences. Hence, research in this field has focused largely on finding relations between words or word sequences, inferring probabilities for unseen events from seen events. In this work we focus on a new approach to cluster words by examining their translations in multiple languages. That is, if two words share the same translation in many languages, they are likely to be (near) synonyms. By adding some context to the hypothesized synonyms and by filtering out those that do not belong to the same part of speech, we are able to find meaningful word clusters. The clusters are incorporated into an n-gram language model by means of class expansion i.e. the contexts of similar words are shared to achieve more reliable statistics for infrequent words. We compare the new model to a baseline word n-gram language model with interpolated Kneser-Ney smoothing.

# Tree models, syntactic functions and word representations

Simon Suster
University of Groningen
s.suster@rug.nl

Gertjan van Noord
University of Groningen

Ivan Titov
University of Amsterdam

In this talk, we propose a method for learning word representations from trees using syntactic functions. Our point of departure is EM-style learning of hidden Markov models, which have been shown to provide effective features, either discrete or continuous, for many NLP tasks. We first revisit the relationship between chain models and those induced from unlabeled dependency trees, and show that the advantage of tree representation over chains is not self-evident. In a bare-bones tree model, the tree only defines the context for a word, whereas we seek to include the nature of the relationship between a child and its parent, as this is indicative, for example, of different semantic roles an argument takes. The proposed extension concerns the introduction of syntactic functions as an additional observed variable in the hidden Markov tree model. The learning and inference procedures remain similar. When using selected syntactic functions, these models improve on plain tree models in certain experiments. We evaluate the word representations, trained on Dutch and English, on two tasks – named entity recognition and frame-semantic parsing. In general, our models have proven to be competitive with other state-of-the-art word representation techniques.

# Tweet Stream Analysis for Flood Time Estimation

Ali Hürriyetoğlu
Radboud University
a.hurriyetoglu@let.ru.nl

Antal van den Bosch
Radboud University
a.vandenbosch@let.ru.nl

Nelleke Oostdijk
Radboud University
n.oostdijk@let.ru.nl

Estimating the remaining time to natural disasters, in time, can prevent the loss of lives and properties. Therefore, we analyze the Twitter stream and focus on floods, in order to identify predictive features that signal the time of the impending event. As features we use the text, time of the post and mentioned times, user information, and geolocation. The spatiotemporal distribution of the tweets, linguistic features, and communicative patterns are compared to actual flood information with the aim of building a predictive model for estimating the time of future floods. We use tweet data and flood information that are provided by FloodTags and Twiqs.nl. Tweets are collected by using related key terms and filtered to eliminate noisy tweets that can interfere with the event time signal. The content quality (veracity) provided by each user is taken into consideration as well. We discuss how these types of naturally occurring, unscheduled events differ in their time-to-event predictability from organized events, which we investigated in previous work.

# Twitter Ngram Frequencies

Gosse Bouma
University of Groningen
g.bouma@rug.nl

Although collections of Twitter messages are accessible for linguistic research at various sites, obtaining frequency figures for large numbers of ngrams can still be challenging. We present a database that contains ngram frequencies obtained from a corpus of over 2 billion Dutch tweets, and illustrate some examples of its use for linguistic research. Ngram frequencies are based on a corpus of Dutch tweets collected by the University of Groningen. We used a language guesser to filter non-Dutch tweets, removed duplicates in each monthly section of the corpus, and used the Alpino tokenizer for tokenization. We did not convert words to lower case and left hashtags in place, but usernames and URLs were replaced by generic placeholders. Using a frequency cut-off of 10, the database contains approximately 6.5M distinct unigrams, 50M bigrams, 124M trigrams, 136M 4-grams, and 118M 5-grams. The search interface allows users to study spelling mistakes (i.e. by comparing frequencies for "hij wordt" vs. "hij word", spelling variants (i.e. "ze%ma" finds spelling variants of the Moroccan discourse particle "zehma" and typical syntactic patterns ("[een,het] %je [dat,die]" finds cases where a neuter (diminutive) noun is followed by a relative pronoun that either agrees with the noun ("dat") or not ("die")). Although no linguistic annotation is available, some linguistic phenomena can still be studied by querying for typical ngrams exemplifying the phenomenon (i.e. frequencies for the use of "hun" as subject can be approximated by searching for "hun [willen,hebben,kunnen,..]") and in some cases the corpus provides information not available in smaller corpora (i.e. frequencies for individual neuter nouns occurring with an non-agreeing relative pronooun allow us to study the influence of semantic class over and above that of the variation in non-agreement of individual nouns).

# User types in Dutch Twitter

Hans van Halteren
Radboud University Nijmegen
hvh@let.ru.nl

By crawling internet or social media platforms, we manage to collect massive amounts of language data, such as TwiNL (Tjong Kim Sang and van den Bosch, 2013), but without reliable metadata. This means that it is impossible, or at least much harder, to extract subsets conforming to the design needed for any specific investigation. In several investigations on TwiNL data, we noticed that the results were somehow polluted. There are quite a number of non-Dutch tweets in the collection, as well as tweets from institutional feeds and/or bots rather than human individuals. And even when we manage to include only individuals, we see unbalanced results because the data appears to be greatly biased towards social chitter-chatter between school kids. For this talk, I examine the more than 30 million users who produced tweets crawled for TwiNL in the period Januari 2011 to June 2013. I take various overall measurements, such as the type-token-ratio, and also create top-N lists of several classes of words. On the basis of such user profiles, I first attempt to identify users producing predominantly Dutch tweets. Within that group I then attempt to identify human individuals. Finally, I will try to distinguish subgroups within the individuals, such as e.g. the abovementioned school kids. If such user typing is successful, it will become possible to construct suitable subcollections of TwiNL for research. In the talk, I describe the measurements and classification methods used, and the degree of success in distinguishing various user types.

# Using computational semantics and computer vision to hack the brain - preliminary results and discussion of ongoing work

Alessandro Lopopolo
CLTL - Vrije Universiteit Amsterdam
a.lopopolo@vu.nl

In recent years, distributional semantic models have been proven useful tools for the investigation of how the human brain encodes semantic representations (Mitchell, 2008). In the present work, two families of distributional semantic models have been considered: a traditional corpus-based lexical model (Baroni & Lenci, 2010), and a visual image-based model built out the representation of concrete concepts represented by a large collection of pictures depicting them (Bruni et al., 2014}. These computational models are contrasted with patterns of brain activity elicited by a set of words. The aim of the analysis is to see how well different computational semantic models can explain brain activity in different areas of the brain. In order to study the interaction of computational and brain representation of the semantic space we have used an exploratory method called representational similarity analysis, implemented in a searchlight fashion (Kriegeskorte et al. 2006}. The results confirm neuroscientific theories about the organisation of the 'semantic network' in the brain. The image-based model tends to show higher similarity with the activity in visual areas of the brain, whereas the corpus-based lexical model is more strongly associated with amodal areas of the left hemisphere, areas that are considered processing 'linguistic' information. Giving these promising results, we also discuss some ongoing work on the role of context, both textual and perceptual, in shaping brain representation and computational modelling.

# Using lexicalized parallel treebanks for STSG induction

Vincent Vandeghinste
University of Leuven
vincent@ccl.kuleuven.be

For the SCATE project (Smart Computer-Aided Translation Environment) we are investigating how we can induce an synchronous tree substitution grammar from a parallel aligned lexicalized treebank. In a lexicalized tree, each node in the tree contains information about its head-token, head-lemma and head-category. Combining these features with the syntactic category and dependency relations of the trees provides us with five sources of information, which we can unify in order to estimate the probability of unseen cases, based on the probabilities of its features and combination of features. Additionaly, we are inspired by the Collins parser to estimate the probability of an STSG rule by multiplying the probabilities of generating the heads with the probabilities for generating the other dependents. We will discuss this model and show how it advances over the model used in the PaCo-MT project.

# V-v sequences in Odia - Parsing with a DFA

Kalyanamalini Sahoo
University of Antwerp
Kalyanamalini.Sahoo2@uantwerpen.be

This paper discusses the parsing of multi-verb constructions in Odia, in a Deterministic Finite State Automaton (DFA). It deals with two verb (V-v) sequences, and in the case of passivization, 3 verb sequences. These V-v sequences are formed by combining a main verb and a fully or partially bleached 'light' verb; where the main verb carries the lexical semantic information; the Tense, Aspect, Mode marking occurs on the light verb. The verbs cannot change their position, and are always spelled as one single word. Such double positional slots for the verbs poses constraint for the processing of the string by DFA, as it would be different from a standardly accepted single verb string. Moreover, both the verbs vary in their grammatical and semantic functions, and hence, choose a particular type of verb to co-occur with. This paper proposes morphological parsing of such multi-verb constructions in a DFA. I specify the co-occurrence restrictions of both the verbs in a verbal form and use the DFA to solve the problem of morphological recognition; determining whether an input string of morphemes makes up a legitimate Odia word or not. The EMILLE/CIIL Corpus (ELRA-W0037) is consulted for the data. Out of the 11079 V-v sequences, the proposed morphological parser analyzes 10567 (95.378%) sequences correctly. Such a morphological parser will help us to build a computational lexicon structured as a list of stems and affixes and also can be used in various applications like morphological analyzer, spell-checker, machine translation, information retrieval, etc.

# Visualizing complex linguistic data using GTFL: a case study for Fluid Construction Grammar

Miquel Cornudella
Sony Computer Science Laboratory Paris
cornudella@csl.sony.fr

Grammar engineering can be a daunting challenge. One of the key ingredients for successfully managing the complexity of this task is to properly visualize linguistic data in an informative way. In this presentation, I will show how GTFL (a Graphical Terminal for Lisp [1]) can be used for these purposes and illustrate it through a web interface that was developed for Fluid Construction Grammar (FCG) [2]. The web interface provides users with a readable and intuitive representation of grammars. It shows a detailed representation of complex linguistic data that can be customised by the users to adapt to different needs and aesthetic preferences. It serves as a powerful tool to help both designing and debugging grammars. Fluid Construction Grammar is an interesting case study for developing a visualiza- tion. An FCG grammar consists of constructions which are basically complex feature structures. As these structures can be very elaborate, a comprehensible visualization is needed to improve human understanding. The ideas presented in this talk can directy be applied to other grammar formalisms using complex feature structures.

# Weakly supervised concept tagging: combining a generative and a discriminative approach

Janneke van de Loo
CLiPS, University of Antwerp
janneke.vandeloo@uantwerpen.be

Guy de Pauw
CLiPS, University of Antwerp
guy.depauw@uantwerpen.be

Jort F. Gemmeke
ESAT, KU Leuven
jgemmeke@amadana.nl

Walter Daelemans
CLiPS, University of Antwerp
walter.daelemans@uantwerpen.be

In previous work, we presented FramEngine, a system that learns to map utterances onto semantic frames in a weakly supervised way, as the training data does not specify any alignments between sub-sentential units in the utterances and slots in the semantic frames. Moreover, the semantic frames used for training often contain redundant information that is not referenced to inside the utterance. FramEngine uses hierarchical hidden Markov models (HHMMs) to model the association between the slots in the semantic frame and the words in the utterance. As such, the trained HHMMs can be used to tag utterances with concepts, viz. slot values in the semantic frames, and the resulting slot value sequences can be converted into semantic frames with filled slots. Previous experiments have shown that FramEngine achieves high semantic frame induction performances with small amounts of training data. In our current work, we show that we can further improve its performance by adding a retraining step with a discriminative concept tagger. Concept tagging in the retraining phase is performed in two substeps, in order to incorporate particular generalisations, which have already proven to be effective in FramEngine. In the presented experiments, we use orthographic transcriptions of spoken utterances as input. Improvements are especially made when the utterances contain disfluencies such as interjections or restarts, which makes the combined system particularly useful for speech-based input.

# p(conclusion | Skipping *2*): Cross-domain Bayesian Language Modelling with Skipgrams

Louis Onrust
Radboud University Nijmegen
l.onrust@let.ru.nl

Language models are important modules in many NLP applications such as machine translation and automatic speech recognition. To avoid overestimation and overcome sparsity, many language models have been built around the modified Kneser-Ney algorithm, which was one of the state-of-the-art algorithms for over a decade. More recently, there has been a revived interest encouraged by Teh's 2006 paper on Bayesian language models, and more recently, Mikolov's 2010 paper on recurrent neural network language models, both improving over the modified Kneser-Ney algorithm. In this talk we focus on Bayesian language models, and how we can improve cross-domain language modelling by adding skipgrams as input features. We give a brief introduction to Bayesian non-parametric language models and how they relate to the Kneser-Ney algorithm, and compare our results to other existing language models.

# 2 BIBLIOGRAPHY


14 A. Baron, P. Rayson, and D. Archer (2009). Automatic Standardization of Spelling for Historical Text Mining. In: Proceedings of Digital Humanities 2009, University of Maryland, USA, 22-25 June 2009.

35 Arbib, Michael A. (2012). How the Brain Got Language. The Mirror System Hypothesis. Oxford: OUP.

35 Bergen, Benjamin K. and Chang, Nancy (2005). Embodied Construction Grammar in Simulation-Based Language Understanding. In: J.-O. Östman & M. Fried (eds.), Construction Grammars: Cognitive Grounding and Theoretical Extensions. Amsterdam: John Benjamins.

35 Boas, Hans C. and Sag, Ivan A. (2012, eds.). Sign-Based Construction Grammar. Chicago: University of Chicago Press.

35 Goldberg, Adele E. (2006). Constructions At Work. The Nature of Generalization in Language. Oxford: OUP.

55 Hendrickx Iris, Véronique Hoste, Walter Daelemans.'Semantic and syntactic features for coreference resolution for Dutch.' In: Proceedings of the CICLing 2008 Conference, Haifa, Israel, Berlin: Springer, 2008, p. 351- 361

91 Housen, A., Kuiken, F., & Vedder, I. (2012). Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Language Learning & Language Teaching. Amsterdam: John Benjamins.

53 Koehn and Schroeder. 2007. Experiments in domain adaptation for statistical machine translation.

29 Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In Proceedings of the 35th Annual Meeting of the ACL (pp. 64–71). Stroudsburg: Association for Computational Linguistics.

91 Lu, Xiaofei. "Automatic Analysis of Syntactic Complexity in Second Language Writing." International Journal of Corpus Linguistics 15, no. 4 (2010).

91 Lu, Xiaofei. "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives." The Modern Language Journal 96, no. 2 (June 2012): 190–208.

14 M. Reynaert (2014a) TICCLops: Text-Induced Corpus Clean-up as online processing system. In: Proceedings of COLING 2014 System Demonstrations, Dublin, Ireland, pp. 52–56.

14 M. Reynaert (2014b) On OCR ground truths and OCR post-correction gold standards, tools and formats. In: Proceedings of DATeCH '14. ACM:New York, NY, USA, pp.159-166.




103

53 Meng et al. 2014. Modeling Term Translation for Document-informed Machine Translation.

53 Nakazawa et al. 2014. Overview of the 1st Workshop on Asian Translation.

35 Nathan Schneider and Reut Tsarfaty. 2013. Book review: Design patterns in Fluid Construction Grammar. Computational Linguistics, 39(2):447–453.

72 Peter McClanahan, George Busby, Robbie Haertel, Kristian Heal, Deryle Lonsdale, Kevin Seppi, Eric Ringger (2010) A Probabilistic Morphological Analyzer for Syriac. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 810–820. MIT, Massachusetts, USA, 9-11 October 2010

53 Skadins et al. 2013. Application of Online Terminology Services in Statistical Machine Translation.

35 Steels, Luc (2011, ed.). Design Patterns in Fluid Construction Grammar. Amsterdam: John Benjamins.

14 T. Vobl, A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz (2014) PoCoTo - an open source system for efficient interactive postcorrection of OCRed historical texts. In: Proceedings of DATeCH '14. ACM:New York, NY, USA, pp. 57-61.

64 Taatgen, Niels A., and John R. Anderson. "Why do children learn to say "broke"? A model of learning the past tense without feedback." Cognition 86.2 (2002): 123-155.

55 Tim Van de Cruys. 2010. Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text. PhD thesis. University of Groningen, The Netherlands

53 Tsvetkov and Wintner. 2012. Extraction of MWE from small parallel corpora.

14 VU DNC Corpus: VU University Diachronic News text Corpus https://portal.clarin.inl.nl/vu-dnc/

91 Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity. Hawaii: University of Hawaii.

33 Zuidema, J. & A. Neijt (2012) Verkennend onderzoek naar de wenselijkheid en de haalbaarheid van een verrijking van de Woordenlijst Nederlandse Taal ten behoeve van spellingonderwijs. Rapport in opdracht van de Nederlandse Taalunie http://taalunieversum.org/sites/tuv/files/downloads/rapport%20VWS%2015022013.pdf.

100 [1] M. Loetzsch. GTFL - A graphical terminal for Lisp http://martin-loetzsch.de/gtfl/..

51 [1] M. Ruiter, L. Beijer, C. Cucchiarini, E. Krahmer, T. M. Rietveld, H. Strik, and Van hamme Hugo, "Human language technology and communicative disabilities: requirements and possibilities for the future," Lang. Resour. Eval., vol. 46, no. 1, pp. 143–151, Mar. 2012.

51 [2] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. ICASSP-96, Atlanta, GA, 1996, vol. 1, pp. 373–376.

100 [2] L. Steels. Design patterns in fluid construction grammar, volume 11. John Benjamins Publishing, 2011.

51 [3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, Y. Yamagishi, and K. Oura, “Speech Synthesis Based on Hidden Markov Models,” Proc. IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.

60 [Odijk 2011] Odijk, J. (2011), "User Scenario Search", internal CLARIN-NL document, April 13, 2011 http://www.clarin.nl/system/files/User%20scenario%20Serach%20110413.docx.

60 [Odijk 2014] CLARIN-NL: Search Illustration 1, presentation LOT Summerschool, Nijmegen, 2014 http://www.clarin.nl/sites/default/files/CLARINforLinguistsLOT2014Summerschoo pdf

# 3 | AUTHOR INDEX



107

# INDEX