

NLPCS 2013

Bernadette Sharp and Michael Zock (Eds.)

Natural Language Processing and Cognitive Science

Proceedings of NLPCS 2013
10th International Workshop on Natural Language
Processing and Cognitive Science
Marseille, France - October 2013



Table of Content

Preface	1
Program	3
Invited Talk - resumé	5
Tutorials - resumé	6
Talks - resumé	7
Coherence and Cohesion for the Assessment of Text Readability.	11
<i>Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat and Delphine Bernhard.</i>	
Towards a Cognitive-Linguistic Reconstruction of "Mental Maps" from Ancient Texts. The Example of Dionysios Periegetes.	20
<i>Guenther Goerz, Ekaterina Ilyushechkina and Thiering Martin</i>	
Automatic Sentence Clustering to help Authors to Structure Their Thoughts.	35
<i>Michael Zock and Debela Tesfaye.</i>	
Semantic Types, Lexical Sorts and Classifiers.	49
<i>Bruno Mery and Christian Retore.</i>	
Hidden Structure and Function in the Lexicon.	65
<i>Olivier Picard, Mélanie Lord, Alexandre Blondin-Massé, Odile Marcotte and Stevan Harnad.</i>	
From Stimulus to Associations and Back.	78
<i>Reinhard Rapp.</i>	
Can Humain Association Norm Evaluate Latent Semantic Analysis ?	92
<i>Izabela Gatkowska, Michael Korzycki and Wieslaw Lubaszewski.</i>	
A Cognition-Oriented Approach to Fundamental Frequency Estimation.	105
<i>Ulrike Glavitsch and Klaus Simon.</i>	
On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness?	117
<i>Eva Lasarcyk, Charlotte Wollermann, Bernhard Schröder and Ulrich Schade.</i>	
Supervised Learning Model for Parsing Arabic Language.	129
<i>Nabil Khoufi, Souhir Louati, Chafik Aloulou and Lamia Hadrich Belguith.</i>	
Disambiguation of the Semantics of German Prepositions: a Case Study.	137
<i>Simon Clematide and Manfred Klenner.</i>	
(Fore)seeing Actions in Objects. Acquiring Distinctive Affordances from Language.	151
<i>Irene Russo, Irene De Felice, Francesca Frontini, Fahad Khan and Monica Monachini.</i>	
Extracting Opinion and Factivity from Italian Political Discourse.	162
<i>Rodolfo Delmonte, Daniela Gifu and Rocco Tripodi.</i>	
Use of Language and Author Profiling: Identification of Gender and Age.	177
<i>Francisco Rangel and Paolo Rosso.</i>	

NLPCS 2013
10th International Workshop on
Natural Language Processing and Cognitive Science

Preface

It is our great pleasure to welcome you to the 10th International Natural Language Processing and Cognitive Science workshop, which is part of a series of workshops previously organised in Porto (2004), Miami (2005), Paphos (2006), Funchal (2007), Barcelona (2008), Milan (2009), Madeira (2010), Copenhagen (2011) and Wrocław (2012). The aim of this workshop is to foster interactions among researchers and practitioners in Natural Language Processing (NLP) by taking a Cognitive Science perspective, hence the workshop's name NLPCS. What characterises this kind of approach is the fact that NLP is considered from various viewpoints (linguistics, psychology, neurosciences, artificial intelligence), and that a deliberate effort is made to reconcile or integrate them into a coherent whole. This workshop is an excellent opportunity to encourage cross-fertilization which may possibly lead to the creation of true semiotic extensions, i.e. the development of brain inspired (or brain compatible) cognitive systems.

We believe that this is necessary, as the modelling of the process is simply too complex to be addressed by a single discipline. No matter whether we deal with a natural or artificial system (people or computers) or a combination of both (interactive NLP), systems rely on many types of quite different knowledge sources. Hence, strategies vary considerably depending on the person (novice, expert), on the available knowledge (internal and external), and on the nature of the information processor: human, machines or both (human-machine communication).

The workshop was opened by the keynote speech entitled “What makes language processing difficult or easy?”, talk given by Philippe Blache, director at the CNRS and Director of the Brain and Language Research Institute (<http://www.blri.fr/accueila.html>), Aix-Marseille Université.

The papers covered a wide range of topics namely:

- Automatic sentence clustering,
- Cognitive linguistic approaches to the construction of texts,
- Semantic issues such as word-sense disambiguation, latent semantic analysis, lexical semantics, affordance acquisition,
- Revelation of the hidden structure and functions of words,
- Word associations and co-occurrences,
- Author profiling and opinion extraction,
- Parsing via supervised learning, and
- Frequency estimation in speech.

The papers were followed by a one-day tutorial covering the following two topics: empirical approaches to machine-translation and the role of neurosciences to make us understand language as a process and to help us develop tools to assist this kind of processing. The titles were as follows:

- *Empirical Translation Process Research: a methodology for investigating cognitive processes in translation and machine translation post-editing.* This was led by M. Carl from the Copenhagen Business School, Denmark.
- *NLPCNS: Natural Language Processing from a Cognitive NeuroScience perspective.* This was led by M. Besson and F.-X. Alario, from the Laboratoire of Cognitive Neuroscience at the University of Aix-Marseille, France.

Putting together NLPCS 2013 was a team effort. We would like to thank the authors for providing the content of the programme. We are grateful to the programme committee who worked very hard in reviewing papers and providing feedback for authors. We would like also to thank CIRM for hosting the workshop and, in particular, Lih-Juang Fang for her help with the administrative tasks.

We hope that you will find this programme interesting and thought-provoking and that the symposium will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world.

We look forward to seeing you at NLPCS 2014.

October 2013

Co-chairs of the workshop:

Bernadette Sharp, Staffordshire University, U.K.
 Michael Zock, CNRS, LIF, University Aix-Marseille, France

10th International Workshop on NLPCS Program

Tuesday, October 15th

09:00 - 09:20 - Registration - CIRM - Main Hall

09:20 - 09:30 - Opening - Conference Room

09:30 - 10:30 - Invited Talk

Philippe Blache

[What Makes Language Processing Difficult or Easy?](#)

10:30 - 10:45 - Tea / Coffee Break

10:45 - 12:45 - Cluster I

10:45 - 11:25 - Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat and Delphine Bernhard.

[Coherence and Cohesion for the Assessment of Text Readability.](#)

11:25 - 12:05 - Guenther Goetz, Ekaterina Ilyushechkina and Thiering Martin.

[Towards a Cognitive-Linguistic Reconstruction of "Mental Maps" from Ancient Texts. The Example of Dionysios Periegetes.](#)

12:05 - 12:45 - Michael Zock and Debela Tesfaye

[Automatic Sentence Clustering to Help Authors to Structure Their Thoughts.](#)

12:45 - 14:30 - Lunch Break

14:30 - 18:45 - Clusters II + III

14:30 - 15:10 - Bruno Mery and Christian Retore.

[Semantic Types, Lexical Sorts and Classifiers.](#)

15:10 - 15:50 - Olivier Picard, Mélanie Lord, Alexandre Blondin-Massé, Odile Marcotte and Stevan Harnad

[Hidden Structure and Function in the Lexicon](#)

15:50 - 16:30 - Reinhard Rapp.

[From Stimulus to Associations and Back.](#)

16:30 - 16:45 - Tea / Coffee Break

16:45 - 17:25 - Izabela Gatkowska, Michael Korzycki and Wieslaw Lubaszewski.
[Can Humain Association Norm Evaluate Latent Semantic Analysis ?](#)

17:25 - 18:05 - Ulrike Glavitsch and Klaus Simon.
[A Cognition-Oriented Approach to Fundamental Frequency Estimation.](#)

18:05 - 18:45 - Eva Lasarcyk, Charlotte Wollermann, Bernhard Schröder and Ulrich Schade.
[On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness?](#)

18:45 - End of session

Wednesday, October 16th

09:00 - 11:00 - Cluster IV

- 09:00 - 09:40** - Nabil Khoufi, Souhir Louati, Chafik Aloulou and Lamia Hadrich Belguith.
[Supervised Learning Model for Parsing Arabic Language.](#)
- 09:40 - 10:20** - Simon Clematide and Manfred Klenner
[Disambiguation of the Semantics of German Prepositions: a Case Study.](#)
- 10:20 - 11:00** - Irene Russo, Irene De Felice, Francesca Frontini, Fahad Khan and Monica Monachini.
[\(Fore\)seeing Actions in Objects. Acquiring Distinctive Affordances from Language.](#)

11:00 - 11:15 - Tea / Coffee Break

11:15 - 12:35 - Cluster V

- 11:15 - 11:55** - Rodolfo Delmonte, Daniela Gîfu and Rocco Tripodi.
[Extracting Opinion and Factivity from Italian Political Discourse.](#)
- 11:55 - 12:35** - Francisco Rangel and Paolo Rosso.
[Use of Language and Author Profiling: Identification of Gender and Age.](#)

12:35 - Closure

What makes language processing difficult or easy?

Philippe Blache

LPL, Laboratoire Parole et Langage

CNRS & Université de Provence, 5 Avenue Pasteur
13604 Aix-en-Provence, France

What makes language processing easy or difficult? The answer to this question and our knowledge concerning the non-linearity of human language processing in general are still very incomplete. What we do know though is that some constructions are harder to process than others: object relative clauses, semantic or syntactic incongruities, being examples in case. Several psycholinguistic models have been proposed to explain, at least partially, the complexity of this phenomenon (Gibson98, Hawkins03, Lewis05).

More recently, a computational approach has emerged, proposing to predict difficulty on the basis of probabilistic information (Hale01). However, (a) we do not know yet precisely how to evaluate these parameters automatically (if possible) in a natural setting and (b) to what extend they are only parts of the picture, next to facilitation effects. Indeed, just as some constructions make processing difficult, others seem to facilitate it. Hence, a general account of language processing has to integrate both approaches.

Structural complexities are of great importance for the understanding of the human language processing by and large and its underlying architecture. For example, it is interesting to study to what extent this applies to idioms: several eye-tracking and evoked potential experiments have shown that idioms are processed faster than non-idiomatic expressions. This suggests that they are processed holistically, which contradicts the incrementality hypothesis. Idioms being stored holistically in long-term memory, their access is easy as it is direct, somehow like the access to a lexical unit.

I will substantiate these claims by presenting in this talk an overview of our experimental knowledge concerning this question. Next, I will present an unifying approach concerning complexity, integrating in the same framework penalizing and facilitative phenomena. Doing so allows us not only to relax certain constraints inherent to incremental processing, but also to account for language processing in a natural setting (or context).

- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, 1–8.
- Hawkins, J. A. (2004). Efficiency and Complexity in Grammars. Oxford University Press, USA.

Empirical Translation Process Research:

a methodology for investigating **cognitive processes** in translation and machine translation post-editing

Michael Carl

Dept. of Int. Business Communication, Copenhagen Business School,

Dalgas Have 15 DK-2000 Frederiksberg, Denmark

This tutorial introduces empirical translation process research, as it is conducted at the Centre for Research and Innovation in translation and Translation Technology (CRITT). Since its beginning, CRITT has been developing software to record and investigate **human translation** and post-editing behaviour and elaborated methodologies for translation process research and post-editing, including the study of cognitive processes underlying these tasks. This tutorial introduces the methods, technology, and recent findings of the CRITT group.

NLPCNS : Natural Language Processing from a Cognitive NeuroScience perspective

Mireille Besson (1) and **F.-Xavier Alario** (2)

LNC(1) & LPC(2), CNRS and Aix Marseille University, Marseille France

Cognitive science rests on the assumption that the understanding and modelling of natural language processing is too complex to be addressed by a single discipline. While ever since the beginning *cognitive neuroscience* was considered to be part of the enterprise having a huge potential to make an important contribution, time was not ripe to produce the desired results. Things have changed over the last decade or two. This is what this talk is about.

This talk will be in a dialogue form, and its goal is to provide an overview of the methods developed in cognitive neuroscience to help us understand natural language processing. Concretely speaking, we will try to answer questions like the following: what is involved when processing language, or, what takes place in our brain when we try to understand or produce language ?

We will start the talk by discussing the advantages and disadvantages of brain imaging methods (EEG / ERP, MEG, MRI, fMRI and DTI, TMS, NIRS) used by cognitive neuroscientists. We will then illustrate how these methods are used to study language in action (natural language processing). In the last part, we will discuss the potential contributions of this line of research for interdisciplinary work. If the potential of cognitive neuroscience has been recognized already quite some time ago, its use in NLP has been scarce. Yet time seems to have become ripe for a change.

Resumé talks

Coherence and Cohesion for the Assessment of Text Readability.

Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat and Delphine Bernhard.

Text readability depends on a variety of variables. While lexico-semantic and syntactic factors have been widely used in the literature, more high-level discursive and cognitive properties such as cohesion and coherence have received little attention. This paper assesses the efficiency of 41 measures of text cohesion and text coherence as predictors of the text readability for the French language. We compare results manually obtained on two corpus including texts with different difficulty levels and show that some cohesive features are indeed useful predictors.

Towards a Cognitive-Linguistic Reconstruction of "Mental Maps" from Ancient Texts. The example of Dionysios Periegetes.

Guenther Goerz, Ekaterina Ilyushechkina and Thiering Martin

We present an interdisciplinary approach to implicit knowledge of spatial cognition in common sense geography. Structures such as, e.g., distance, scale, topology, trajectory and frames of reference are believed to be encoded as mental models. Furthermore, we refer to common sense as 'naive' perception and descriptions of space and the use of 'intuitive' arguments in geographical contexts. Our empirical data sets comprise of ancient written texts from different periods and sources. The idea is to combine annotating and parsing techniques based on corpus data analysis as a (semi-) automatic analysis and cognitive parameters applied in cognitive linguistics. These parameters are based on gestalt-psychological principles such as figure-ground asymmetries. We argue that the survey of antique texts provides further insights whether there are basic epistemological expressions of spatial orientation that might be candidates for universals.

Automatic Sentence Clustering to help Authors to Structure Their Thoughts.

Michael Zock and Debela Tesfaye.

To produce written text can be a daunting task, presenting a challenge not only for high school students or second language learners, but actually for most of us, including scientists and PhD students writing in their mother tongue. Text production involves several tasks: *message planning* (idea generation: what to say?), *text structuring* (creation of an outline), *expression* (mapping of content onto linguistic form) and *revision*. We will address here only *text structuring* which is probably the most challenging task implying the grouping, ordering and linking of messages which at the onset of doing so (the moment of providing the conceptual input) lack this kind of information. We are particularly interested in the answer to the following question: on what grounds do writers 'see' connections between message, ideas or thoughts to impose some kind of order, allowing them to group messages into categories? As this is a very complex problem on which we hardly have begun to work, we will present here only preliminary results based on a very simple example, applying mainly to descriptions, one of the many text-types.

Semantic Types, Lexical Sorts and Classifiers.

Bruno Mery and Christian Retore.

We propose a cognitively and linguistically motivated set of sorts for lexical semantics in a compositional setting: the classifiers in languages that do have such pronouns. These sorts are needed to include lexical considerations in a semantical analyser such as Boxer or Grail.

Indeed, all proposed lexical extensions of usual Montague semantics to model restriction of selection, felicitous and infelicitous copredication require a rich and refined type system whose base types are the lexical sorts, the basis of the many-sorted logic in which semantical representations of sentences are stated. However, none of those approaches define precisely the actual base types or sorts to be used in the lexicon. In this article, we shall discuss some of the options commonly adopted by researchers in formal lexical semantics, and defend the view that classifiers in the languages which have such pronouns are an appealing solution, both linguistically and cognitively motivated.

[Hidden Structure and Function in the Lexicon.](#)

Olivier Picard, Mélanie Lord, Alexandre Blondin-Massé, Odile Marcotte and Stevan Harnad.

How many words are needed to define all the words in a dictionary? Graph-theoretic analysis reveals that about 10% of a dictionary is a unique Kernel of words that define one another and all the rest, but this is not the smallest such subset. The Kernel consists of one huge strongly connected component (SCC), about half its size, the Core, surrounded by many small SCCs, the Satellites. Core words can define one another but not the rest of the dictionary. The Kernel also contains many overlapping Minimal Grounding Sets (MGSs), each about the same size as the Core, each part-Core, part-Satellite. MGS words can define all the rest of the dictionary. They are learned earlier, more concrete and more frequent than the rest of the dictionary. Satellite words, not correlated with age or frequency, are less concrete (more abstract) words that are also needed for full lexical power.

[From Stimulus to Associations and Back.](#)

Reinhard Rapp.

Free word associations are the words human subjects spontaneously come up with upon presentation of a stimulus word. In experiments comprising thousands of test persons large collections of associative responses have been compiled. In previous publications it could be shown that these human associations can be resembled by statistically analyzing the co-occurrences of words in large text corpora. In the current paper for the first time we consider the reverse question, namely whether the stimulus can be predicted from the responses. By presenting an algorithm which produces surprisingly good results our answer is clearly affirmative.

[Can Human Association Norm Evaluate Latent Semantic Analysis ?](#)

Izabela Gatkowska, Michael Korzycki and Wieslaw Lubaszewski.

This paper presents the comparison of word association norm created by a psycholinguistic experiment to association lists generated by algorithms operating on text corpora. We compare lists generated by Church and Hanks algorithm and lists generated by LSA algorithm. An argument is presented on how those automatically generated lists reflect real semantic relations.

[A Cognition-Oriented Approach to Fundamental Frequency Estimation.](#)

Ulrike Glavitsch and Klaus Simon.

This paper presents an efficient, two-phase fundamental frequency detection algorithm in the time-domain. In accordance with the human cognition process it computes base fundamental frequency estimates first that are verified and corrected in a second step. The verification proceeds from high-energy stable segments where reliable estimates are expected to lower-energy regions. Irregular cases are handled by computing a series of fundamental frequency variants that are evaluated for highest plausibility adopting the hypothesis testing principle of

human thinking. The algorithm was evaluated on a clean speech database as a proof of concept where it shows significantly lower error rates than a comparable reference method.

[On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness?](#)

Eva Lasarcyk, Charlotte Wollermann, Bernhard Schröder and Ulrich Schade.

In this paper we present work on the modelling of uncertainty by means of prosodic cues in an articulatory speech synthesizer. Our stimuli are embedded into short dialogues in question-answering situations in a human-machine scenario. The answers of the robot vary with respect to the intended level of (un)certainty, the independent variables are intonation (rising vs. falling) and filler (absent vs. present). We perform a perception study in order to test the relative impact of the prosodic cues of uncertainty on the perception of uncertainty and also of naturalness. Our data indicate that the cues of uncertainty are additive. If both prosodic cues of uncertainty are present, the perceived level of uncertainty is higher as opposed to the deactivation of a single cue. Regarding the relative contribution of intonation vs. filler our results do not show a significant difference between judgments. Moreover, the correlation between the judgment of uncertainty and of naturalness is not significant.

[Supervised Learning Model for Parsing Arabic Language.](#)

Nabil Khoufi, Souhir Louati, Chafik Aloulou and Lamia Hadrich Belguith.

Parsing the Arabic language is a difficult task given the specificities of this language and given the scarcity of digital resources (grammars and annotated corpora). In this paper, we suggest a method for Arabic parsing based on supervised machine learning. We used the SVMs algorithm to select the most probable syntactic labels of the sentence. Furthermore, we evaluated our parser following the cross validation method by using the Penn Arabic Treebank. The obtained results are very encouraging.

[Disambiguation of the Semantics of German Prepositions: a Case Study.](#)

Simon Clematide and Manfred Klenner.

In this paper, we describe our experiments in preposition disambiguation based on a - compared to a previous study - revised annotation scheme and new features derived from a matrix factorization approach as used in the field of distributional semantics. We report on the annotation and Maximum Entropy modelling of the word senses of two German prepositions, "mit" (with) and "auf" (on). 500 occurrences of each preposition were sampled from a treebank and annotated with syntacto-semantic classes by three annotators. Our coarse-grained classification scheme is geared towards the needs of information extraction, it relies on linguistic tests and it strives to separate semantically regular and transparent meanings from idiosyncratic meanings (i.e. of collocational constructions). We discuss our annotation scheme and the achieved inter-annotator agreement, we present descriptive statistical material e.g. on class distributions, we describe the impact of the various features on syntacto-semantic and semantic classification and focus on the contribution of semantic classes stemming from distributional semantics.

[\(Fore\)seeing Actions in Objects. Acquiring Distinctive Affordances from Language.](#)

Irene Russo, Irene De Felice, Francesca Frontini, Fahad Khan and Monica Monachini.

In this paper we investigate if conceptual information concerning objects' affordances as possibilities for actions anchored to an object can be at least partially acquired through language. Considering verb-noun pairs as the linguistic realizations of relations between actions performed by an agent and objects we collect this information from the ImagAct dataset, a linguistic resource obtained from manual annotation of basic action verbs, and

from a web corpus(itTenTen). The notion of affordance verb as the most distinctive verb in ImagAct enables a comparison with distributional data that reveal how lemmas ranking based on a semantic association measure that mirror that of affordances as the most distinctive actions an object can be involved in.

[Extracting Opinion and Factivity from Italian Political Discourse.](#)

Rodolfo Delmonte, Daniela Gifu and Rocco Tripodi.

The success of a newspaper article for the public opinion can be measured by the degree in which the journalist is able to report and modify (if needed) attitudes, opinions, feelings and political beliefs. We present a symbolic system for Italian, derived from GETARUNS, which integrates a range of natural language processing tools (also available in the public domain) with the intent to characterise the print press discourse from a semantic and pragmatic point of view. This has been done on some 500K words of text, extracted from three Italian newspapers in order to highlight their stance on the deep political crisis situation which brought to the change of government that took place at the end of 2011. We tried two different approaches: a lexicon-based approach for semantic polarity using off-the-shelf dictionaries with the addition of manually supervised domain related concepts. Another one is a feature-based semantic and pragmatic approach, which computes propositional level analysis on the basis of the verbal complex and other semantic markers to process factuality and subjectivity. Results are quite revealing and confirm the otherwise common knowledge about the political stance of each newspaper on such topic.

[Use of Language and Author Profiling: Identification of Gender and Age.](#)

Francisco Rangel and Paolo Rosso.

"In the beginning was the Word, and the Word was with God, and the Word was God". Thus, John 1:11 begins his contribution to the Holy Bible, the importance of the word lies in the essence of human beings. The discursive style reflects the profile of the author, who decides, often unconsciously, about how to choose and combine words. This provides valuable information about the personality of the author. In this paper we present our approach to identify age and gender of an author based on his use of language. We propose a representation based on stylistic features and obtain encouraging results with a SVM-based approach on the PAN-AP-132 dataset.

Coherence and Cohesion for the Assessment of Text Readability

Amalia Todirascu⁽¹⁾, Thomas François⁽²⁾, Núria Gala⁽³⁾,
Cédric Fairon⁽²⁾, Anne-Laure Ligozat⁽⁴⁾, Delphine Bernhard⁽¹⁾

⁽¹⁾ FDT, LiLPA, Université de Strasbourg

⁽²⁾ CENTAL, UCLouvain, Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique

⁽³⁾ LIF-CNRS, AMU, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France

⁽⁴⁾ LIMSI-CNRS, Orsay, France

E-mail: todiras@unistra.fr, thomas.francois@uclouvain.be, nuria.gala@univ-amu.fr, cedrick.fairon@uclouvain.be,
annlor@limsi.fr, dbernhard@unistra.fr

Abstract. Text readability depends on a variety of variables. While lexico-semantic and syntactic factors have been widely used in the literature, more high-level discursive and cognitive properties such as cohesion and coherence have received little attention. This paper assesses the efficiency of 41 measures of text cohesion and text coherence as predictors of text readability. We compare results manually obtained on two corpora including texts with different difficulty levels and show that some cohesive features are indeed useful predictors.

1 Introduction

Although reading is considered as a crucial skill in education, reaching a sufficient level is a complex challenge for a significant part of the population. A recent publication from the Council of the European Union reports that “on average in the EU-27 in 2009, 19.6 % of [15-year-old] students were low achievers in reading” (De Coster et al., 2011: 22). One way to sustain the growth of one’s reading skills is to offer him/her opportunities for practice, whether guided or independent. Various experiments indicate that regular practice improves reading skills (Mastropieri et al., 1999). For this practice to be profitable, it is also necessary that the texts suit the level of students (O’Connor et al., 2002), which is not always the case.

To assist teachers or readers themselves to more easily find adequate texts, tools have been developed since the 1920’s in the field of readability. They are called readability formulas and aim to match readers of various reading abilities with texts that are within their reach, using various textual characteristics for prediction.

Classic formulas such as Flesch’s (1948) first focused on a few number of lexico-syntactic characteristics (e.g. the average number of words per sentence or the average number of syllables per word). In the 1980’s, the structuro-cognitivist approach of readability stressed the importance of higher textual dimensions such as the inference load (Kintsch, 1979; Kemper, 1983), the conceptual density (Kintsch and Vipond, 1979), or organisational aspects (Meyer, 1982). However, these new dimensions were hardly investigated at that time due to the complexity of the linguistics models involved. Even since, only a few studies -that will be covered in more details in Section 2- focused on those high-level textual dimensions.

Among those high-level dimensions, the level of coherence of the texts is an important one and will be the focus of this paper. It has been shown that a higher level of coherence between a pair of related sentences decreases their reading time and improves their recall (Kintsch et al., 1975). Myers et al. (1987) focused on causal relations and compared the reading speed and the recall of four similar pairs of sentences (expressing the same cause and consequence), ranging from an incoherent version to a very coherent one. They obtained surprising results: while the reading time decreases as the coherence level increases, the recall follows a quadratic function in the shape of an inverted U. In other words, moderately connected sentences are the best remembered ones. Such sentences generally require the reader to make an inference to explicit their relationship barely sketched in the text. This inference generation process produces a higher reading time, but also a richer connection network between the representations of both sentences in memory, leading to a better recall. Mason and Just (2004) used functional magnetic resonance imaging (fMRI) to test this hypothesis with subjects reading the sentences of Myers et al. (1987). They observed activation patterns consistent with Myers et al. (1987)’s findings.

These studies confirm the idea that the more coherent a sequence of sentences, the better these are understood. From these findings, it appears that readability models should benefit from taking into account high-level textual dimensions such as coherence or cohesion. However, as detailed in Section 2.2, current explorations of the issue have failed to achieve a consensus on their efficiency, as they are based on automatic parameterization procedures, prone to errors. In this paper, we propose to investigate whether various measures of text cohesion and coherence are useful to assess the readability of texts, when their parameterization is manually performed. Section 2 further discusses the concepts of coherence and cohesion, and summarizes previous approaches of those dimensions in the readability literature. Section 3 presents the methodology applied in the paper to assess the usefulness of several measures of coherence and cohesion for the prediction of text readability. We also describe the tools and the corpora used in our tests. Section 4 reports a preliminary experiment exploring how features based on cohesion device such as reference chains vary between a normal and a simplified version of the same texts. Based on these results, Section 5 investigates a larger set of variables measuring text coherence and text cohesion, assessing their efficiency to predict French as a foreign language (FFL) text difficulty.

2 Coherence and Cohesion in Readability

2.1 Coherence and Cohesion

Coherence and cohesion are two important properties of texts. Text coherence is considered as a “semantic property of discourses, based on the interpretation of each individual sentence relative to the interpretation of other sentences” (Van Dijk, 1977: 93). A text is realised as a sequence of related utterances. Some theories describe coherence relations by the existence of explicit linguistic markers reinforcing cohesion (Charolles, 1997; Hobbs, 1979). However, cohesive markers are not mandatory elements to obtain coherent texts, although they contribute to the overall text interpretation (Charolles, 1997).

Halliday and Hasan (1976) identified several cohesive devices helpful for the semantic interpretation of the whole text: coreference relations (various expressions referring to the same entity), discourse connectives, lexical relations such as synonymy, hypernymy, hyponymy, meronymy, and thematic progressions. Among these cohesive devices, coreference relations are expressed via anaphoric chains (Kleiber, 1994) or reference chains (Schneedecker, 1997). Anaphoric chains consist of two elements: the anaphor (an expression semantically related to a discourse entity already introduced in the text) and its antecedent (the referred or related entity). The anaphor and its antecedent might be related by various semantic relations such as referential identity or meronymy. Reference chains contain at least three referring expressions, related to the same entity (Schneedecker, 1997). The following example (fig.1) contains two reference chains (un lion étranger ‘a foreign lion’/s’/l’intrus ‘the intruder’; le chef de la tribu ‘the chief of the tribe’/il ‘it’/le dominant ‘the dominant male’), but one anaphoric chain (un combat ‘a fight’/s’).

Fig.1. An example of reference and anaphoric chains

Lorsqu'[un lion étranger]_1 au groupe [s']_1 approche, [un combat]_2 [s']_2 engage (parfois jusqu'à la mort) entre [le chef de la tribu]_3 et [l'intrus]_1. S'il]_3 gagne, [le dominant]_3 reste dans le groupe.

'When a foreign lion approaches the group, a mortal fight involves the chief of the tribe and the intruder. If he wins, the dominant male stays within the group'.

Referring expressions introducing new entities are proper names, indefinite noun phrases, or definite noun phrases. Anaphors referring to known entities are mainly represented by personal pronouns, reflexive pronouns, possessive determiners, demonstrative determiners.

The use of anaphoric or reference chains reinforces the presence of the same entity along the text (Hobbs, 1979). More recent studies such as the Centering theory (Grosz, *et al.*, 1995) claim that some entities used in an utterance are more important than others (centre). This theory proves that local coherence is influenced by the centering properties of the utterance and by the selection of various referring expressions. Referring expressions should verify complex morpho-syntactic (gender and number agreement) and syntactic constraints (syntactic parallelism) to remain the centre of the discourse unit. Along

with lexical repetition, such chains contribute to preserve the main topic of the paragraph or of the document.

2.2 The Use of Coherence and Cohesion Measures for Readability

As mentioned above, the level of coherence and cohesion of texts impacts the understanding of readers. However, these aspects were initially not considered in classic readability models (Flesch, 1948), which were limited to lexical and syntactic characteristics. Bormuth (1969) is probably the first to explore the issue. For him, resolving anaphoric relations correctly is a prerequisite to a good understanding of a text. Therefore, he defined 10 classes of anaphora and computed their proportion, as well as the density of anaphora in the text and the mean distance between each anaphora and its antecedent. These two latter features appeared to be the best predictors of text readability among the 12, with respectively a correlation $r = 0.532$ for density and $r = 0.392$ for the mean distance. Later, Kintsch (1979) analysed the impact of inferences on understanding and found out that the mean number of inferences required in a text was not well correlated with text difficulty.

Another approach of coherence in readability is based on the latent semantic analysis (LSA) developed by Landauer et al. (1998). This method projects sentences in a semantic space in which each dimension roughly corresponds to a semantic field. Therefore, it better allows assessing the semantic similarity between sentences, since it can capture lexical repetitions, even through synonyms or hyponyms. However, this method is not sensitive to cohesive clues such as ellipsis, pronominal anaphora, substitution, causal conjunction, etc. The application of this technique to readability was first investigated by Folz et al. (1998), who computed the average similarity between each pair of sentences in a text as a proxy of the text overall coherence. This variable was also included in Coh-Metrix (Graesser et al., 2004), along with variations such as word overlap, noun overlap, stem overlap, and argument overlap. However, the efficiency of this variable was not assessed before Pitler and Nenkova (2008), who measured its association with text difficulty and obtained a non-significant $r=-0.1$. Later, McNamara et al. (2010) reached a similar conclusion, showing that an LSA-based variable has not much predictive power. François and Fairon (2012) obtained a higher correlation for French ($r = 0.63$), but it was due to some specificities of their corpus. They used FFL (French as a Foreign Language) texts from textbooks, including some texts from beginner's textbooks that were merely a list of disconnected sentences. Therefore, the LSA-based feature tended to consider disconnected texts as easy ones, increasing the strength of the correlation and inverting its direction.

An alternative approach to LSA was suggested by Barzilay and Lapata (2008), who view a text as a matrix of the discourse entities¹ present in each sentence. The cohesive level of a text is then computed based on the transitions between those entities. Pitler and Nenkova (2008) implemented this model through 17 readability variables, but none was significantly correlated with difficulty. Feng et al. (2009) also replicated this technique, without getting more efficient features.

Finally, Pitler and Nenkova (2008) drew from statistical language models to propose a cohesion model in which texts are viewed as a bag of discourse relations (temporal, comparison, etc.). These relations are either explicit (when marked) or implicit. The authors computed the likelihood of a text based on its discourse relations, having trained their model on the Penn Discourse Treebank. They obtained interesting correlations for this variable ($r = 0.48$), which is their best feature.

To conclude, we see that only a few studies focused on using coherence and cohesion measures as predictive variables for readability purposes and mostly for English. It also appears that most variables experimented in the literature were not found significantly correlated with text difficulty. Our study aims to further investigate this issue, focusing on French and taking advantage of (a) several linguistic studies about specific cohesion devices such as reference chains (Schnedecker, 2005) and (b) the availability of RefGen (Longo and Todirascu, 2010), a tool that can help us to capture cohesion and coherence information for French texts.

¹ They define a “discourse entity” as nominal phrases being part of a co-reference relation and having a function (subject, object, etc.).

3 Methodology to Assess whether Coherence and Cohesion Correlate with Readability

Our goal is to investigate the use of several cohesive and coherence properties to evaluate the difficulty of French texts. To this aim, we built two annotated corpora to be used in our experiments. Then, drawing on the literature reported in Section 2, we defined 41 variables aiming at measuring text coherence and cohesion (see Section 4). Although we intended to annotate all of them manually, some of them were eventually computed with RefGen (a tool that we introduce in Section 3.2), when the error rate of their annotation process was deemed low enough. Finally, the efficiency of these variables as predictors of text difficulty was assessed on the corpora (see Sections 4 and 5).

3.1. The Corpora

Two corpora were collected for our experiments, both being annotated in terms of text difficulty. The first one is a corpus of comparable texts from Wikipedia and Vikidia² (a simplified encyclopaedia targeted at children between 8 and 13 years old). We collected 13 informative texts from Wikipedia, describing animals or geographic areas (7,597 tokens) and selected texts on the same subject from Vikidia (5,308 tokens). This corpus was used as a way of detecting interesting features for the rest of the analysis and to gather significant differences between simplified and original texts. To analyse significant features for readability, we manually annotated the corpus' reference chains.

The second is a subset of the corpus of FFL texts gathered by François (2009). This corpus consists of 2,160 texts, selected from 28 FFL textbooks, as long as they are related to a reading comprehension task. All textbooks considered comply with the Common European Framework of Reference for Languages (CEFR), a standard scale for foreign language education in Europe that uses 6 levels (A1 to C2). Therefore, each text was assigned the level of the textbook it came from. In this study, we only used texts from levels A2 to C1 and selected only informative texts to control for the genre of the texts across both experiments. A1 texts were rejected because several of them were just a collection of unconnected sentences. C2 texts were not considered either because there were not enough informative texts for this level in François (2009)'s corpus.

3.2 Annotation of Discourse Entities and of Reference Chains

The computation of our variables for both corpora would require a large amount of manual work, which led us to consider the automation of some tasks (e.g. POS-tagging or detection of entities), provided that their error rate remains low. Few tools are available for coreference resolution in French (Victorri, 2005; Popescu-Belis, *et al.*, 1999) but most of them focus on specific anaphora type (Lassalle and Denis, 2011) or specific domains or tasks (human-machine dialogue systems (Salmon-Alt, 2001)). RefGen is a rule-based system for French which performs the automatic annotation of reference chains (Longo and Todirascu, 2010b), but also entity detection and-POS tagging. RefGen tags and lemmatizes the texts using TTL (Ion, 2007) and it annotates potential referring expressions such as: complex noun phrases (simple NP modified by several PP or relative clauses), named entities (persons and organisations), definite or indefinite noun phrases. In addition, the tool applies several heuristics to label syntactic functions (subject, object, and others). After deleting impersonal occurrences of the 3rd person singular pronoun ('il'), the tool identifies a set of referring expressions as possible starters of a reference chain. Then, RefGen computes a set of antecedent and anaphor pairs by checking several morpho-syntactic and semantic features. Finally, the system groups the candidate pairs into reference chains.

Longo and Todirascu (2010) evaluated RefGen by using a corpus of 7,230 tokens and obtained good results for the entity annotation module (for the module identifying complex noun phrases: recall = 0.87 and precision = 0.91, for the named entity recognition: recall = 0.85 and precision = 0.91) and promising results for the reference chain identification module (recall = 0.58 and precision = 0.70). Reference chain identification is known to be a difficult task, which explains the lower results obtained for this second task. As a consequence, we decided to use this tool to identify discourse entities, but we manually annotated the relations between the referring expressions as well as their syntactic functions.

² This corpus was build and annotate by Ratiba Khobzi, University of Strasbourg.

4 Reference Chains in Wikipedia and Vikidia

As a first investigation of the usefulness of coherence and cohesion variables for text readability prediction, we studied the behaviour of reference chains in a corpus of original texts and their simpler version. It should be mentioned that reference chains have a specific behaviour according to text types or genres. Schnedecker (2005) and Schnedecker and Longo (2012) identify specific properties of reference chains in newspapers genres, such as portrays and news. These studies investigated properties such as the length (the number of referring expressions composing the reference chain), the distance (the number of sentences separating the expressions composing the same chains), the types of referring expressions, and the type of the first element starting a chain. The same properties have been studied in several text genres: law texts, editorials, novels, public reports (Longo et Todirascu, 2013). The study shows significant variations in these properties: longer chains characterize novels, newspaper articles contain medium-sized chains, while law texts contain very short ones. The types of referring expressions composing reference chains also differ from one genre to another: news contain more proper names and personal pronouns, while law texts and public reports contain more indefinite and definite noun phrases. To control as much as possible for this variation across genres, we restricted our analysis to one genre: informative texts.

The properties highlighted by the above studies were manually annotated in our first corpus (Vikidia and Wikipedia). In addition, we compared the number of reference chains, the syntactic functions of the referring expressions composing the chains and the relation between the reference chains and the text topic.

We noticed that the number of reference chains was slightly more important in simple texts (49) than in the original (44). For most of the texts, the average length of the reference chains found in simple texts is shorter than the length of the chains in the original texts. To give an example, 'Le lion' is the main referent in both of the following excerpts; the Wikipedia text contains four expressions referring to it while the Vikidia one has only two (pronouns) (fig.2):

Fig.2. An example of annotated reference chains in Wikipedia et Vikidia texts.

[Le lion]_1 (*Panthera leo*) est un mammifère carnivore de la famille des félidés du genre *Panthera* (félins). [Il]_1 est surnommé " [le roi des animaux]_1 " car [sa]_1 crinière [lui]_1 donne un aspect semblable au Soleil, qui apparaît comme " le roi des astres ". (Wikipedia)

'The lion is a carnivore mammal, member of the family Felidae, in the genus *Panthera* (felins). It is named «king of animals» due to its mane, which gives it the aspect similar to the Sun, which is « the king of asters »'

[*Le lion*]_1 est un mammifère carnivore ressemblant au chat. [Il]_1 fait partie, comme lui, des félins. [*Son*]_1 nom scientifique est *Panthera leo*. (Vikidia)

'The lion is a carnivore mammal similar to a cat. It is member of the felins. Its scientific name is *Panthera leo*'

The distribution of the referential expressions types shows that while the relative frequencies of indefinite (0.2 for Wikipedia and 0.35 for Vikidia) or definite noun phrases (2.59 vs 2.83) are similar in both corpora, several categories are more frequent in simple texts than in the original: proper names (0.07 for Wikipedia; 0.26 for Vikidia), personal pronouns (2.23 for Wikipedia; 3.69 for Vikidia) and demonstrative pronouns (0.04 for Wikipedia; 0.2 for Vikidia).

It should also be noticed that the first element opening a reference chain is more likely to be a definite noun phrase or a NP without determiner for Wikipedia texts, while we observed a preference for indefinite noun phrases in simple texts. In both cases, however, the entity referred to within the longest reference chain is generally the global topic of the document.

Finally, we studied the syntactic function of the referring expressions contained in the chains. We investigated the subject, object and other syntactic functions of the mentions contained in chains. We counted all the transitions (subject-object, subject-subject; subject-other function etc.) between two consecutive sentences containing mentions of the same entity (e.g. part of the same reference chain). The most interesting cases are those with the same syntactic function kept in two consecutive sentences. We observed that this happens more frequently in complex texts than in simple ones. The number of subject pronouns is also more important in simple texts than in Wikipedia texts. In other words, we noticed several variations between the behaviour of reference chains between simple texts and their Wikipedia

counterparts. To confirm these trends, we then performed a more quantitative investigation, described in the next section.

5 Cohesion and Coherence for the Readability of FFL Texts

5.1 Variables of Text Coherence and Cohesion

At the end of our preliminary study on Wikipedia and Vikidia texts, several characteristics of text coherence and cohesion appeared to be valuable for readability prediction. Therefore, based on the literature in readability and the work of Schnedecker (1997, 2005), we defined 41 variables, divided up within five classes as follows:

1. *P.O.S. tag-based variables*: Pronouns and articles are crucial elements of coherence and cohesion. We computed 9 variables based on these part-of-speeches, namely (1) the ratio between pronouns and nouns; the average proportion of pronouns per sentence (2) and per word (3); the average proportion of personal pronouns per sentence (4) and per word (5); the average proportion of possessive pronouns per sentence (6) and per word (7); and the average proportion of definite articles per sentence (8) and per word (9). We also computed the ratio of proper names per word (10).
2. *Lexical coherence measures*: We also replicated several methods based on lexical cohesion, namely (11) the average similarity – measured with cosinus – between adjacent sentences projected in a LSA space, (12) the word overlap (number of common words in two consecutive sentences), (13) the lemma overlap, and the noun and pronouns overlap, based either on lemmas (14), or inflected forms (15). More precisely, every text from the corpus was transformed in a list of bag-of-words vectors (one per sentence), before these vectors were weighted. In the case of the various “overlap” variables, *tf-idf* (term frequency-inverse document frequency) was used for the weighting, while we applied a singular value decomposition (SVD) for LSA³.
3. *Entity coherence*: consecutive sentences can share similar arguments (the subject of the sentence n is also the subject of the sentence $n+1$, the object of the sentence n becomes the subject of the sentence $n+1$, etc.). We followed Pitler and Nenkova (2008) by counting the relative frequency of the possible transitions between the four syntactic functions played by the entity in sentence $n+1$: subject (S), object (O), other complements (C), and (N) when the entity is absent (variables 16 to 28).
4. *Entity density*: we computed the average proportion of entities (simple and complex noun phrases, pronouns, etc.) per document (29), the average number of entities per sentence (30), the average proportion of unique entities per document (31), and the average number of words per entity (32). These features were obtained with the automatic annotation provided by RefGen.
5. The last class gathers features corresponding to various properties of the reference chains: the proportion of the various types of expressions included in a reference chain : indefinite NP (33), definite NP (34), personal pronouns (35), possessive determiners (36), demonstrative determiners (37), demonstrative pronouns (38), reflexive pronouns (39), or proper nouns (40); the average length of reference chains (41).

5.2 Analysis of the Variable Efficiency

We saw that findings in the literature about the efficiency of coherence and cohesion-based variables for readability are not consistent: some of the studies report non-significant correlations, while others show significant correlations. An explanation for this variation could be the fact that most of those studies rely on an automatic approach of coherence and cohesion, which are notoriously difficult to automatize.

To better control for this aspect, we opted, in this study, for a manual approach of all variables whose automatic annotation would have been impaired by a significant error rate. These experiments were performed on the FFL corpus, that includes texts with a larger spectrum of difficulty. However, since manual annotation requires a much larger amount of resources, we restricted the experiment to 5 texts per

³ To compute the *tf-idf* and the LSA, we used a large amount of texts from the François (2009)'s corpus that were not used for this study. For the LSA, we compared various sizes for the reduced space with a cross-validation procedure that led us to retain a small 15-dimensional space.

level, for a total of 20 texts. We manually annotated the reference chains and their syntactic functions, and then computed all variables described in Section 5.1. Their efficiency as readability predictors was then assessed through Spearman correlations⁴ between each variable and the levels of the texts. Table 1 reports the most significant correlations.

Table 1. The most significant correlations obtained from the manually annotated corpus. The numbers preceding the variables refer to numbers used in Section 5.2

Variable	Corr. and p-value	Variable	Corr. and p-value
35. PRON	-0.59 (p = 0.005)	3. Pers. Pro./S	-0.41 (p = 0.07)
33. Indef NP	-0.50 (p= 0.02)	10. Names/W	-0.4 (p = 0.08)
18. S → O	0.46 (p = 0.04)	9. # def. art./W	0.38 (p = 0.1)
22. O → O	-0.44 (p= 0.048)	17. S → S	-0.36 (p = 0.12)

Interestingly, two variables based on reference chains are significant: the proportion of transitions of the type subject (S) to object (O) between sentences, as well as the proportion of object (O) to object (O). S-O transitions seem to appear more frequently in harder texts, while the O-O (and also S-S) are typical of easier texts⁵. This finding is interesting, since neither Pitler and Nenkova (2008) nor Feng and al. (2009) were able to show the efficiency of the class of variables for readability, using an automatic approach.

Considering the type of referring expressions used in the chains also seems promising. Our two best features are indeed the proportion of personal pronouns and indefinite NP in the chains. Both types of phrases tend to be more present in easier texts. As regards the average length of the chains, it was surprising to notice that long chains are represented similarly in simple texts and in complex ones.

6 Conclusion and Future Work

The experiments in this paper demonstrated that some variables of text coherence and text cohesion are interesting predictors of text readability. We showed that variables based on syntactic transitions present a different profile in simple and complex texts, with more transitions keeping the same function from one sentence to the next one in simpler texts. This is already an interesting finding, since previous approaches of the issue, based on automatic modelling, obtained non-significant correlations. Furthermore, based on the work of Schneedecker (2005) and Schneedecker and Longo (2012), we suggested new features for readability, like the proportion of the type of referring expressions in the chains. Our most interesting finding is that among those features, two of them (PRON, Indef NP) appeared to be good variables, actually our best ones. Therefore, it is useful not only to consider the function of the referring expressions, but also their type. Simpler texts from our corpus indeed tend to use more pronouns and indefinite NPs.

Our manual approach confirmed the interest to consider textual dimensions, such as coherence and cohesion, to assess the readability of informative texts. Several of our variables indeed were able to discriminate between L1 texts (Wikipedia and Vikidia) and FFL texts of various levels. Since, previous work, based on an automatic analysis, were more mitigated on this issue, especially regarding variables based on the syntactic transitions, our findings could be interpreted in two ways: (1) either the significant correlations we observed are due to some specificities of our corpora (genre of the texts, small amount of observations, etc.), or (2) the fact that previous work had trouble to demonstrate the efficiency of coherence and cohesion variables for readability is mostly due to errors in the annotation procedure, performed automatically.

⁴

Spearman correlation formula is described among others in Howell (2008). We did not use the Pearson correlation here, since readability variables often do not have a linear relationship with difficulty.

⁵ This second feature is also rarely observed and it is not obvious that its efficiency would scale to a larger set of data.

To decide between these two conclusions, our analysis should be replicated on a larger corpus, on one side, but should also be performed via an automatic annotation procedure. This would allow to check whether our best variables remain efficient once they are extracted via an automatic system such as RefGen. In further experiments, we plan to investigate if the use of automatic annotations of reference chains, and the inherent annotation errors would impact the efficiency of our coherence and cohesion variables. A last step to our investigation would be to test whether coherence and cohesion dimensions really bring new information to a readability model, as regards to information already contained in lexico-syntactic features.

Acknowledgements

This project is partly financed by the Programme Hubert Curien (PHC) Tournesol 2013 (France and Fédération Wallonie-Bruxelles). We would like to also acknowledge the help of Bernadette Dehottay during the collection of the FFL corpus and Ratiba Khobzi for the annotation of the Vikidia and Wikipedia corpora. RefGen has been developed during Laurence Longo's Ph.D.thesis ("Vers de moteurs de recherche sémantiques"), University of Strasbourg.

References

- Barzilay, R., Lapata, M. (2008) Modeling Local Coherence: An Entity-based Approach, *Computational Linguistics*, 34(1):1-34.
- Bormuth, J. (1969) *Development of Readability Analysis*. Rapport technique, Projet n°7 – 0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.
- Charolles, M. (1995) Cohesion, coherence et pertinence de discours, *Travaux de Linguistique*, 29:125-151.
- De Coster, I., Baidak, N., Motiejunaite, A., and Nooran, S. (2011) Teaching Reading in Europe: Contexts, Policies and Practices. *Education, Audiovisual and Culture Executive Agency, European Commission*.
- Feng, L., Elhadad, N. et Huenerfauth, M. (2009) Cognitively motivated features for readability assessment. *Proceedings of EACL 2009*, pp. 229-237.
- Flesch, R. (1948) A new readability yardstick. *Journal of Applied Psychology*, 32(3):221-233.
- Foltz, P., Kintsch, W. and Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2):285-307.
- François, T. (2009) Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. *Proceedings of EACL 2009: Student Research Workshop*, pp. 19-27.
- François, T. and Fairon, C. (2012) An "AI readability" formula for French as a foreign language. *Proceedings of EMNLP 2012*, Jeju, pp. 466-477.
- Graesser, A., McNamara, D., Louwerse, M. and Cai, Z. (2004) Coh-Metrix : Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, 36(2):193-202.
- Grosz, B., Joshi, A., Weinstein, S. (1995) Centering: A Framework for Modeling the Local Coherence of Discourse.
- Halliday, M.A.K. and Hasan, R. (1976) *Cohesion in English*. London: Longman.
- Hobbs, J. (1979) Coherence and Coreference. *Cognitive Science*, 3(1):67-90.
- Howell, D. (2008). *Méthodes statistiques en sciences humaines*, 6ème édition. De Boeck, Bruxelles.
- Ion, R. (2007) *Metode de dezambiguizare semantică automată. Aplicații pentru limbile engleză și română*, Teză de doctorat, București: Academia Română.
- Kemper, S. (1983) Measuring the inference load of a text. *Journal of Educational Psychology*, 75(3):391-401.
- Kintsch, W. (1979) On modeling comprehension. *Educational Psychologist*, 14(1):3-14.
- Kintsch, W., Kozminsky, E., Streby, W., McKoon, G. et Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196-214.
- Kintsch, W. and Vipond, D. (1979) Reading comprehension and readability in educational practice and psychological theory, In: Nilsson, L. (ed) *Perspectives on Memory Research*, Hillsdale, NJ: Lawrence Erlbaum, pp.329-365.
- Kleiber, G. (1994) *Anaphores et Pronoms*. Louvain-la-Neuve: Duculot.
- Landauer, T., Foltz, P. et Laham, D. (1998) An introduction to latent semantic analysis. *Discourse processes*, 25(2):259-284.

- Lassalle, E and Denis, P. (2011). Leveraging different meronym discovery methods for bridging resolution in French In Anaphora Processing and Applications, Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Selected papers. LNAI, Springer.
- Longo, L. and Todirascu, A. (2010) Genre-based Reference Chains Identification for French, *Investigations Linguisticae*, 21:57-75.
- Longo, L., Todirascu, A. (2013) Une étude de corpus pour la détection automatique de thèmes. *Actes des 6e journées de linguistique de corpus (JLC 09)*, pp. 143-155.
- McNamara, D., Louwerse, M., McCarthy, P. et Graesser, A. (2010) Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292-330.
- Mason, R. and Just, M. (2004). How the brain processes causal inferences in text. *Psychological Science*, 15(1):1-7.
- Mastropieri, M. A., Leinart, A., and Scruggs, T. E. (1999) Strategies to increase reading fluency. *Intervention in School and Clinic*, 34:278-283.
- Meyer, B. (1982). Reading research and the composition teacher : The importance of plans. *College composition and communication*, 33(1):37-49.
- Myers, J., Shinjo, M. and Duffy, S. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, 26(4):453-465.
- O'Connor, R. E., Bell, K. M., and Harty, K. R. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, 94: 474-485.
- Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of EMNLP 2008*, pp. 186-195.
- Popescu-Belis A., Robba I. & Sabah G. (1998) Reference Resolution Beyond Coreference: a Conceptual Frame and its Application. *Proceedings of Coling-ACL'98 (International Conference on Computational Linguistics - Meeting of the Association for Computational Linguistics)*, Montreal, Canada, p.1046-1052.
- Salmon-Alt, S. (2001) *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*, Thèse de doctorat, Université Henri Poincaré, mai 2001
- Schnedecker C. (1997) *Nom propre et chaînes de référence*. Recherches Linguistiques 21. Klincksieck, Paris.
- Schnedecker, C. (2005) Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de Linguistique*, 2: 85-133.
- Schnedecker, C., Longo, L. (2012) Impact des genres sur la composition des chaînes de référence : le cas des faits divers, In : Neveu, F., Muni Toke, V., Blumenthal, P., Klingler, T., Ligas, P. Prévost, S., and Teston-Bonnard, S. (Eds.) *3e Congrès Mondial de Linguistique Française*, Lyon, France, July 2012, pp. 1957-1972.
- Van Dijk, T. (1977) *Text and Context: Exploration in the Semantics and Pragmatics of Discourse* . London: Longman.
- Victorri, B. (2005) Le calcul de la référence, *Sémantique et traitement automatique du langage naturel*, Patrice Enjalbert (Ed.) (2005) 133-172

Towards a cognitive-linguistic reconstruction of “mental maps” from ancient texts – The example of Dionysios Periegetes

Guenther Goerz^{1,2}, Ekaterina Ilyushechkina^{3,5}, Martin Thiering^{2,4,5}

¹ University of Erlangen-Nuremberg, Computer Science Dept., 91052 Erlangen,
goerz@cs.fau.de

² Max Planck Institute for the History of Science, 14195 Berlin

³ Free University Berlin, History Department, Friedrich-Meinecke-Institut

⁴ Technical University Berlin, Linguistics Department

⁵ Topoi Excellence Cluster, C: Perception and Representation of Space, Berlin

Abstract. We present an interdisciplinary approach to implicit knowledge of spatial cognition in common sense geography. Structures such as, e.g., distance, scale, topology, trajectory and frames of reference are believed to be encoded as mental maps. Furthermore, we refer to common sense as ‘naive’ perception and descriptions of space and the use of ‘intuitive’ arguments in geographical contexts. The empirical data sets comprise of ancient written texts from different periods and sources. Our methodology combines annotating and parsing techniques based on corpus data analysis as a (semi-) automated analysis and cognitive parameters applied in cognitive linguistics. These parameters are based on gestalt-psychological principles such as figure-ground asymmetries. We argue that the survey of ancient texts provides further insights whether there are basic epistemological expressions of spatial orientation that might be candidates for universals. As a first example, we investigate Dionysios Periegetes’ Description of the World.

1 Problem statement

An important goal of the Berlin research cluster TOPOI is the investigation of conceptualizations of space in the ancient world. Our research group (C-5) studies “Common Sense Geography”, referring to the part of historical geography concerned with implicit or tacit knowledge in ancient cultures (Geus/Thiering [13]). We specifically implement the idea of mental models as abstract cognitive representations. Common sense geography denotes a ‘lower’ geography, to be distinguished from ‘professional’ or higher’ geography, i.e., the phenomenon of the spread and application of geographical knowledge outside of expert circles and disciplinary contexts. Furthermore, common sense geography refers to a ‘naive’ perception and description of space and the use of ‘intuitive’ arguments in geographical contexts.

We survey common sense conceptualizations of geographic concepts and relations – sometimes called ‘naive geography’⁶ – and how they are used in ancient non-scientific texts with methods of cognitive linguistics with corpus construction, annotation, and parsing. Cognitive linguistic text analysis will result in formal two-level representations: A first, linguistic level, representing particular language-bound word meanings represented by semantic/logical forms, and a second, conceptual level of abstract conceptual knowledge represented by object schemata. For the latter part, the analysis should not only deliver propositional representations, but also analog representations which can be conceived of as sketches, showing the interaction of figure-ground asymmetries. There are several projects in cognitive science which emphasize such a duality, e.g. SRM by Ragni, Knauff and Nebel⁷ [43, 3]. Based on these results we aim to reconstruct cognitive maps where in ideal cases should be possible to generate sketches with a limited degree of ambiguity.⁸ Persons, landmarks, buildings and other artifacts, historical and fictitious events are recognized at certain places, associated with them, and henceforth memorized and retrieved; temporal courses of events are mapped into spatial relations.⁹

2 Our approach

2.1 Cognitive linguistics and Gestalt theory

This project examines space as a cognitive, linguistic and operative category.¹⁰ The focus is on reconstructing the mental models of ancient peoples and investigating their dealings with and movement in their environment as a knowledge system of implicit assumptions about space. Implicit knowledge is externalized here through linguistic representation. The cognitive representation of spatial relations and knowledge structures is organized and modified through mental models. Using an approach that involves cognitive linguistics, the group will devise a mental model of “common sense geography” on the basis of the Neo-Whorfian hypothesis, which argues that linguistic expressions of concepts has some degree of influence over conceptualization in cognitive domains (Levinson

⁶ Egenhofer and Mark [9]

⁷ But even in their ACT-R implementation [3] the analogical representation is constructed manually, not by a parser; personal communication by M. Knauff, 2012.

⁸ As opposed to many reconstructions by 19th century historians of cartography which show a lot of details based on implicit assumptions and presuppositions which cannot be found in the sources.

⁹ Tolman [47] was the first to introduce the term “cognitive map” to denote the cognitive representation of space and of spatial relations, first of all on the individual level, later on augmented by a layer of social communication. Kitchin and Blades [26] and MacEachren [37] provide a representative and comprehensive overview from the viewpoint of psychology and cognitive science.

¹⁰ cf. mental spaces theory (Fauconnier [10, 11]); Knauff [27] on spatial layout models based on Johnson-Laird’s mental model theory [22]

[35]). Common sense geography can be characterized as an especially illuminating example of implicit knowledge. The project will address implicit processes of mental orientation that depict culture-specific – but also eventually universal – representations of cognitive structures. The material assembled and edited by the group members (e.g. on landmarks, distance specifications, frames of reference systems, scales) will enable the group to investigate the linguistic representation of implicit and shared knowledge. The goal is to create mental models that hold not only for ancient texts, but also represent eventually universal structures.

As is argued here, mental models are based on universal cognitive mechanisms and gestalt-theoretical principles (Thiering [46, 45]). Of particular relevance is the spatial division of figure and background in identifying and locating objects (Talmy [44]). This asymmetry corresponds roughly to the linguistic categories of subject and object. Spatial relations are represented through grammatical markers and semantic fields. Mental models consequently play a decisive role in spatial representation. As abstract cognitive representations, they store information of the events and objects of the external world.¹¹ This holds for all types of experiences that are stored non-verbally and mentally in memory, but especially for orientation in space and references to places, as well as for topological and geometrical knowledge. These categories will be identified and compared so that fundamental spatial knowledge structures can be determined.

2.2 Propositional and analog-depictional representations

A basic theoretical assumption of the LILOG project¹² which we share is that there are two connected levels of representation: a linguistic and a conceptual one (cf. Jackendoff [21]). The generic, conceptual level abstracts from linguistic features of a particular language, but avoiding “the mistake of ascribing (whatever sort of) ‘psychological reality’ to formal models...” ([29], p. 10). At the conceptual level we find cognitive maps encoding various spatial parameters that get into language. Languages differ in the way they parse-up spatial relations linguistically (Levinson [35]; Levinson and Wilkins [36]; Thiering [46]). Linguistic analysis results in semantic constructions represented in a Description Logic

¹¹ Fauconnier [10, 11], Johnson-Laird [22], Knauff [27].

¹² IBM Germany’s LILOG (“LIinguistics and LOGic”) project (1986–1992) has a lot in common with our approach to reconstruct cognitive maps. LILOG’s objective was “to develop a text comprehension system that extracts knowledge from texts resulting in representations used to answer questions about those texts in a natural language dialogue” (Lang et al. [29], p. 6). To our knowledge LILOG is the only system which fully implemented the linking of propositional with analog (“depictional”) representations of spatial knowledge. “This strategy encompasses the structural analysis of linguistic expressions of motion and localization (typical of route descriptions, tourist guides, etc.), which primarily draw on topological relations in large-scale space environments. This is where the present work enters the picture by complementing topology with geometry, thereby reflecting the fact that spatial knowledge is organized by the interaction of topological (localizations of and distances between objects in large-scale space) and geometric principles (axes and positions of objects in small-scale space).” (Lang et al. [29], p. 6).

language. The spatial subcomponent consisted of two interacting modules:¹³ an inference engine equipped with rules for spatial reasoning to derive propositional knowledge and a depictorial component which operated on an analogical representation (Khenkar [25, 24]).

Our plan is to adapt and develop tools which take up the results of a broad coverage parser and generate similar two-level representations according to our specifications (see below). In our case, depictional representations will serve as sketches of cognitive maps to represent and process reifications of cognitive objects on an epistemological level, i.e. frame of reference, topology, direction, trajectory, distance, and shape.¹⁴

2.3 Preliminary work: Linguistic preprocessing

The architecture for cognitive linguistic analysis of texts comprises the following steps, which should in principle be interactive, i.e. information should flow on both directions, bottom-up and top-down. In fact, it is more or less unidirectional due to the heterogeneity of software modules we are employing. The steps are in general: (1) lemmatization, (2) part-of-speech (POS) tagging, (3) chunking / syntactic parsing, (4) semantic role labeling and constraint-based construction with (a) word-sense disambiguation and (b) co-reference resolution, (5) cognitive linguistic description and markup. For ancient Greek and Latin, only tools for the first two steps exist; not even for Latin are more than are more than a limited number of experimental parsers. This is why at this stage we use translations.¹⁵

A representative ancient text of common sense geography is Dionysios Periegetes' (2nd century CE) "Description of the World" in hexameter verse [48].¹⁶ In this case, we digitized the English translations by Greaves [14] and Khan [23]¹⁷ and two German translations, Brodersen [4] and Fruhwirth [12]. All of them have been encoded in TEI¹⁸ so that they can be aligned for comparison – certainly this is a necessary task for checking the translation and use of propositions and other words which are immediately relevant for spatial entities and relations.

First of all, we ran a few utilities on the English translation for heuristic purposes, to get an overview of the vocabulary, and of the use of prepositions

¹³ Habel and Pribbenow [19, 42, 41], see also Latecki et al. [31–33]

¹⁴ But, of course, the interaction between both parallel representations, the propositional and the analogical, is still a research problem.

¹⁵ For some texts, the Perseus Digital Library (<http://www.perseus.tufts.edu/hopper/>; 02.06.2013) holds for some texts a tree bank with word-by-word morphological and dependency-based syntactic information (AGDT), but unfortunately for none of those which are pertinent for our purpose.

¹⁶ Plain text in Thesaurus Linguae Graecae, <http://www.tlg.uci.edu/>; 02.06.2013.

¹⁷ A closer look at both translations showed that both translations are not close enough to the poetic text; so we are planning to elaborate a new one which is better suited for alignment.

¹⁸ <http://www.tei-c.org/>; 02.06.2012.

and other lemmas which are immediately relevant for spatial entities and relations. To get precise ideas about the text, some tools¹⁹ have been applied to the raw text: Word (form) lists, inverse word form lists (for the endings), bigrams, trigrams, frequency lists, concordances (KWIC index), and word (form) cooccurrences. Here, only a selection of quantitative results about the text can be given: The Greek text consists of 1186 verses made up from 7658 word tokens (of 3405 types). The English prose translation has 11396 word tokens (2217 types) in 437 sentences and there are 1047 full and 269 auxiliary verb tokens. The most frequent of 1245 preposition tokens are *in* (135), *toward* (89), *from* (75), *to* (57), *into* (51), and *on* (48). The most frequent content words are *sea* (124), *land* (107), *mountain* (45), *island* (43), *river* (36), *stream* (29), *ocean* (28) and the four main directions *east* (56), *north* (52), *south* (43), *west* (35). The three continents are mentioned *Asia* (27), *Libya* (= Africa! 23), *Europe* (18). Some results of semantic tagging, considering 572 words tagged as unmatched, where 427 are proper names, we have 540 toponyms, 544 geographical terms, 492 words of location and direction, 132 of moving, and 132 denoting places; 44 words refer to distance. For cognitive linguistic analysis, directional and distance expressions with source and goal play an important role. Therefore, a brief look at the most frequent trigrams is instructive: e.g., “*as far as*” occurs 22 times, “*toward the [direction]*” 40 times.

Based on these investigations and the parsing results described below we set up linguistic annotation of the TEI encoded texts – POS-Tags and lemmata as well as sentence boundaries while taking account of authority files, for disambiguation, also a hierarchical lexicon such as WordNet. We are currently marking up all toponyms and will check with common gazetteers such as Pleiades in Pelagios²⁰ or the Getty Thesaurus of Geographic Names²¹ in order to identify places and extract their coordinates. So it will be possible to show them on a modern map by means of GIS visualization tools, e.g. Europeana4D²², but also on historic maps.

3 Workflow for the semi-automatic analysis

The whole process would be at best semi-automatic due to the lack of computational linguistics tools for the classic languages. The parsing results as described below will have to be revised and corrected manually and can be turned into TEI markup by applying a few scripts. The linguistically marked up TEI files are the basis for the task of constructing semantic and cognitive representations. Important for this step is the recognition of frames of reference that can either be relative, intrinsic or absolute and the encoding of spatial relations, in particular

¹⁹ Command line Unix tools as well as the corpus tools Antconc and Voyant: http://www.antlab.sci.waseda.ac.jp/antconc_index.html and <http://voyeurtools.org>; 02.06.2013.

²⁰ <http://pelagios-project.blogspot.de/>; 02.06.2013.

²¹ <http://www.getty.edu/research/tools/vocabularies/tgn/>; 02.06.2013

²² <http://www.tinyurl.com/e4d-project>; 02.06.2013.

the recognition of figure and ground (Levinson [34, 35]). In the long run, we are aiming at a further transformation into a treebank representation (cf. Mambrini [38]) for which tools are currently under development.

3.1 Parsing

We applied syntactic and semantic taggers²³ which provide wordclass and some morphosyntactic information. The next step was to run grammatical parsers on the text to get constituent structures and dependency structures which are important for semantic representation. Among available parsers, we used the Stanford Parser²⁴, TsujiLab’s Enju²⁵, and the mate-tools parsers (Lund Univ.), in particular the semantic parser SRL²⁶ als well as the University of Illinois Semantic Role Labeler²⁷. The last three parsers also generate predicate-argument structures. Considering the syntactic complexity of the text and due to adverse tagging preferences, in a number of cases with first two parsers preferred attachment of substructures in constituent structures which are wrong from a semantic point of view. Lund and Illinois SRL delivered the best results so far.

3.2 Semantic structures

To give an example, we take verses 93 sqq. of Dionysios Periegetes’ “Description of the World”: *“From there, Adriatic brine widens, extending into the north, and it creeps back again toward a westerly nook, which the neighboring peoples also named the Ionian Sea. It disgorges itself upon two lands: as you enter this sea, the Illyrian land appears on the right hand, and above is Dalmatia, land of warlike men; and on the left side extends the immense isthmus of the Ausonians, far-stretching”*. In the poem, the author takes the reader to an imaginary voyage, where in these sentences the Ionian Sea is approached from the Adriatic Sea. In these verses, he demonstrates a directional feature in the localization of geographical objects from the perspective of an imaginary observer.

We focus on the first sentence and present only the predicate-argument structure (PAS) as generated by SRL (fig. 1). “Widen” in its first reading 01 is the main predicate with one argument, A1 (in the verb frame lexicon “the thing becoming wider”), and an adverbial modifier, which contains the predicate “extend” with two arguments, A0 (verb frame: “agent”) and A2 (“EXT: how much”), etc.

For semantic representations, we aim at a (flat) variant of Discourse Representation Structures (DRS) with which we gained experience in previous

²³ Stuttgart Treetagger (<http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html>; 02.06.2013), Lancaster CLAWS and USAS (semantic) taggers (<http://ucrel.lancs.ac.uk/wmatrix/>; 02.06.2013).

²⁴ <http://nlp.stanford.edu/software/lex-parser.shtml>; 01.06.2013. The stand-alone tool “grammarscope”, using the same grammar, can be downloaded from this website.

²⁵ <http://www.nactem.ac.uk/enju/>; 02.06.2013.

²⁶ Semantic Role Labeler, <http://nlp.cs.lth.se/semantics/>; 01.06.2013.

²⁷ <http://cogcomp.cs.illinois.edu/demo/srl/>; 08.08.2013.



Fig. 1. Example: predicate-argument structure, split in two halves

projects²⁸. The nesting of the PAS as shown in the example together with the reference structure which is implicit in fig. 1 (e.g. A0 “it”) contains all the required information; only a slight syntactic transformation to extract the discourse referents (e.g., “Adriatic brine”) and to explicitly represent the verb frame with named roles has to be applied²⁹. As for an automatic construction of DRSs, we are currently evaluating the recently released “C & C tools” which contain a parser based on Combinatory Categorial Grammar (CGG) and the computational semantics tool Boxer³⁰.

For semantic enrichment of PASs, the USAS tags, which provide a thesaurus-based semantic word classification, are included. If there is a need for even richer semantic indexing, domain models and “formal ontologies”³¹ can be integrated which provide much more content about the conceptualization of space and the geographic domain than the semantic lexicon entries.³² In particular for the

²⁸ Discourse Representation Theory is the most well known dynamic semantic theory which is able to represent the meaning of more than just one sentence – what is required to resolve reference. Discourse referents, introduced by e.g. indefinite noun phrases, named entities, etc., are explicitly represented.

²⁹ Furthermore, for DRSs, we introduced a neat way for an event-based representation inspired by Donald Davidson: i.e., for events introduced by verbs, generate a discourse referent e and then have $\text{particularevent}(e)$ & $\text{actor}(e,A)$ & $\text{goal}(e,G)$ & $\text{place}(e,P)$ &... instead of n-place predicates with reification. Such structures are particularly appropriate for representations with Description Logics, which we prefer for ontological reasons, see below.

³⁰ <http://svn.ask.it.usyd.edu.au/trac/candc/wiki>; 08.08.2012.

³¹ A formal ontology defines the terminological system of a domain of discourse (Noy [40]) as the kernel of appropriate domain theories. In ideal cases it draws on their resp. context of justification and uses their underlying abstraction methods (e.g., Guarino [15, 39]).

³² For the case of thesauri implemented in SKOS (Simple Knowledge Organization System, W3C Recommendation: <http://www.w3.org/TR/skos-primer/>; 08.08.2013)

level of generic knowledge, e.g., on time and space, so-called reference ontologies provide a framework in which specific domain knowledge can be incorporated. The CIDOC Conceptual Reference Model³³ is one such reference ontology. Using the CRM opens up a wide spectrum of interoperability and linking to many web resources, such as the gazetteers mentioned above. Ontological enrichment with CRM, which in its basic design is event-based and hence compatible with the representation scheme just mentioned, would provide a generic “assignment event” which has open positions to be filled or linked with the PAS, resp., for agent, (material and immaterial) ingredients, time-span, and place.

At least at this point, it will be necessary for semantic evaluation and disambiguation to apply some formal reasoning to the semantic/cognitive representations. For ontologies implemented in OWL-DL powerful reasoning engines³⁴ are available which can efficiently solve consistency checking, classification and retrieval problems. Reasoning also helps to identify implicit information. Most of the processing will still have to be done manually; finally we expect to end up with annotated logical forms which express the spatial relations of objects described in the text.

3.3 Cognitive analysis

The example sentence has shown some significant numbers of spatial occurrences encoded in, e.g., spatial adpositions, toponyms, and landmarks. The following imaging parameters apply here with respect to the figure-ground asymmetry:

- Figure: Adriatic brine_{FIG1} / it (sc. Adriatic brine)_{FIG2} / it (sc. Ionian Sea)_{FIG3} / you_{FIG4} / Illyrian land_{FIG5} / Dalmatia_{FIG6} / isthmus_{FIG7} / each sea_{FIG8} / [Tyrrhenian_{FIG9} / Sicilian_{FIG10} / Ariatic_{FIG11}]
- Ground: there (sc. Iapigian land)_{GND1} / nook_{GND2} / two lands_{GND3} / sea_{GND4} / three seas_{GND5} / wind_{GND6} / [the west wind_{GND7} / the south (sc. wind)_{GND8} / the east (sc. wind)_{GND9}]

Clearly, this short text example presents already a variety of parameters to set a mental map using various reference points based on toponyms and landmarks. More specifically, also two frames of reference are at work, the absolute and the relative (arguably, also the intrinsic frame): absolute (cardinal points: “north/west(erly)” and winds); relative (from the viewpoint of the imaginary traveler) + deictic (this). Another important semantic information is encoded via the trajectory and motion event between a figure and a ground, or rather between the source and goal.

and ontologies implemented in OWL-DL, the SKOS Primer shows in section 5.2 how to bridge lexical concepts and ontological concepts, a solution we provided in our Virtual Research Environment WissKI, cf. Goerz/Scholz [17]

³³ <http://www.cidoc-crm.org/>; 02.06.2013; since 2006 ISO standard 21127, originally defined for the cultural heritage sector. We implemented it in a Description Logic Language, the Semantic Web Ontology language OWL-DL [16], cf. <http://erlangen-crm.org/>; 02.06.2013.

³⁴ e.g., Pellet <http://clarkparsia.com/pellet/>; 02.06.2013

- (1) X goes from point Y in direction of Z = Trajectory (source and goal)
 + motion (not linear, but crooked: “into the north” / and then “back toward the west(erly nook)”)

The region is profiled by part-whole constructions, e.g., one sea devided into two parts as presented by the two names (“*Adriatic/Ionian*”). Another example is the *Illyrian* land (= land₁) + *Dalmatia* (land₂ = *isthmus*).

Also, the region’s scope is encoded as bounded/unbounded (bounded = “neighboring peoples”). The unbounded region entails information about the frontier land that implicitly demonstrates the large extension of the sea itself (“extends”, “immense”, “far-streching”). Here some implicit information of the boundaries and shape is profiled (“encircled”). These rather static spatial relations are accompanied by dynamic relations based on mental mapping processes. One such relation is encoded as trajectory, e.g. a horizontal movement from a source to a goal.

- (2) X Enter Y = Trajectory (source + goal) + motion

The preposition (X enter Y) profiles the trajectory from a starting point to a (real or imagined) endpoint. Hence, here the mapping process is what Thiering calls mental triangulation process of figure-ground asymmetries [46].

Finally, spatial adpositions such as “above” profile a non-contact relation between a figure and a ground implying a cartographic view/bird’s-eye view and by that eventually a specific perspective. A vertical alignment between figure and ground is encoded. Prototypically this alignment entails no contact and no movement, and a specific immediate region and scope of the figure-ground alignment (Levinson [35]). Actually, the cognitive-semantic markup is done semiautomatically with the help of scripts applied to the POS-USAS-tagged XML text, the predicate-argument structure and, in addition for adpositions, the dependency structure as well.

The next and final step is a mental map representation of the analogue systems involved here. Fig. 2 shows a (manually made) sketch which could be taken as a starting point for the reconstruction of a mental map. We base our description on Langacker’s [30] terminology and graphic representations using the various imaging parameters introduced above. He presents specific images for the description of, e.g., bounded/unbounded region, sequential and serial scanning etc.

4 Outlook on cognitive maps: reconstruction and inference

We conclude with an outlook on the formal methods to be applied to reconstruct map sketches, i.e., analogue representations similar to the ones mentioned with LILOG.

“In the last analysis all maps are cognitive maps” – this thesis put up by Blakemore und Harley [2] indicates an important topic of recent research in the

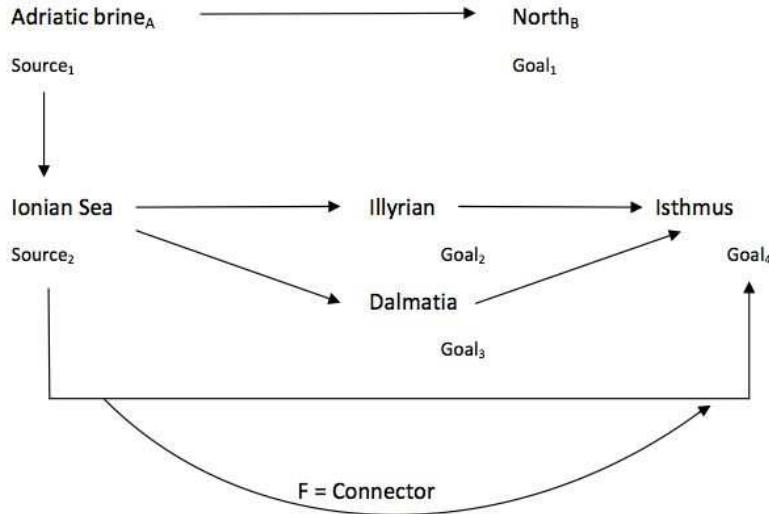


Fig. 2. Mental map sketch of the example

history of cartography. Georeferencing, i.e., reference to geographical locations, is the underlying principle for organizing and presenting all kinds of information in maps.

There are several formal approaches to qualitative theories of geographic (Euclidean) space which are suitable for spatial reasoning³⁵. Vieu [49] has elaborated a theory which is particularly well suited for our approach. On the basis of mereology as an axiomatized part-whole relation, she provides a formalization of topological concepts as well as geometrical concepts, in particular distance and orientation, in first-order logic.

Easy access to cognitive mapping can be achieved by investigating the questions of “where”, “what”, and “when” systematically.³⁶ For the “where”, i.e. spatial information in a proper sense, naming is an elementary means to determine identity³⁷. For a description of places, depending on the frame of reference, a specification of states or processes, and of distance and direction must be added. An example of a process specification would be a route description, how a certain place can be reached. “What” – the objective – and “when” become important for the solution of spatial problems: a set of suitable properties must

³⁵ e.g. Egenhofer [9], Vieu [50], Hernandez [20].

³⁶ cf. Landau and Jackendoff [28].

³⁷ It should be noted that not all languages endorse “where” and “what” questions. Brown clearly shows that Tzeltal does not use where and what [5]. Hence, it is not clear how universal these markers are. For an extensive crosslinguistic overview see Levinson and Wilkins [36].

be given which are useful to find a solution by means of cognitive mapping. In other words, first of all, we have to identify the elements which are necessary for an epistemological organization of spatial knowledge. In a second step we need to develop a analogical/depictional representation suitable for computational processing. Obviously, regions and their relative positions play a key role as well as directions or orientation, resp., and distance. To perceive these elements, to identify them and to refer to them in discourse is an accomplishment in abstraction which has in any case also a cognitive foundation. So, basically we are considering the assembly of maps and their description in terms of those primarily qualitative categories with (qualitative) spatial reasoning in mind.

So far, our heuristic approach comprised the following processing steps with the goal to provide all information items for the construction of mental maps:

1. Lemmatization and POS tagging
2. Semantic tagging
3. Grammatical parsing into dependency structures (and also constituent structures), in particular focussing on prepositions
4. Semantic construction (predicate-argument structures)

For the final step, the generation of a cognitive representation, the salient data to be considered are

- Toponyms (1, 2)
- Enumeration of ethnogeographical terms (1, 2) including modifiers – size, shape, etc. – (3)
- Spatial relations, directions, etc. from prepositional phrases (3)
- Subject / object from predicate-argument structures
- Frame of reference with the help of movement and position verbs (2) and toponyms

In the long run, it is required to combine spatial reasoning with Description Logic inferences. On the technical level, topological and orientation relations as well as distance and size can be represented by specific datatypes and processed by special constraint solvers. The “Region Connection Calculus” (RCC-8), an elementary topological theory with regard to qualitative spatial reasoning, has been developed by Cohn et al. [7]. From the – logically formulated – theorems of the so called RCC-8 theory a composition table can be derived which can very efficiently solve such specific tasks. An interesting question is whether the theoretically identified primitives (such as in RCC-8) are not only epistemologically plausible, but also cognitively.

The integration with a logical representation framework leads to a system of hybrid reasoning: With Pellet Spatial³⁸ a reasoning system has been created which combines Description Logic inferencing with spatial reasoning. In a previous project in the domain of the history of cartography we also implemented prototypically an extension of a cartographic ontology in Description Logics³⁹. A

³⁸ <http://clarkparsia.com/pellet/spatial/>; 02.06.2013.

³⁹ Görz, Scholz [18] and Deang [8].

lot of research has been done recently to further integrate orientation, distance and shape which we will have to take up to describe maps in terms of these primarily qualitative categories⁴⁰. We expect that the investigation of ancient texts can provide insights whether there are basic epistemological expressions of spatial orientation.

Acknowledgement. We would like to thank the Berlin Common-Sense-Geography Group, especially Klaus Geus and Elton Barker for suggestions and comments. We are also grateful to Martin Scholz for his advice on the visualization of parsing results.

References

1. Batsakis, S., Petrakis, E.: SOWL: A framework for handling spatio-temporal information in OWL 2.0. In: 5th International Symposium on Rules: Research Based and Industry Focused (RuleML 2011). pp. 242–249. Barcelona (July 2011)
2. Blakemore, M., Harley, J.: Concepts in the History of Cartography — A Review and Perspective, *Cartographica — International Publications on Cartography*, vol. 17/4, Monograph 26. University of Toronto Press, Toronto (1980)
3. Boeddinghaus, J., et al.: Simulating spatial reasoning using ACT-R. In: Proceedings of the 7th International Conference on Cognitive Modeling. pp. 62–67. ICCM, Trieste (April 2006)
4. Brodersen, K. (ed.): Dionysios von Alexandria: Das Lied von der Welt. Zweisprachige Ausgabe. Georg Olms Verlag, Hildesheim and Zürich and New York (1994)
5. Brown, P.: A sketch of the grammar of space in tzeltal. In: *Grammars of Space*, pp. 230–272. Cambridge University Press, Cambridge (2006)
6. Christodoulou, G., Petrakis, E., Batsakis, S.: Qualitative spatial reasoning using topological and directional information in OWL. In: 24th International Conference on Tools with Artificial Intelligence (ICTAI 2012). pp. 1–7. Athens (November 2012)
7. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Representing and reasoning with qualitative spatial relations about regions. In: *Spatial and Temporal Reasoning*, pp. 96–134. Kluwer Academic Publishers, Dordrecht and Boston and London (1997)
8. Deang, D.M.: Geometrical and logical modelling of cartographic objects. Master thesis in computational engineering, University of Erlangen-Nuremberg, Erlangen (October 2000)
9. Egenhofer, M., Mark, D.: Naive geography. In: *Spatial Information Theory. Proceedings of Conference on Spatial Information Theory (COSIT'95): A Theoretical Basis for GIS*. Semmering, Austria, Lecture Notes in Computer Science, vol. 988, pp. 1–15. Springer, Berlin (1995)
10. Fauconnier, G.: Mental Spaces: Aspects of Meaning Construction in Natural Language. MIT Press, Cambridge, MA (1994/1985)
11. Fauconnier, G.: *Mappings in Thought and Language*. Cambridge University Press, New York (1997)
12. Fruhwirth, A.: Die Periegese des Dionysios von Alexandria. Einführung und Übersetzung. Tech. rep., Institut für Klassische Philologie, Universität Graz, Graz (1990), Diplomarbeit zur Erlangung des Grades eines Magisters der Philosophie

⁴⁰ Among others Batsakis, Christodoulou et al., e.g. [1, 6].

13. Geus, K., Thiering, M. (eds.): Common Sense Geography and Mental Modelling, Preprints, vol. 426. Max Planck Institute for the History of Science, Berlin (2012)
14. Greaves, D.D.: Dionysios Periegetes and the Hellenistic Poetic and Geographical Traditions. Phd thesis, Department of Classics, Stanford University, Stanford (June 1994)
15. Guarino, N.: Formal ontology and information systems. In: Guarino, N. (ed.) Formal Ontology in Information Systems. Proceedings of FOIS-98, Trento, Italy, 6–8 June 1998. pp. 3–15. IOS Press, Amsterdam (1998)
16. Görz, G., Oischinger, M., Schiemann, B.: An implementation of the cidoc conceptual reference model (4.2.4) in owl-dl. In: Proceedings CIDOC 2008 — The Digital Curation of Cultural Heritage. Athen, Benaki Museum, 15.–18.09.2008. pp. 1–14. ICOM CIDOC, Athen (September 2008)
17. Görz, G., Scholz, M.: Adaptation of nlp techniques to cultural heritage research and documentation. Journal of Computing and Information Technology (CIT) 18(4), 317–324 (2010), extended version of conference paper in Luzar-Stiffler, Vesna and et al.: Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, June 21–24, 2010, Cavtat/Dubrovnik, 79–84
18. Görz, G., Scholz, M.: Semantic annotation for medieval cartography: The example of the Behaim globe of 1492. e-Perimetron 8(1), 26–36 (2013), <http://www.e-perimetron.org/>
19. Habel, C.: Processing of spatial expressions in LILOG. In: Text Understanding in LILOG. Integrating Computational Linguistics and Artificial Intelligence. Final Report on the IBM Germany LILOG-Project, pp. 598–608. No. 546 in Lecture Notes in Computer Science, Springer, Berlin etc. (1991)
20. Hernández, D.: Qualitative Representation of Spatial Knowledge, Lecture Notes in Computer Science, vol. 804. Springer, New York (1994)
21. Jackendoff, R.: Semantics and Cognition. MIT Press, Cambridge, MA (1983)
22. Johnson-Laird, P.N.: Mental Models. Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge, MA (1983)
23. Khan, Y.Z.: A commentary on Dionysios of Alexandria's guide to the inhabited world, 174–382. Phd thesis, University College, London (2010), complete translation contained
24. Khenkar, M.N.: Eine objektorientierte darstellung von depiktionen auf der grundlage von zellmatrizen. In: Repräsentation und Verarbeitung räumlichen Wissens. pp. 99–112. No. 245 in Informatik-Fachberichte, Springer, Berlin etc. (1990)
25. Khenkar, M.N.: Object-oriented representation of depictions on the basis of cell matrices. In: Text Understanding in LILOG. Integrating Computational Linguistics and Artificial Intelligence. Final Report on the IBM Germany LILOG-Project, pp. 645–656. No. 546 in Lecture Notes in Computer Science, Springer, Berlin etc. (1991)
26. Kitchin, R., Blades, M.: The Cognition of Geographic Space. I.B. Tauris Publishers, London and New York (2002)
27. Knauff, M. (ed.): Space to Reason: A Spatial Theory of Human Thought. The MIT Press, Cambridge, MA (2003)
28. Landau, B., Jackendoff, R.: “what” and “where” in spatial language and spatial cognition. Behavioral and Brain Sciences 16, 217–238 (1993)
29. Lang, E., Carstensen, K.U., Simmons, G. (eds.): Modelling Spatial Knowledge on a Linguistic Basis. Theory – Prototype – Integration. No. 481 in Lecture Notes in Artificial Intelligence, Springer, Berlin and Heidelberg (1991)
30. Langacker, R.W. (ed.): Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites. Stanford University Press, Stanford (1987)

31. Latecki, L., Pribbenow, S.: On hybrid reasoning for spatial expressions. In: Proceedings of the 10th European Conference on Artificial Intelligence, ECAI-92. pp. 389–393. John Wiley & Sons, New York (1992)
32. Latecki, L., Pribbenow, S.: Orientation and qualitative angle for spatial reasoning. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-93, Chambéry, France. pp. 1–6. Morgan Kaufman, San Mateo (1993)
33. Latecki, L., Röhrlig, R.: Räumliches schließen und berechenbarkeit. Künstliche Intelligenz 7(4), 35–43 (1994)
34. Levinson, S.C.: Language and space. Annual Review of Anthropology 25, 353–382 (1996)
35. Levinson, S.C.: Space in Language and Cognition. Exploration in Cognitive Diversity. Cambridge University Press, Cambridge et al. (2003)
36. Levinson, S.C., Wilkins, D. (eds.): Grammars of Space. Cambridge University Press, Cambridge (2006)
37. MacEachren, A.M.: How Maps Work. Representation, Visualization, and Design. The Guilford Press, New York and London (1995)
38. Mambrini, F., Passarotti, M.: Will a parser overtake achilles? first experiments on parsing the ancient greek dependency treebank. In: Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11). pp. 133–144. Edições Colibri, Lisboa (2012)
39. Menzel, C.: Ontology theory. In: Euzenat, J., Gomez-Perez, A., Nicola, G., Stuckenschmidt, H. (eds.) Ontologies and Semantic Interoperability, Proc. ECAI-02 Workshop. CEUR-WS, vol. 64, pp. 61–67. ECCAI, Lyon (2002)
40. Noy, N.: Ontologies. In: Farghaly, A. (ed.) Handbook for Language Engineers, pp. 181–211. CSLI Publications, Stanford, CA (2003)
41. Pribbenow, S.: Interaktion von propositionalen und bildhaften repräsentationen. In: Repräsentation und Verarbeitung räumlichen Wissens. pp. 156–174. No. 245 in Informatik-Fachberichte, Springer, Berlin etc. (1990)
42. Pribbenow, S.: Phenomena of localization. In: Text Understanding in LILOG. Integrating Computational Linguistics and Artificial Intelligence. Final Report on the IBM Germany LILOG-Project, pp. 609–620. No. 546 in Lecture Notes in Computer Science, Springer, Berlin etc. (1991)
43. Ragni, M., Knauff, M., Nebel, B.: A computational model for spatial reasoning with mental models. In: Proceedings of the 27th Annual Cognitive Science Conference. pp. 1797–1802. L. Erlbaum, Mahwah, NJ (2005)
44. Talmly, L.: Towards a Cognitive Semantics, Vol. I and II. MIT Press (2000)
45. Thiering, M.: Figure-ground reversals in language. Gestalt Theory 33(3/4), 245–276 (2011)
46. Thiering, M.: Degree of specificity in spatial semantics. In: Variation in Language and Language Use: Linguistic, Socio-Cultural and Cognitive Perspectives, pp. 367–420. No. 96 in Duisburg Papers on Research in Language and Culture, Lang, Frankfurt/Main (2013)
47. Tolman, E.: Cognitive maps in rats and men. Psychological Review 55, 189–208 (1948)
48. Tsavari, I.O. (ed.): Dionysiou Alexandreos Oikumenes periegesis. Ioannina (1990)
49. Vieu, L.: A logical framework for reasoning about space. In: Spatial Information Theory. A Theoretical Basis for GIS. European Conference, COSIT'93. Lecture Notes in Computer Science, vol. 716, pp. 25–35. Springer-Verlag, Berlin, etc. (September 1993)

50. Vieu, L.: Spatial representation and reasoning in artificial intelligence. In: Spatial and Temporal Reasoning, pp. 5–41. Kluwer Academic Publishers, Dordrecht and Boston and London (1997)

Automatic Sentence Clustering to help authors to structure their thoughts

Michael Zock¹ Debela Tesfaye²

¹ Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France
michael.zock@lif.univ-mrs.fr

²IT doctoral program, Addis Ababa University, Addis Ababa, Ethiopia
dabookoo@yahoo.com

Abstract : To produce written text can be a daunting task, presenting a challenge not only for high school students or second language learners, but actually for most of us, including scientists and PhD students writing in their mother tongue. Text production involves several tasks: *message planning* (idea generation: what to say?), *text structuring* (creation of an outline), *expression* (mapping of content onto linguistic form) and *revision*. We will address here only *text structuring* which is probably the most challenging task implying the grouping, ordering and linking of messages which at the onset of doing so (the moment of providing the conceptual input) lack this kind of information. We are particularly interested in the answer to the following question: on what grounds do writers 'see' connections between message, ideas or thoughts to impose some kind of order, allowing them to group messages into categories? As this is a very complex problem on which we hardly have begun to work, we will present here only preliminary results based on a very simple example, applying mainly to descriptions, one of the many text-types.

1. The problem

Sentence generation involves three major tasks : idea generation, translation into language,(words, sentences) and expression in spoken or written form (Levelt, 1989; Reiter & Dale, 2000). Text production, in particular the written mode implies another task, *discourse structuring* (Andriessen et al. 1996, de Beaugrande, 1984, Flower & Hayes, 1980). Ideas¹ have to be grouped, ordered and linked, as if not the reader may misunderstand or not understand at all. Being unable to see the connection between the parts, he cannot make sense of the 'whole (text). The document is perceived as a set of unrelated segments, the discourse being incoherent.

Discourse structuring is a real challenge, because the ideas to be conveyed generally lack the links we need for discourse structuring, and ideas tend to come to our mind in any order, i.e. via association (Iyer et al., 2009). Hence, 'order' is in this

¹ We will use the following terms synonymously: ideas, thoughts, messages, conceptual input. Functionnally speaking they are the same for us, as all of them are basically conceptual and linguistically unspecified, i.e. pre-linguistic.

case a by-product, depending only on the relative associative strength between two items, a prime (doctor) and a probe (target: e.g nurse). This kind of order is of course very different from the one see or expect in normal texts where ideas are linked conceptually, i.e. via a common denominator (topic), via a temporal or causal link or rhetorically (concession, rebuttal, ...). Indeed, it is quite rare that text elements (propositions or sentences) are related only on the basis of statistical considerations (weight, frequency, ...). In conclusion, ideas come to our mind in an order that is fundamentally different from the one in which they will appear in the final output, well structured text. What makes things worse is the fact that the information needed to impose order on these data is generally absent in the conceptual input, i.e. the messages to be conveyed. This information has to be inferred. This is where the problem lies, and this is why discourse structuring is so hard, presenting a challenge not only for high school students or second language learners, but for most of us, including scientists and PhD students writing in their mother tongue. This being so one may wonder whether and to what extent computers could help.

Before taking a look at the work done by computational linguists, one may consider the work done by cognitive psychologists and rhetoricians on whose theories these applications may rest. Clearly, a lot has been written on writing.² Yet, despite the vast literature on composition and despite the recognition of the paramount role played by idea structuring (outlining) for yielding readable prose, little has been produced to clarify what it takes concretely speaking to achieve this goal (i.e. to help authors). No doubt, a lot of good and interesting research on writing can be found in the mentioned literature and the book series 'Studies in Writing', edited by G. Rijlaarsdam,³ but even there, one will find next to nothing on the subject we are interested in, the grouping of ideas or the discovery of links between messages or thoughts.

2. Related Work

One may take a look at the work done by another community, computational linguists working on *text* generation (Reiter & Dale, 2000, Bateman & Zock, 2003). Their ambition consists in the automatic production of texts based on messages and goals.⁴ Since everyone seems to agree with the fact that texts are structured (Mann & Thomson, 1987), this seems the right place to go. Alas, even there one will be disappointed. To avoid misunderstandings, the work produced by this community is important and impressive in many ways. Nevertheless, it seems to be based on assumptions incompatible with respect to our goal, which is to assist a writer in text

² Among others: Alamargot & Chanquoy, 2001; Bereiter & Scardamalia, 1987; Flower & Hayes 1981, 1980; Kellogg, 1999; Levy & Ransdell, 1996; Matsuhashi, 1987; Olive & Levy, 2001; Rijlaarsdam, van den Bergh, & Couzijn, 1996; Rijlaarsdam & Van den Bergh, 1996; Torrance & Jeffery, 1999. For more pointers see: <http://www.writingpro.eu/references.php>

³ <http://www.emeraldinsight.com/products/books/series.htm?id=1572-6304>

⁴ This is often seen as a top-down process : goals triggering ideas, i.e. messages, which trigger words, which are inserted in some sentence frame, to be adjusted morphologically, and so forth and so on.

production, i.e. help her or him to organize a set of ideas that prior to that point were a more or less random bunch of thoughts (at least for the reader). Here are some of the reasons why we believe that this kind of work is not compatible with our goal. First of all, interactive generation (our case) is quite different from automatic text generation. Next, most text generators are based on assumptions that hardly apply in normal writing : (a) *all the messages* to be included in the final document are available at the very moment of building the text plan (Hovy, 1991); (b) ideas are retrieved *after* a text plan has been determined (McKeown, 1985), or the two are done more or less in parallel (Moore & Paris, 1993); (c) the links between ideas (messages) or the topics to be addressed are all known at the onset of building the text plan. This last point applies both to Marcu's work (Marcu, 1997) and to data-based generators (Reichenberger et al., 1995).

Practically all these premises can be challenged, and none of them accounts for the psycholinguistic reality of composition, i.e. text production by human beings (Bereiter & Scardamalia, 1987; Andriessen et al. 1996). For example, authors often do not know the kind of links holding between ideas,⁵ neither do they always know the topical category of a given message.⁶ Both have to be inferred. Authors have to discover the link(s) between messages and the nature of the topical category to which a message or a set of messages belongs. Both tasks are complex, requiring a lot of practice before leading to the skill of good writing (coherent and cohesive discourse).

The above mentioned work also fails to model the dynamic interaction between idea generation (messages) and text structure, Simonin's work (1988) being arguably an exception. Indeed, a topic may trigger a set of ideas (top-down generation), just as ideas may evoke a certain topic (bottom-up), and of course, the two can be combined, a bottom-up strategy being followed by a top-down strategy (see figure 1). This kind of interaction occurs often in spontaneous writing where ideas lead to the recognition of a topical category, which in turn leads to the generation of new data of the same kind. Hence ideas or messages may have to be dropped. Not having enough conceptual material, the author may decide either no to mention a given fragment, to put it in a footnote, or to continue searching for additional material.

In the remainder of this paper, we will present a small prototype trying to emulate the first strategy : discovery of structure in data (messages). Yet before doing so, we would like to spell out in more detail some of the assumptions underlying our work and show how they relate to what is known about the natural writing process.

3. Assumptions concerning the writing process and the building of the intended tool

As mentioned already, authors tend to use different strategies when writing: they

⁵ The following two messages [(a) get married (x), (b) become pregnant (x)] could be considered as a *cause, consequence* or as a natural *sequence*.

⁶ What we mean by topic is the following. Suppose you were to write the following « foxes hide underground ». In this case a reader may conclude that you try convey something concerning the foxes' 'habits' (hide) or 'habitat' (underground).

start from topics or goals (top-down), from initially unrelated data or ideas (bottom-up), or they use mixed strategy, i.e. they use both.

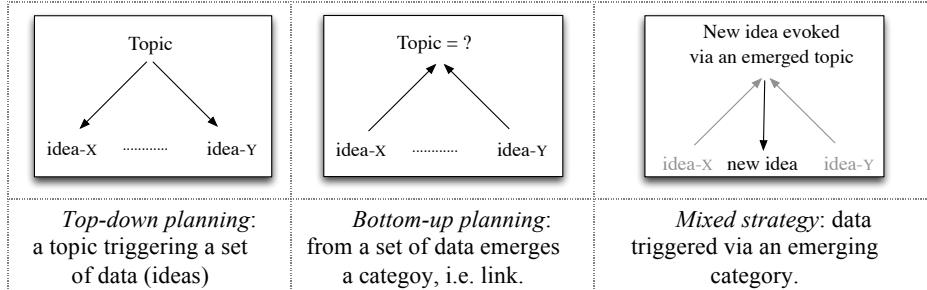


Fig. 1: Discourse planning top-down, bottom-up or both?

Starting from a goal they seek relevant content (messages), organize it according to topical and rhetorical criteria, translate it into language and revise it. This is known as top-down planning. Note that revision can take place at any stage, and that during the initial stage of conceptual planning (brainstorming) little filtering takes place. Authors are mainly keen on potentially interesting ideas. (Incidentally, this is why the term 'brainstorming' better captures the reality of this situation than 'idea planning'.) It is only at the next step that contents are thoroughly examined. This may lead to a modification of the message base : some ideas will be dropped, others added. The result of this is generally a so called outline or textplan, i.e. a definition of what is to be expressed (since it hasn't been so far) when, i.e. in what order.

Another strategy consists in going the opposite way. Starting from the ideas coming spontaneously to the authors' mind (bottom-up planning), she will try to group them into sets (topical trees) and to link these clusters. In this kind of bottom-up planning, the structure or topic emerges from the data. These topics may act as seeds, triggering eventually additional material (mixed strategy). Bottom-up planning is a very difficult problem (even for people). Yet this is the one we are interested in. Remains the question, on the basis of what knowledge writers know which ideas cohere, that is, what goes with what and in what specific way?

Suppose you have an assignment asking you to write a small document about foxes and their similarities and differences compared to wolves or coyotes, two animals with which they are sometimes confused. This might trigger search for information concerning 'foxes', possibly yielding a set of messages like the one shown in figure 2a. For the time being it does not really matter where these ideas come from (author's brain, external resource, or else), what we are interested in here is to find an answer to the following questions : (a) How does the author group these messages or ideas into topical categories ? (b) How does he order them within each group ? (c) How does he link and name these chunks ? (d) How does he discover and name the relations between each sentence or chunk ?

We will focus here only on the first question (topical clustering), assuming that messages will be grouped if they have something in common, and assuming further that there are good chances that messages or message elements do indeed have something in common. The question is how to show this. Actually, this can be either hard or fairly trivial, as in the case of term identity (co-refence : messages [offer (x,

`dog1, y) & like_to_chase (dog1, milkman])`). Remains the question of how to reveal commonalities or links between data in the non-obvious cases. One can think of several methods.

For example, one could try to enrich input elements by adding information (features, attribute-values, etc.) coming from external knowledge sources: corpora (cooccurrence data, words associations), dictionaries (definitions), etc. Another method could consist in determining similarities between message elements (words). This is the one we have used, and we will explain it in more depth here below (section 4). Once such a method has been applied, we should be able to cluster messages (since they have something in common) by category, even though we may not be able to name it. The name is implicit and requires other methods for revealing it.

The result of this will be one or several topic trees, grouping (ideally) all inputs. While different trees may achieve different rhetorical goals (the focus being different), all of them ensure coherent discourse. The effect of these variances can probably only be judged by a human user, who shall pick the one fitting best his or her needs. While our program will not be able to achieve this goal, that is, build a structure that conceptually and rhetorically matches the authors' goals, it should nevertheless be able to help the user perceive conceptual coherence, hence allow him to create a structure (topic tree) where all messages cohere, something that not all grown up human beings are able to do.

One other point concerning goals and bottom-up planning. Goals can be of various sorts. They can be coarse grained ("Convince your father to lend you his car") or more fine-grained, relating to a specific topic: describe an animal and show how it differs from another one with which it is often confused (alligator-crocodile; fox-coyote/wolve). Messages may also feed back and clarify goals or generate new goals. This cyclic process between top-down and bottom-up (emergence) processing is a very frequent case in human writing. We will focus here only on the latter, confining ourselves to very simple cases (2-place predicates) and assuming that there are common elements between the different messages. Of course, in reality this is not always the case —(Be careful, the road may be dangerous. They've just announced a typhoon.),— but since such cases require a different approach —(inferencing: causes can be conceived as the systematic correlation between two events or a state and an event),— we will not deal with them here.

4. Methodology

We present in this section a description of the method used to allow for the kind of grouping mentioned here above. Concerning the method one could have considered the following strategy: take a set of well written texts of a specific type (description), extract its sentences, normalize and scramble them and have the computer try to reorganize them to produce a coherent whole, matching as good as possible the initial document (gold standard). Having thought about this strategy too late, we used a different approach (see here below).

Messages can be organized on various dimensions and according to various viewpoints : *conceptual* relations (taxonomic, i.e. set inclusion, causal, temporal,...),

rhetorical relations (concession, disagreement), etc. We will focus here only on the former, assuming that messages can at least to some extent be organized via the semantics⁷ of their respective constituent elements.

Put differently, in order to reveal the relative proximity or relation between a set of messages we may consider the similarity of some of their constituent elements. Summing similarity values is a typical component of a vector-space model (<http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html>) and has been well described in books by D. Widdow⁸ and Manning, Raghavan and Schütze.⁹ Concerning ‘similarity’ one needs to be careful though, as the words’ similarity does not guarantee relatedness, even it may be one of its preconditions. Indeed, many researchers have used this feature for sentence similarity detection, but most of them based their analysis on the surface form which may lead to erroneous results, because similar meanings can be expressed via very different forms (for example: ‘use for’ vs. ‘instrument’, ‘her’). Likewise a given form or linguistic resource, say, the possessive adjective may encode very different meanings. Compare, —his car vs. his father vs. his toe,— which express quite different relations : ownership, family relationship, inalienable part of the human body.

What we present here is very preliminary work. Hence, our method is designed to address only very simple cases, two-place predicates, i.e. sentences composed of two nouns (a subject and an object) and a (linking) predicate. Given a set of this kind of inputs our program determines their proximity regardless of their surface forms. The sentences will be clustered on the basis of semantic similarity between the constituent words. This yields a tree whose nodes are categories (whose type should ideally be expressed explicitly, for example : food, color, ...) and whose leaves are the messages or propositions given as input. In the following sections we will explain in more detail our approach taking the inputs shown in figure 2a to illustrate our purpose.

- | | |
|--|--|
| 1° resemble (foxes, dog)
2° live_in (foxes, woods)
3° steal (foxes, chicken)
4° be (foxes, red)
5° eat (foxes, fruits)
6° hide (foxes, underground)
7° eat (foxes, eggs) | <pre> graph TD Foxes[Foxes] --> appearance[appearance] Foxes --> livingPlace[living place] Foxes --> habits[habits] appearance --> 1[1] appearance --> 4[4] livingPlace --> 2[2] livingPlace --> 6[6] habits --> 3[3] habits --> eatingHabits[eating habits] eatingHabits --> 5[5] eatingHabits --> 7[7] </pre> |
|--|--|

Fig. 2a Conceptual input (messages)

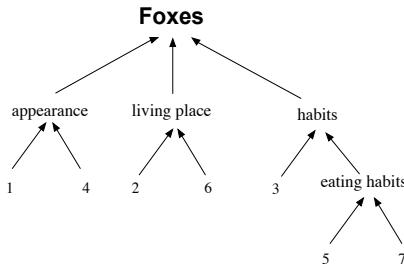


Fig. 2b Clustered output (topic tree)

The goal is to cluster this set of messages by topics. Indeed, {1, 4} address

⁷ Of course the term of semantics can mean many things (shared elements between a set of words, associations, etc.), and which of these uses an author is referring to needs to be made explicit.

⁸ <http://www.puttypeg.net/book/>

⁹ <http://nlp.stanford.edu/IR-book/pdf/06vect.pdf>

physical features (appearance), {2, 6} provide spatial information, the place where foxes live or hide (*habitat*), while {3, 5, 7} deal with their *habits*. This last category can be split into subtopics, in our case, 'theft' {3} and 'consumption' {5, 7}. The result of this analysis can be displayed in the form of a tree.¹⁰

In order to achieve this result we have defined an algorithm carrying out the steps referred to in table 2. We will describe and explain them in more depth in the following sections. Note that what we called messages here above, is now called sentence which is processed by a parser.

Table 2. Mains steps for topic clustering

- | |
|--|
| <ol style="list-style-type: none"> 1. Determine the role of words, i.e. perform a syntactic analysis; 2. Find potential seed-words; 3. Align words playing the same role in different sentences; 4. Determine the semantic proximity between the aligned words; 5. Determine the similarity between sentences; 6. Group sentences according to their semantic affinity (similarity). |
|--|

4.1 Identification of the syntactic structure

The goal of this step is to identify the dependency structure of the sentence. This information will be used later on (a) to identify the semantic seeds (see section 4.2), (b) to align words playing a similar role and (c) to identify the role of the different elements of the underlying proposition, i.e. the respective predicate, subject or object. To obtain this information we used the Stanford parser.¹¹ For example, the input 'Foxes eat fruits' would yield the following output:

Tagging

Foxes/NNS **eat**/VBP **fruits**/NNS ./ .

Dependencies

N_{subj} (eat-2, foxes-1)
D_{obj} (eat-2, fruits-3)

Of all these outputs we are concerned here only with N_{subj} and D_{obj} in order to determine the main elements of the message: the subject, object and the main verb, or, in propositional terms [predicate (argument₁, argument₂)]. Next, we used the similarity of the parts (words) to determine the similarity of the wholes (sentences).

4.2 Identification of the semantic seed-words

As mentionned already, in order to reveal the proximity or potential relation between two or more sentences one can try to identify the similarity between the respective constituent words. One needs to be careful though. If one does this by taking into

¹⁰ Note, that generally one can come up with more than one tree, any set of data allowing for multiple analyses. Much depends on the point of view.

¹¹ <http://nlp.stanford.edu/software/lex-parser.shtml>

account only the similarity values of the respective (pair of) words one may bias the analysis and get incorrect results.

There are several problems at stake. For example, the number of identical words does not necessarily imply relatedness or similarity. Actually, two sentences may be composed of exactly the same words, and still mean quite different things, compare: "Women without their men are helpless" vs. "Men without their women are helpless". Given the fact that such cases are quite frequent in natural language we decided not to rely on (all) the words occurring in a sentence, or to use a "bag of words" approach (sentence without stop words). We prefered to rely only on specific words, called seeds, to compare the similarity of different sentences. We consider seed words to be elements conveying the core meaning of a sentence. For example, for the two sentences here above we could get the following seeds : (a) without (man, women); (b) without (women, men), which reveal quite readily their difference.

Our idea of choosing seed words seems all the more justified as different kind of words (lexical categories) have different statuses: some words conveying more vital information than others. Nouns and verbs are generally more important than adjectives and adverbs, and each one of them conveys normally more vital information than any of the other parts of speech.¹² We assumed here that the core information of our sentences is presented via the nouns (playing different roles : subject, object) and the verb linking them (predicate). We assumed further that dependency information was necessary in order to be able to carry out the next steps. Compare, "Foxes hide underground" vs. "foxes hide *their prey* underground". A 'bag of word'-method or a simple surface analysis would not do, as neither of them reveals the fact that the object of hiding ('fox' vs. 'prey') is different in each sentence, a fact that needs to be made explicit.

To avoid this problem we used the dependency information produced by the parser which allowed us to determine the role of the nouns (*deep-subject*, *deep-object*) and the predicate (verb) linking the two. For example, this reveals the fact that the following two sentences are somehow connected: '*Foxes eat eggs*' and '*Foxes eat fruits*'. In both cases the concept 'fox' is connected to some object ('egg' vs. 'fruit') via some predicate, the verb 'eat'. The core of these two sentences is identical. Both of them tell us something about the foxes' diet or eating habits (egg, fruits). Note, that while this method does not reveal the nature of the link (diet, food), it does suggest that there is some kind of link (both sentences talk about the very same topic: food). Hence, syntactic information (part of speech, dependency structure) is precious as it allows us to identify potential seed words which will be useful for subsequent operations.

4.3 Word alignment

In order to compare sentences in terms of similarity, we need not only a method for doing so, but we need also the data to be in a comparable form. Hence we need the

¹² Note that we do not consider 'connectors' (yet, despite, because) here, as they are not known at this stage.

input to be given in a standardized form or we need to carry out some normalization. This latter can be accomplished to some extent via a dependency parser which reveals the roles played by different words. We can now align the words of the various sentences and compare those playing the same semantic role.

Word alignment consists not simply in finding identical words in different sentences, but rather in finding and aligning words playing the same role in these sentences. This means in our case that we have to compare, say, the subject of one sentence with the subject of another, and do the same for the other syntactic categories or semantic roles (verbs, deep-objects, ...). To allow for this we rely again on the dependency information produced by the parser (section 4.1). Note, that our example showed only surface relations (subject, object, etc.), while ideally we need information in terms of deep-case roles : agent, beneficiary, etc. (Fillmore, 1968). Applied to our examples, "Foxes eat fruit" and "Foxes eat eggs", it is clear that "fruit" can be aligned with "eggs", since both nouns play the same role.

Note that we also need to be able to detect synonyms or semantic equivalences : 'instrument \equiv is used for'; "resemble \equiv be alike", "for example \equiv somehow \equiv like", etc. These words are very useful and could be used as topic-signatures (Lin & Hovy, 2000), hence seed words. Note that such information is obtained indirectly in our approach via the *vector space model* which is briefly described in the next section.

4.4 Determination of the similarity values of the aligned words

While there are various ways to detect links between sentences or words (for example, shared features or associations), two obvious ones are coreference and class-membership. See our example in figure 2a, where the two sentences —("Foxes eat eggs"; "Foxes eat fruits")— have an identical referent, the actor 'fox', and two different instances of the same class, the generic element 'food'.

As mentioned already, in order to compare sentences in terms of their meaning the compared structures must have a common format. Similarity of meaning supposes of course that we are able to extract somehow the meaning of the analyzed objects (sentences, words). Yet word meanings depend on the context in which a word occurs. Words occurring in similar contexts tend to have similar meanings. This idea, known as the 'distributional hypothesis' has been proposed by various scholars (Harris, Firth, Wittgenstein, 1922). For surveys, see (Sahlgren, 2008, Dagan et al. 1999), or (http://en.wikipedia.org/wiki/Distributional_semantics).

Since we try to capture meaning via word similarity, the question of how to operationalize this notion arises. One way of doing so is to create a vector space composed of the target word and its neighbors (Lund and Burgess, 1996). The meaning of a word is represented as a vector based on the n-gram value of all co-occurring words. In the following two sentences, —"Foxes eat eggs"; "Foxes eat fruits",— we have 4 distinct tokens or words: foxes, eat, fruits and eggs. Hence we constructed a vector for each one of them by considering their co-occurrences in COHA (Corpus of Historical American English), a part of speech tagged n-gram corpus of 400 million words (Mark, 2011). This allowed us to apply the vector-space model in order to compute the degree of similarity between a set of words. To this end we

computed the distance (cosine) of the respective vectors. Let us suppose that there are only 4 words co-occurring with ‘fruit’ and ‘egg’ (“juice, vitamin, price, eat” and “chicken, protein, eat and oval”), then the vector for fruit would be “juice, vitamin, price, eat” and the vector for egg would be: chicken, protein, eat, oval.

Note that we will also count the frequency of the co-occurrence. To compute the similarity between two vectors we computed the cosine of their angle. Hence, we constructed such vectors for all major words of our sentences. For the example here above we have four vectors, one for each of the words occurring in both sentences : fox, eat, fruit and eggs. Note that the words in the vectors are replaced by their weight, that is a numerical value representing the meaning of the respective concepts. For instance, for the *fruit* vector, all of the following words, “juice, vitamin, price, eat” are replaced by a numerical value (weight). For details see step 1 here below. We have used this method already for another task, the automatic extraction of part-whole relations (Tesfaye and Zock, 2012). Since then we have extended it to allow for the computation of similarity between words. The method consists basically in two operations : creating a vector for all words and identifying the similarity of the aligned words.

Step 1 : Creation of a vector for all words based on co-occurrence information

The co-occurrence information is gleaned from COHA. The vector is built on the basis of a word's co-occurrence within a defined window (phrase, sentence, and paragraph). Since different words make different contributions we assign weights to reflect their relative contribution in terms of relevance allowing to determine the meaning of a sentence or word. Hence meanings are expressed via a weight which may depend on the context of a given term, a factor which needs to be taken into account. In order to determine the weight (W) we use the percentage of the term's co-occurrence frequency (TCF) with respect to a given concept out of the **total number of co-occurrences** (TotNbc) of the term with any other term in the corpus. For example, in order to determine the weight of the term *egg* (one of the words in the ‘chicken vector’) in defining the meanings of *egg* we count the total frequency of the co-occurrences of *chicken* with *egg* and divide then this result by the total number of co-occurrences of the term *chicken* with any other term in the corpus. We did the same for all relevant terms. The above described operations can be captured via the following formula which is used to determine the weight (W) of a given word for defining the meaning of another co-occurring term:

$$W = TCF-yz / \text{TotNbc-xy}, \text{ where}$$

TCF-*yz* is the frequency, i.e. the number of times *y* co-occurs with *z* ;

TotNbc-*xy* is the total number of times *x* co-occurs with *y*.

To build a vector for a given concept we use the weighted value of all co-occurring words. Hence, we calculated the relative weight of each word of the vectors in defining the meaning of the term for which the vector was built.

Step 2 : Identification of the cosine similarity between vectors of the aligned words

The similarities between words are computed on the basis of the cosine of the words' vectors. Note that the similarity value is calculated only for the aligned words.

One could object that we simply rely on a TF*idf approach. This is true only to

some extent. While there are some similarities, our approach is different in at least two respects. We are interested in co-occurrence frequencies rather than in term or document frequencies. Not needing the terms' occurrence/frequency, we do not use at all the inverse document frequency. Hence, if the TF*idf approach yields a term-document matrix, ours yields a term-term matrix.

4.5 Determination of the similarity between sentences

The meaning of a sentence can (at least to some extent) be obtained via the combined meanings of the constituent elements, words. Having identified in 4.4 the similarity values between the aligned words, we build now a matrix showing their respective similarity values. The rows and columns of the vectors are built on the basis of co-occurring words and the cells contain their similarity values. In order to identify the similarity between two sentences we add the similarity values of the component words and compute then the average to derive a single similarity value between the sentences. Hence, in order to identify the degree of similarity between a pair of sentences we sum up the respective similarity values of their subjects, verbs and objects, to divide then the result by 3 in order to get the average.

Table 3. Sample word similarity matrix of co-occurrences

	resemble	eat	are	live	steal	hide
resemble						
eat	0.293					
are	0.550	0.152				
live	0.210	0.365	0.139			
steal	0.428	0.392	0.210	0.306		
hide	0.527	0.430	0.240	0.631	0.450	

With respect to our fox example (figure 2) all inputs apart from sentence 3, are clustered this way. Sentence 3 is clustered with the sentences 2 and 6 according to the algorithm presented in the next section.

4.6 Sentence clustering based on their similarity values

As mentioned already, our strategy consists in the creation of a tree based on the similarity values of the sentences given as input. Sentences are clustered in three steps on the basis of the similarity value of the subjects, the verb and the objects. Accordingly, sentences talking about different topics, say 'foxes' and 'fruits', are placed in different clusters. At the next cycle each group is further clustered depending on the topic (habit, physical appearance, etc.) which may be signalled via the verb or the object. Here below is the clustering algorithm.

Table 4. *The clustering algorithm:*

- | |
|---|
| <ol style="list-style-type: none"> 1. Take any sentence of the considered pool of inputs and search for another one whose topic similarity (i.e. value of the subject) is closer to the target than any of its competitors to form a cluster. Similarity values are obtained via the word-similarity-matrix (see step 4.5 and table-3). 2. Continue to cluster the sentences of the groups obtained so far by using the similarity values of the verb linking the subject and the object. 3. Perform the same operation as in step 2 based on the similarity of the objects. 4. Repeat steps 1 to 3 until all the sentences, or the greatest possible number of sentences are clustered. 5. Create a link between the clusters based on the respective similarity values of the verb and the object. |
|---|

5. Experiment and Evaluation

In order to test our system we used a text collection of 28 sentences. The test set contains 4 groups of sentences talking respectively about 'foxes', 'fruits', and 'cars'. The last set, called rag bag, is composed of topically unrelated sentences. It is used only for control purposes.

Table 4. Our test sentences

1. Topic ₁ Fox	2. Topic ₂ Fruits
<ol style="list-style-type: none"> 1. Foxes resemble dogs. 2. Foxes live in the woods. 3. Foxes steal chicken. 4. Foxes are red. 5. Foxes eat fruits. 6. Foxes hide underground. 7. Foxes eat eggs. 	<ol style="list-style-type: none"> 8. An apple a day keeps the doctor away. 9. Apples are expensive this year. 10. Oranges are rich in vitamin C. 11. The kiwi fruit has a soft texture. 12. Grapes can be eaten raw. 13. Grapes can be used for making wine. 14. The strawberries are delicious.
3. Topic ₃ Cars	4. Topic ₄ ragbag
<ol style="list-style-type: none"> 15. A car is a wheeled motor vehicle. 16. Cars are used for transporting passengers. 17. Cars are mainly designed to carry people. 18. The first racing cars amazed the automobile world. 19. Cars typically have four wheels. 20. Cars are normally designed to run on roads. 21. Cars also carry their own engine. 	<ol style="list-style-type: none"> 22. Olive oil is a fat obtained from olives. 23. Playboys usually have a lot of money. 24. A finger is a limb of the human body. 25. Apple is the name of a software company. 26. Eau sauvage is a famous perfume. 27. Wine is an alcoholic beverage made from fermented grapes or other fruits. 28. IBM is an American corporation manufacturing computer hardware.

The system's task is now to integrate as many messages as possible. This will yield a tree containing as many branches as there are topics, in our case four. At the next cycle the system will try to create subcategories, that is, branches are further divided, or, viewed in the opposite direction, messages or clustered in more specific categories (habit, living place, etc. in the fox group). Whether this is feasible depends of course on the message elements. Since the function of the control group (the set of

topically unrelated sentences) is only to check the system's accuracy, none of its sentences should appear in any other group than the 'control group'.

Once this clustering is done, we can determine the system's performance by counting the number of sentences assigned properly in the tree. For the evaluation we used the classical metrics, defining *precision*, *recall*, and *F-measure* in the following way: *recall* (Number of sentences correctly assigned to the valid cluster/ Total number of sentences); *precision* (Number of sentences correctly assigned to the valid cluster/ Number of sentences clustered); *f-measure* ($2 / [(1/\text{precision}) + (1/\text{recall})]$).

We obtained the following results: [22 out of the 28](#) sentences are placed in the correct cluster, 6 occur in the wrong place. For example, the sentence 25 and 27 are both placed in the fruit cluster (topic 2), while they should not. This is due to the fact that our method does not take into account the semantics of the string 'apple' which can refer both to the fruit or the computer manufacturing company located at Cupertino. The same holds for 'wine' which can be both an alcoholic drink or a fruit.

We have also evaluated our system by further clustering the sentences within each topic based on their similarity. All the sentences of topic 1 are clustered correctly, the sentences 1-4, 2-6 and 5-7 forming three clusters. Sentence 3 is more closely related to the clusters containing 5 and 7 than to any other cluster. All the sentences in group 2 are clustered and two of them (10 and 14) are grouped in a more specific category. However, we do have some problems as some items appear in the wrong place. 25 and 27 are intruders, dealing with a different topic. Hence they should not be in this category. The same holds for sentence 9 which is put in the (sub)group of the sentences 10 and 14. This is clearly wrong, since it is about Apple the *computer company* and not about 'apple', the *fruit*. On the other hand, sentence 11 should be there, that is in the same cluster as 10 and 14. Yet it is not. It is placed elsewhere, standing like a loner, which of course is a mistake. In topic 3 (cars) all the sentences are clustered correctly, and two sentences (15, 16) are placed in a more specific category. However, the system failed to cluster Sentence 19 and 21 together.

Given the above described results, the system has a recall, precision and f-measure [of 78.6%](#) if we consider only the sentences placed in the correct position in the tree. It is interesting to note that some of the sentences placed in the wrong category have a similarity value fairly close to the one of the correct cluster. Actually, most of them have the second highest similarity value. Note also that at this point, the clusters, i.e. the nodes of the tree, are not labeled. Whether this could be achieved via topic signatures (Lin & Hovy, 2000) remains to be shown and is clearly work for the future.

6. Outlook and conclusion

We tried to present a new approach concerning a component of text production. While most work on computerized language production is fully automatic, our approach is interactive. Given some input by the user (messages) the system helps her to structure the data in order to produce a coherent output. Since this is a quite complex task we have started with a very simple set of data. Nevertheless, we have tried to reveal not only the obvious (co-references), but also the more hidden and distributed information. Our next steps will consist in dealing with more complex

cases, including other conceptual relations and the problem of labeling the topical categories. We would like to push further the idea of the information associated to seed words. Any set of data is likely to be grouped according to various criteria or view points, and it remains to be shown where the information (additional data) allowing for this kind of grouping comes from. While not being trivial, this will certainly be very interesting. Last, but not least, we would like to take a closer look at some of the work devoted to sentence ordering (Wang, 2006), work we were not aware of at the very moment of writing this paper.

References

5. Alamargot, D., & Chanquoy, L. (2001). Through the models of writing. Dordrecht: Kluwer.
6. Andriessen J., deSmedt K. & M. Zock. (1996) Discourse Planning: Empirical Research and Computer Models. In T. Dijkstra & K. de Smedt (Eds). *Computational Psycholinguistics: AI and Connectionist Models of Human Language processing*, Taylor & Francis, London, pp. 247-278
7. Bateman, J. & Zock, M. (2003). Natural language generation. In R. Mitkov (Ed.), *Handbook of computational linguistics*. London: Oxford University Press pp. 284-304
8. Beaugrande, R. de (1984). *Text production: Towards a science of composition*. Norwood, NJ: Ablex. (http://www.beaugrande.com/text_production.htm)
9. Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.
10. Dagan, I., Lee, L. and F. Pereira. (1999). Similarity-based models of cooccurrence probabilities, Machine Learning, Vol. 34(1-3) special issue on Natural Language Learning, pp. 43-69
11. de Marneffe, M.C. and Manning, C. (2008). Stanford typed dependencies manual
12. Fillmore, C. (1968). The Case for Case. In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88
13. Flower, L. & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive Processes in writing* (pp. 39–58). Hillsdale, NJ: Erlbaum.
14. Hovy, E. H. (1991). Approaches to the planning of coherent text. In C. L. Paris, W. R. Swartout, & W. C. Mann (Eds.), *Natural language generation in artificial intelligence and computational linguistics* (pp. 83–102). Boston: Kluwer Academic.
15. Iyer, L.R., Doboli, S., Minai, A.A., Brown, V.R., Levine and D.S., Paulus (2009). Neural dynamics of idea generation and the effects of priming. *Neural Networks*, Volume 22, Issues 5–6, pp 674-686.
16. Kellogg, R. (1999). *Psychology of writing*. New York: Oxford University Press.
17. Levy C. M. & S. Ransdell (Eds.), (1996). *The science of writing. Theories, methods, individual differences and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.
18. Lin, C.-Y. & E.H. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. Proceedings of the COLING Conference. Saarbrücken

19. Lund, K. and Burgess, C. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2), 203-208.
20. Mann, W. C., & Thompson, S. A (1987) Rhetorical Structure Theory: A Theory of Text Organization, in: Polanyi, L. (ed.) *The Structure of Discourse*, Norwood, N.J.: Ablex
21. Marcu, D. (1997). From Local to Global Coherence: A Bottom-up Approach to Text Planning. *The Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 629-635, Providence, Rhode Island, July 1997.
22. Mark, D. (2011). N-grams and word frequency data from the Corpus of Historical American English (COHA).
23. Matsuhashi, A. (1987). Revising the plan and altering the text. In A. Matsuhashi (Ed.), *Writing in real time* (pp. 197–223). Norwood, NJ: Ablex.
24. McKeown, K. R. (1985). Text generation: Using discourse strategies and focus constraints to generate natural language text. Cambridge University Press.
25. Moore, J. D. and C.L. Paris. (1993). Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics*, 19(4).
26. Olive, T., & Levy, C. M (Vol. Eds.) (2001). *Studies in writing: Vol. 10. Contemporary tools and techniques for studying writing*. Dordrecht: Kluwer.
27. Reichenberger, K., Rondhuis, K. J., Kleinz, J., & Bateman, J. A. (1995). Communicative Goal-Driven NL Generation and Data-driven Graphics Generation: an architectural synthesis for multimedia page generation. 9th International Workshop on Natural Language Generation. Niagara on the Lake, Ontario, Canada
28. Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge University Press.
29. Rijlaarsdam, G., & Van den Bergh, H. (1996). The dynamics of composing – An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 107–125). Hillsdale, NJ: Erlbaum.
30. Rijlaarsdam, G., H. van den Bergh, & M. Couzijn (Eds.), Effective teaching and learning of writing. Current trends in research (pp. 253–273). Amsterdam: Amsterdam University Press.
31. Sahlgren, M. (2008). "The Distributional Hypothesis". *Rivista di Linguistica* 20 (1): 33–53.
32. Simonin N. (1988). An Approach for Creating Structured Text, in Zock & Sabah (Eds.) *Advances in Natural Language Generation: an Interdisciplinary Perspective*, co-edition Pinter, London, Ablex, Norwood, N.J, Vol. 1, 146-160
33. Torrance, M., & Jeffery, G. (Vol. Eds.) (1999). *Studies in writing: Vol. 3. The cognitive demands of writing*. Amsterdam: Amsterdam University Press.
34. Wang, Y. W. (2006) Sentence ordering for multi-document summarization in response to multiple queries, Simon Fraser university.
35. Zock, M. & Tesfaye, D. (2012). Automatic index creation to support navigation in lexical graphs encoding part_of relation. Cogalex-III, Coling workshop, Mumbai, India, pp. 33-52

Semantic Types, Lexical Sorts and Classifiers

Bruno Mery & Christian Retoré

Université de Bordeaux, IRIT-CNRS, LABRI-CNRS*

Abstract. We propose a cognitively and linguistically motivated set of sorts for lexical semantics in a compositional setting: the classifiers in languages that do have such pronouns. These sorts are needed to include lexical considerations in a semantical analyser such as Boxer or Grail. Indeed, all proposed lexical extensions of usual Montague semantics to model restriction of selection, felicitous and infelicitous copredication require a rich and refined type system whose base types are the lexical sorts, the basis of the many-sorted logic in which semantical representations of sentences are stated. However, none of those approaches define precisely the actual base types or sorts to be used in the lexicon.

In this article, we shall discuss some of the options commonly adopted by researchers in formal lexical semantics, and defend the view that classifiers in the languages which have such pronouns are an appealing solution, both linguistically and cognitively motivated.

Introduction

One of the most difficult aspect of the automated processing of human language is the phenomenon of *polysemy*, the ability for words to be used for different meanings in different contexts. Relatively recent studies, such as Pustejovsky (1995), have held the view that polysemy is a feature that enables creativity in linguistic acts, and that the meaning of words might be deduced by the application of generative mechanisms from their contexts, via processes refining semantical composition. Instead of thinking of all words denoting individual objects as sharing the same semantic types (of *entities*), advanced lexical semantics could class them along *lexical sorts* according to their contextual behaviour, and a process of type-checking could infer the correct meaning from any combination of predicate and object.

For the computational linguist, the problem of lexical semantics thus becomes twofold:

1. How does the semantic composition have to be modified ?
2. How should the base types, the lexical sorts, be defined ?

The first point has been the subject of many different and related proposals, including the authors' own framework. This paper is concerned with the second part of the problem, and propose a linguistically-motivated solution.

* This research has benefitted from grants and inputs by ANR Polynomie, Project Itipy (Région Aquitaine), and CoLAn. We are indebted to the reviewers for their input.

1 Including Lexical Considerations into Syntactical and Semantical Parsers

There are some wide coverage analysers that produce complete semantic analyses expressed as logical formulae, like Boxer by Johan Bos (English) and Grail by Richard Moot (spoken Dutch, written French). In both cases, the grammar, that is, a lexicon mapping each word to several semantic categories, is statistically acquired from annotated corpora. It thus has up to one hundred categories per word, hence the parser first computes the most likely sequences of categories and parse the n best. See Bos (2008); Moot (2010b).

In order to compute semantic representation, both use *categorial grammars* (multimodal or combinatory CG) and this is not a coincidence. Indeed, categorial grammars allow easy transformation from syntactic categories to semantic types and from syntactic analyses to semantic analyses.

Both analysers, as well as many other practical and theoretical frameworks, rely on principles of semantical composition along with the tradition of Montague Grammar, specified in Montague (1974) and refined many times since.

Montague Grammar assumes that words have a correspondence with terms of the simply-typed λ -calculus, with applications and abstractions given by the syntactic structure of the utterance, sentence or discourse. Those terms are constructed in a type system that use two types, t for truth-valued formulae, and e for entities. In that way, all single entities share the same sort, e .

Some frameworks and analysers also add the base type s , for indices of possible worlds, and the abstract sort v for events. However, linguistic entities still share the single sort e .

Considerations of lexical semantics provide compelling arguments for different base types. Specifically, the single sort e for entities can be split in several sorts, refining the type system. Consider:

- (1) a. The hound barked.
 b. *The vase barked.
 c. ?The sergeant barked.

Restrictions of selection (what, according to dictionaries, noun phrases can be object to specific verbs) dictate that (1a) is correct, (1b) is difficult to admit without a clear context, and (1c) is acceptable, but indicates a common metaphorical usage of *bark*, implying that the person referred to has certain dog-like qualities.

If the distinction is made by an analyser at the stage of semantic composition, using a singular sort e for all entities does not allow to distinguish between the syntactically similar sentences. Using different sorts for animate and inanimate entities (as commonly used in dictionary definitions) will licence (1a) and reject (1b)¹.

¹ This does not imply that sentences such as (1b) should never receive any semantic analysis. There are some contexts (such as fairy tales or fantasy) that can give meaning

With additional distinctions between, in this case, dogs and humans, and a flexible typing system that detects type clashes and licence certain modification to the typing of lexical entities, the metaphorical usage of the verb in (1c) can be detected and identified.

Lexical semantics also helps with the common problem of *word sense disambiguation*. A common use of words pertaining to organisations such as banks, schools, or newspapers is to represent some unnamed person that is responsible for the conduct of that organisation. Consider:

- (2) The bank has covered for the extra expenses.

(2) means that someone has taken the liberty mentioned. Distinguishing between the normal use of the word (as an organisation) and this specific use (as an agent within that organisation) is only possible if the semantic system has a mean to set them apart, and a way to accomplish this is having *Organisations* and *Agents* being two different sorts of entities in the type system.

Pustejovsky (1995) and other related publications present a broad linguistic framework encompassing those issues and many others related to polysemy and the creative use of words in context. It relies on a rich lexicon with several layers of information, and a many-sorted type system that help distinguish the different sorts of entities using an ontological hierarchy founded on linguistic principles.

The main issue is that this rich ontological type system has not been detailed, and is very much not trivial to construct, let alone that the general composition rules are missing from the original formulation.

1.1 Rich Types and Lexical Modifiers

The authors have defined a system for the inclusion of lexical semantics data (see Mery et al. (2007), Bassac et al. (2010), Mery (2011) and Mery & Retoré (2013)), and some of those results have been implemented in a semantics analyser. Instead of the single sort *e*, we make use of many different sorts for entities that can distinguish between different linguistic behaviours.

Formally, this framework uses a version of the type logic with n sorts, TY_n , detailed in Muskens (1996b). Without detailing functionalities outside the scope of this contribution, those n sorts are used to type the different classes of entities of the lexicon. When a type clash is detected, the analyser searches for available *modifiers* provided by the logical terms that would allow the analysis to proceed, and makes a note of the lexical operations used in order to compute the actual meaning of the sentence. For instance, the following sentences refer to different facets of the entity *bank* (all pertaining to the finance-related concept), identifiable by the predicates used:

to such sentences, and strategies to deal with those and compute a correct semantics with the same compositional analysis. In order to recognise that such a special treatment is needed, however, the system still needs to detect that the use is non-standard; it is as simple as detecting a type clash

- (3) a. The bank is closed today.
 b. The bank is at the next corner.
 c. The bank has gone mad.

(3a) refers to one of the most common use of the word, an *Organisation*, its base type. The type system maintains inferences for commonly used modifications, a very common is to refer to a physical location where the organisation is embodied, and thus the analyser would shift the type of the term to *Location* in (3b). In (3c), the predicate should apply to a person, and thus the type system would look for a way to associate a person to the organisation referred to.

Our framework makes use of abstraction over types (and second-order λ -calculus) in order to keep track of the lexical types involved, of constraints and modifications over those types. With hand-typed lexical entries and sorts defined over a restricted domain, this approach has been implemented and tested. However, we do not have a type system covering an entire language.

As an abridged example of the analyser, consider the sample lexical entry below:

Lexical item	Main λ -term	Modifiers
<i>Birmingham</i>	$birmingham^T$	$Id_T : T \rightarrow T$ $t_2 : T \rightarrow P$ $t_3 : T \rightarrow Pl$
<i>is_a_huge_place</i>	$huge_place : Pl \rightarrow \mathbf{t}$	
<i>voted</i>	$voted : P \rightarrow \mathbf{t}$	

where the base types are defined as follows:

T town

P people

Pl place

The sentence:

- (4) Birmingham is a huge place

results in a type mismatch (the predicate is of type $Pl \rightarrow \mathbf{t}$, argument of type T)

$$huge_place^{Pl \rightarrow \mathbf{t}}(Birmingham^T))$$

The lexical modifier $t_3^{T \rightarrow Pl}$ that turns a town (T) into a place (Pl) is inserted, resulting in:

$$big_place^{Pl \rightarrow \mathbf{t}}(t_3^{T \rightarrow Pl} Birmingham^T))$$

Considering:

- (5) Birmingham is a huge place and voted (Labour).

In order to parse the co-predication correctly, we use a polymorphic conjunction $\&\Box$. After application and reduction, this yields the following predicate:

$$\Lambda \xi \lambda x^\xi \lambda f^{\xi \rightarrow \alpha} \lambda g^{\xi \rightarrow \beta} (\text{and}^{\mathbf{t} \rightarrow \mathbf{t}} \rightarrow \mathbf{t} (huge_place (f x))(voted (g x)))$$

Applying the argument of type T and the correct modifiers t_2 and t_3 , we finally obtain:

(and ($\text{huge_place}^{Pl \rightarrow t}(t_3^{T \rightarrow Pl} \text{Birmingham}^T)$)) ($\text{voted}^{Pl \rightarrow t}(t_2^{T \rightarrow P} \text{Birmingham}^T)$))

1.2 The Difference between our Proposal and related Formulations

There are several related proposals devoted to type-driven lexical disambiguation that share many characteristics, including works by Pustejovsky, Asher, Luo and Bekki, started in Pustejovsky & Asher (2000), elaborated in Asher & Pustejovsky (2005), extensively developed in Asher (2011) and subject of continuing work in Xue & Luo (2012) and Bekki & Asher (2012).

We are indebted to the authors of this proposal and many others. However, our formulation differs from the others in a significant way.

Ontological Types and Meaning Shifts : In Asher (2011) and other proposals, the base types are envisioned as an ontological hierarchy that derive a language-independent system of transfers of meaning. The different possible senses associated to a word are largely dependent on conceptual relations made available by its type.

Lexical-based Transformations : In our model, while base types distinguish between different sorts and drive the disambiguation process, the availability of transformations from a sort to another is defined at the lexical level, and depends on the language. It is thus possible to define idiosyncrasies and keep a closer rein on complex linguistic phenomena. This does not exclude to have some type-level transformations for practical purposes, specifically for the factorisation of common meaning shifts (e.g. transformations that apply to all vehicles also apply to cars).

2 Results on a restricted Domain

As observed by a reviewer, our model does not need a wide coverage generalist semantic lexicon to be tested , and we actually made some experiments for a particular question (in fulfilment of a regional project Itipy), the reconstruction of itineraries from a historical (XVII-XX century, mainly XIX) corpus of travel stories through the Pyrenees of 576.334 words. See Lefevre et al. (2012a,b) for details.

For such a task the grammar ought to be a wide coverage one, including a basic compositional semantics without sorts nor any lexical information. We do have such a grammar, which has been automatically extracted from annotated corpora: it is a wide coverage multimodal categorial grammar that is a lexicalised grammar with an easy interface with compositional semantics à la Montague.

In the absence of manually typed semantic information, the grammar only includes an automatically constructed semantic lexicon with semantic terms

that only depict the argument structure, e.g., *give* has $\lambda s^e \lambda o^e \lambda d^e . give(s, o, d)$ as its semantics. The actual implementation detailed in Moot (2010b,a) uses λ -DRSs of Discourse Representation Theory Kamp & Reyle (1993); Muskens (1996a) rather than plain λ -terms in order to handle discursive phenomena.

As the task is to provide a semantic representation the paths traversed or described by the authors, we focused on spatial and temporal semantics. Temporal semantics is handled by operators à la Verkuyl, that have little to do with lexical semantics, so we shall not speak about this in the present paper. But the semantics of space is modelled by the very framework described in the present paper.

As expected, the sorts or base types are easier to find for a specific domain or task. For space and motion verbs we obviously have two sorts, namely *paths* and *regions*, the later one being subdivided into *villages*, *mountains*, and larger areas like mountains chains. Paths did not need to be further divided, since by the time the stories in our corpus were written people only walked on paths (that could be called trails nowadays). Nowadays for the analysing travel stories one would possibly although consider motorways, roads, trails, etc.

The principal coercion we study in this setting for the analysis of itineraries is the phenomenon known as fictive motion Talmy (1999). One can say "*the path descends for two hours*". In order to interpret such a sentence, one needs to consider someone that would follow the path, although there might be no one actually following the path, and it is often difficult to tell apart whether the narrator does follow the path or not. Such constructions with verbs like "*descendre*, *entrer*, *serpenter*,..." are quite common in our corpus as examples below show:

- (6) (...) cette route qui monte sans cesse pendant deux lieues
 (...) this road which climbs incessantly for two miles
- (7) (...) où les routes de Lux et de Cauterets se séparent. Celle de Lux entre dans une gorge qui vous mène au fond d'un précipice et traverse le gave de Pau.
 (...) where the roads to Lux and to Pau branch off. The one to Lux enters a gorge which leads you to the bottom of a precipice and traverses the Gave de Pau.

Our syntactical and semantical parser successfully analyses such examples, by considering coercion that turn an immobile object like a road into an object of type *path* that can be followed. A coercion introduced by the motion verb that allow fictive motion, e.g. "*descendre*" (*descend*), construct a formula (a DRS) that says that if an individual follows the path then he will goes down. The formula introduces such an individual, bound by an existential quantifier, and it is part of discourse analysis to find out whether it is a virtual traveller or whether the character of the travel story actually followed the path. Moot et al. (2011a,b)

With a handwritten lexicon designed for a more precise analysis of spatial semantics, our framework worked successfully, i.e., automatically obtained the proper readings (and rejected the infelicitous ones when motion event are applied to improper spatial entities).

2.1 The Granularity of the Type System

The obstacle to our framework, and other related proposals, is thus the building of the system of sorts for entities. There is no real consensus on the criteria to be followed. We chose to dismiss the claims that such an endeavour is simply impossible, that compositional semantics should stick to the single-sort Montagovian ϵ , and that any refinements should wait a phase of pragmatics or interpretation left as an exercise to the reader, as made in very blunt terms by Fodor & Lepore (1998) and more reasonably by Blutner (2002), and refuted in Pustejovsky (1998) and Wilks (2001). We assume that a rich lexicon with a refined type system are helpful for a number of theoretical and practical applications.

However, in those cases, the type system is more often than not simply assumed. James Pustejovsky has described how it should behave in a number of details, in publications such as Pustejovsky (2001). It has never been detailed beyond the top level and some examples; as it was outlined, the system was a hierarchical ontology comprising most concepts expressed in natural language, with at least hundreds of nodes. The other proposals range between a dozen high-level sorts (*animated, physical, abstract...*) and every common noun of every language (Xue & Luo (2012)), and even every possible formula with a single free variable (as formulae are derived from types, that last definition is circular). Some others, such as Cooper (2007), propose using a record type system that does away neatly with the granularity problem, as record types are re-defined dynamically²; or even deliberately vague approaches, arguing that a definite answer to that question would be self-defeating.

2.2 Practical Issues with the Controversy

While leaving the issue open is philosophically appealing, as the possibility of a definition of an actual, single metalinguistic ontology contradicts existential principles, there is a very compelling reason to pursue the matter: providing an actual implementation of a compositional lexical semantic analysis. Partial implementations, including ours illustrated in section 2, exist, but without a comprehensive and well-defined type system, they are largely prototypal and rely on a few hand-written types. They do prove the viability of the analysis and the interest for word sense disambiguation, but they cannot provide a really useful analysis outside the scope of very specific domains, up to now. Large-scale generic NLP applications remain out of reach. Manual or semi-automated annotations are difficult, as they have either to be restricted to a very specific domain where it is possible to define base types comprehensively, or to be few in number and thus vague and error-prone. Choices have to be made, not in order to define the essence of lexical meaning, but simply to provide testable, falsifiable models and software that can be refined for actually useful applications.

² However, the inclusive definition of the records type system places it beyond classical type theory, which necessitates further adaptation in the logical framework.

This does not mean that a definite set of sorts can or should be devised once and for all, but a linguistically-motivated system, adaptable and mutable, would be an important step forward.

3 Type Granularity and the Classifier Systems

Sorts should represent the different classes of objects available to a competent speaker of the language. That two words of the same syntactic category have different sorts should mark a strong difference of semantic behaviour.

Our type system should be useable, with a computationally reasonable number of sorts. It should nevertheless be complex enough to model the lexical differences we are looking for.

In short, the set of sorts used as base types should be small in cardinality, with respect to the lexicon; large in the scope of lexical differences covered, if not complete; linguistically and cognitively motivated; adaptable, and immune to idiosyncrasy.

There have been many studies of some linguistic features that can prove interesting candidates for such a set, including grammatical attributes (gender, noun classes...) and meta-linguistic classes proposed by Goddard & Wierzbicka (2007). We have chosen to illustrate some of the properties of the classifier systems, a class of pronominal features common to several language families including many Asian languages and every Sign language.

3.1 The Case of the Classifier Systems

A large class of languages employ a certain category of syntactic items known as classifiers. They are used routinely for quantificational operations, most commonly for counting individuals and measuring mass nouns. Classifiers are also widely used in Sign Language (several variations) for analog purposes.

Classifiers are interesting, as they are used to denote categories of physical objects or abstract concepts and approximate a linguistic classification of entities. The fact that they arise spontaneously in different and wide-reaching language families, their variety and their coverage makes them good candidates for base types. Classifiers are often present in many Asian languages (Chinese, Japanese, Korean, Vietnamese, Malay, Burmese, Thai, Hmong, Bengali, Munda), in some Amerindian and West African languages and in all Sign Languages. They are almost absent for Indo-European languages; in English a trace of a classifier is "head" in the expression "forty heads of cattle" where one can thereafter use "head(s)" to refer to some of them.

They are used as pronouns for a class of nouns which is both linguistically and ontologically motivated. They differ from noun classes in the sense that they are much more classifiers (200–400) than noun classes used for flexion morphology and agreement (≤ 20). Several classifiers may be used for a single noun, depending on the relevant reading. Classifiers are especially developed

and refined for physical objects and can often stand alone with the meaning of a generic object of their class, and some nouns do not have a classifier: in such a case the noun itself may be used as a classifier.

The notions conveyed by classifiers differ somehow from language to language. For instance, in Chinese, classifiers can be used to count individuals, measures, both, or neither (see Li XuPing (2011) for details), the latter case being used to denote a similarity with the referred class. They are some linguistic and cultural idiosyncrasies. However, the main features of the system are common to all languages.

3.2 Classifiers in French Sign Language

Classifiers in sign languages (see Zwitserlood (2012)) are used in the language as distinct pronouns each of them applying to cognitively related nouns, in the sense that their shape evoke their visual shape or the way these entities are used or handled. There are many of them for material objects, humans beings, animals, while ideas and abstract object are gathered into wider classes. Classifiers in sign languages are hand shapes, that are used to express physical properties, size, position, and also the way the classified object moves. Here are a few examples, from French sign language (LSF):

Hand shape	Classifier of ...
horizontal M hand shape	flat object, car, bus, train (not bike)
vertical M hand shape	bike, horse, fish,
Y handshape	plane
C handshape	small round or cylindrical object
forefinger up	person
fist	head of a person
4 hand shape	a line of people
three crouched fingers	small animal

The classifier used for a given object depends on what is said about the noun / entity represented by the classifier. For instance, a line of n people waiting to be served at the bakery may be represented by n fore fingers, in case for example, these n people are individualised and one wants to say they were discussing, or with the 4 hand shape of one wants to says they were waiting, they were numerous etc.

Some linguists, such as Cuxac (2000), call them pro-forms rather than classifiers. Pro-forms are analogous to pro-nouns: they stand for the form (shape) of the object: they refer to an object via its shape or part of its shape i.e. they depend on the aspect that is being referred to, just like the restriction of selection in lexical semantics. Polysemic mechanisms also apply to pro-forms, as different pro-forms can be used to refer to different facets of the same lexeme: e.g., a car might be referred to using a C shape (cylinder) pro-form to indicate that it is thought of as a container, or using a M shape (flat, horizontal hand, palm down) to indicate a moving vehicle.

Classifiers of sign languages are also used to identify how many objects one speaks about.

3.3 Classifiers in Japanese

In Japanese, the classifiers are used as counters, in a syntactic category formally known as “numerical auxiliaries”. They are always used in conjunction with a numeral, or a pronoun referring to a numeral:

- (8) Otoko no Hito ga nan Nin imasu ka ?
Male person SUB how-many counter for people live Q ?
'How many men are there ?'

In (8), *Nin* is the classifier for people. The rest of the sentence makes clear that we are referring to a specific subclass, men.

Japanese classifiers organise a hierarchy of sorts among the lexical entities. Children or people unfamiliar with the language can get by with a dozen common classifiers, mostly used as generic classes. Competent speakers of the language are expected to use the correct classifiers in a list comprising about a hundred entries. There are also a few hundred classifiers used only in specific situations such as restricted trades or professions, or ritualistic settings. Finally, classifiers can be generated from common nouns as a creative, non-lexical use of the word.

Examples of classifiers in that respect include:

Generic classifiers

- *Tsu*: empty semantic content, used to mean any object. Commonly translated as “thing”.
- *Nin*: people (human).
- Order (*Ban*), frequency (*Kai*), amount of time in minutes, hours, days, etc.
- *Hai*: measure. Used to mean “x units of” anything that is a mass concept, and is presented in a container (bottles of water, bowls of rice, cups of tea, etc.)

Common classifiers

- *Mai*: flat or slim objects, including paper, stamps, some articles of clothing, etc.
- *Dai*: vehicles, machines, appliances.
- *Ko*: small things (such as dice, keys, pins) or unspecified things (their classifier is not known to the speaker or does not exist).
- *Hon*: long and thin objects, such as pens, bottles, but also rivers, telephone calls (if they take a long time), etc.

Specialised classifiers

- *Bi*: fritter and small shrimps (for fishmongers).
- *Koma*: frames (for comic strip editors).

A complete discussion of the classifier system of Japanese or any other language falls outside the scope of this publication. What we want to illustrate is that it provides a linguistically sound classification of entities, applicable to any entity in the language (anything that can be referred by a pronoun), and derived from cognitive categories reflected by the etymology of the individual classifiers. In some cases, the classifiers are similar to words used in language that do not have a complete classifier system, such as the English *head* for units of cattle (the counter *Tō* for cattle and large animals is the character denoting “head”). In others, the metaphorical reasoning behind the lexical category is apparent (*Hon*, the character for “book” and “root”, is used to count long things, including objects that are physically long, rivers and coasts that have a similar shape on a map, and abstract things that take a long time such as calls, movies, tennis matches...).

The classifier system is very obviously the result of language evolution. In each language concerned, many classifiers have a different history (linguists have argued that the classifier system in Japanese, as well as in Korean and other languages of the Asia-Pacific region, has been heavily influenced by Chinese, see T’sou (2001) for details). However, the grammatical need to have a categorisation of entities in order for nouns to be countable or measurable has produced classes that share similar characteristics, suggesting that they are derived from natural observation of their salient features. In other words, even if classifiers are not commonly used in linguistics to denote anything other than numerical auxiliaries, we think they provide good candidates for a type system of the granularity we are interested in.

Moreover, classifiers can have a behaviour similar to lexical sorts in formal lexical semantics. Entities with the same denotation can have different classifiers if they are used in different contexts. *Nin* (people) can be used to count persons in many cases, but *Mei* (names) will have to be used in cases the situation calls for dignity and formality. *Hai* (full container) can be used to measure countable nouns, but also boats in a dismissive way (as a populist might refer to “a shipload of migrants”). Inapplicable classifiers can be used for metaphoric usages, puns, or obscure references to the particular etymology of a word or character. The overly obsequious humility of a character might be indicated by his use of the counter for small animals (rather than people) for himself; for other persons, this is considered a grave insult (often translated as “I am an unworthy insect” or “You are a mere ant to me”).

3.4 Classifiers as Base Types: Linguistic or Cognitive Choice ?

What is pleasant in the choice of classifiers as base types is that they are natural both from a cognitive and from a linguistic viewpoint. They definitely are linguistic objects, since they are part of the language, being independent morphemes (words or signs). However these morphemes represent nouns, or, more precisely, refer to the relevant aspect of the noun for a particular predicate (adjective or verb), this is the reason why several classifiers are possible for a given object. Thus they also gather objects (rather than words) that resemble

each other as far as a given predicate is applied to them, and this other aspect is more cognitive than linguistic.

Clearly, the precise classifier system depends on the language, but they obey some common general properties: it suggests that the classifier system is cognitively motivated. An intriguing common property is that physical entities that a speaker interact with have a very precise system of classifiers, with sub classifiers (i.e., a classifier being more specific than another), thus providing a kind of ontology in the language. For example, human beings and animals have classifiers, and there is a richer variety of classifiers for animals usual and closer to the human species: for instance there is a specific classifier in French sign language for small animals (rabbits, rats,...). Although it could seem natural for sign languages, because sign language is visual and gestural that physical entities have very refined classifier systems, as signs recall the visual aspects of objects and the way we handle them, it is surprising that the Asian classifier systems are actually as rich for physical objects as the one for French Sign Language. From what we read, it seems that all classifier system do represent fairly precisely the physical objects.

For this reason we think that the classifier system is halfway between a cognitively motivated set of sorts, and a linguistic system. It is thus a good answer to our initial practical question: what should the base types be in compositional semantics if one wishes to include some lexical semantics (e.g. to limit ambiguities) to a semantic parser.

We propose building, for use by the existing analysers for syntax and semantics, a system of sorts based on the observed classifier systems and adapted to the target languages (English, French...). The common use of the classifier systems indicate that they have a reasonable granularity. The classifier systems also have some limited redundancy and specialisation, that is included in our system as lexical modifiers indicating hyponymy and hyperonymy relations between sorts.

3.5 Integrating Base Types in our Lexicon

Our system requires base types in order to describe lexical sorts, that is, classes of entities that behave differently from one another as semantic units. These sorts are used to categorise nouns that refer to individuals, and form the base types of our hierarchy; predicates, action nouns, adverbs and adjectives are defined by complex or functional types built from those sorts.

We have seen that classifiers have many desirable qualities in the description of such classes, specifically as they apply to individuals. The cover provided is extensive, and the classification is linguistically motivated; some classifiers might have an archaic origin, or other peculiar features that makes them strongly idiosyncratic, but the strength of our system lies in the accurate representation of those idiosyncrasies, and we think classifiers provide a sound entry point for the classification necessary in our lexicon.

4 Conclusion

Our type-theoretical model of lexical semantics is already implemented in analysers for syntax and semantics based on refinements of Montague Grammar and categorial grammars, and has proven useful for the study of several specific linguistic issues, using restricted, hand-typed lexica. A first system being tested uses different sorts for regions, paths and times, as well as a fictive traveller, to analyse itineraries in a specific corpus of travel stories, as illustrated in section 2. The devising of a complete type system for each of the target languages, and thus the definition of a wide-coverage classification of entities into sorts, is a necessity for the next step: the completion of the lexicon and its semantics.

The base types, and the semantics for the transformations necessary for our approach, can be obtained by those methods or a combination thereof:

1. by statistical means (this is, however, a very difficult issue even with a very simple type system, see Zettlemoyer & Collins (2009) for a discussion);
2. by hand (this is possible for restricted domains);
3. by derivation from other linguistic data.

For that last method, we believe that the classifier systems used in various languages present the properties we would expect from such a type system. We propose to use the classifier systems as a template for classifying sorts in the target language, and are currently designing tests in order to confirm that such categories are identified as such by speakers of the language. For those languages that do not have classifiers, we are considering the adaptation of a classifier system of a language that does. Finally, if the kind of semantic analysis we want to perform is oriented towards some sorts, it is possible to use both classifiers and specific sorts.

Bibliography

- Asher, N. (2011). *Lexical Meaning in Context: a Web of Words*. Cambridge University Press.
- Asher, N., & Pustejovsky, J. (2005). Word Meaning and Commonsense Metaphysics. Semantics Archive.
- Bassac, C., Mery, B., & Retoré, C. (2010). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Language, Logic, and Information*, 19(2).
- Bekki, D., & Asher, N. (2012). Logical polysemy and subtyping. In Y. Motomura, A. Butler, & D. Bekki (Eds.) *JSAI-isAI Workshops*, vol. 7856 of *Lecture Notes in Computer Science*, (pp. 17–24). Springer.
- Blutner, R. (2002). Lexical Semantics and Pragmatics. *Linguistische Berichte*.
- Bos, J. (2008). Wide-coverage semantic analysis with boxer. In J. Bos, & R. Delmonte (Eds.) *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, (pp. 277–286). College Publications.
- Cooper, R. (2007). Copredication, dynamic generalized quantification and lexical innovation by coercion. In *Fourth International Workshop on Generative Approaches to the Lexicon*.
- Cuxac, C. (2000). *La Langue des Signes Française. Les voies de l'iconicité*. Ophrys.
- Fodor, J. A., & Lepore, E. (1998). The emptiness of the lexicon : Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry*, 29(2).
- Goddard, C., & Wierzbicka, A. (2007). Semantic primes and cultural scripts in language learning and intercultural communication. In F. Sharifian, & G. B. Palmer (Eds.) *Applied Cultural Linguistics: Implications for second language learning and intercultural communication*, (pp. 105–124). John Benjamins.
- Kamp, H., & Reyle, U. (1993). *From Discourse to Logic*. Dordrecht: D. Reidel.
- Lefevre, A., Moot, R., & Retoré, C. (2012a). Traitement automatique d'un corpus de récits de voyages pyrénéens : analyse syntaxique, sémantique et pragmatique dans le cadre de la théorie des types. In *Congrès mondial de linguistique française*.
- Lefevre, A., Moot, R., Retoré, C., & Sandillon-Rezer, N.-F. (2012b). Traitement automatique sur corpus de récits de voyages pyrénéens : Une analyse syntaxique, sémantique et temporelle. In *Traitement Automatique du Langage Naturel, TALN'2012*, vol. 2, (pp. 43–56).
- URL <http://aclweb.org/anthology/F/F12/>
- Li XuPing (2011). *On the semantics of classifiers in Chinese*. Ph.D. thesis, Bar-Ilan University.
- Mery, B. (2011). *Modélisation de la Sémantique Lexicale dans le cadre de la Théorie des Types*. Ph.D. thesis, Université de Bordeaux.
- Mery, B., Bassac, C., & Retoré, C. (2007). A montague-based model of generative lexical semantics. In R. Muskens (Ed.) *New Directions in Type Theoretic Grammars*. ESSLLI, Foundation of Logic, Language and Information.

- Mery, B., & Retoré, C. (2013). Recent advances in the logical representation of lexical semantics. In *NCLS – Workshop on Natural Language and Computer Science, LiCS 2013*. Tulane University, New Orleans.
- Montague, R. (1974). The proper treatment of quantification in ordinary English. In R. H. Thomson (Ed.) *Formal Philosophy*, (pp. 188–221). New Haven Connecticut: Yale University Press.
- Moot, R. (2010a). Semi-automated extraction of a wide-coverage type-logical grammar for French. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*. Montreal.
- Moot, R. (2010b). Wide-coverage French syntax and semantics using Grail. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*. Montreal.
- Moot, R., Prévot, L., & Retoré, C. (2011a). A discursive analysis of itineraries in an historical and regional corpus of travels. In *Constraints in discourse*, (p. <http://passage.inria.fr/cid2011/doku.php>). Ayay-roches-rouges, France.
URL <http://hal.archives-ouvertes.fr/hal-00607691/en/>
- Moot, R., Prévot, L., & Retoré, C. (2011b). Un calcul de termes typés pour la pragmatique lexicale — chemins et voyageurs fictifs dans un corpus de récits de voyages. In *Traitement Automatique du Langage Naturel, TALN 2011*, (pp. 161–166). Montpellier, France.
URL <http://hal.archives-ouvertes.fr/hal-00607690/en/>
- Muskens, R. (1996a). Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19, 143–186.
- Muskens, R. (1996b). Meaning and Partiality. In R. Cooper, & M. de Rijke (Eds.) *Studies in Logic, Language and Information*. CSLI.
- Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.
- Pustejovsky, J. (1998). Generativity and Explanation in Semantics: a reply to Fodor and Lepore. *Linguistic Inquiry*, 29, 289–311.
- Pustejovsky, J. (2001). Type construction and the logic of concepts.
URL citeseer.ist.psu.edu/pustejovsky01type.html
- Pustejovsky, J., & Asher, N. (2000). The Metaphysics of Words in Context. *Objectual attitudes, Linguistics and Philosophy*, 23, 141–183.
- Talmy, L. (1999). Fictive motion in language and “ception”. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.) *Language and Space*, (pp. 211–276). MIT Press.
- T’sou, B. K. (2001). Language contact and linguistic innovation. In M. Lackner, I. Amelung, & J. Kurtz (Eds.) *New Terms for New Ideas. Western Knowledge and Lexical Change in Late Imperial China*, (pp. 35–56). Koninklijke Brill.
- Wilks, Y. (2001). The “Fodor”-FODOR Fallacy bites back. In P. Bouillon, & F. Busa (Eds.) *The Language of Word Meaning, Studies in Natural Language Processing*. Cambridge University Press.
- Xue, T., & Luo, Z. (2012). Dot-types and their implementation. In *Béchet and Dikovsky (2012), pages 234-249*.
- Zettlemoyer, L. S., & Collins, M. (2009). Learning context-dependent mappings from sentences to logical form. In *ACL-2009*.
- Zwitserlood, I. (2012). Classifiers. In R. Pfau, M. Steinbach, & B. Woll (Eds.) *Sign Languages: an International Handbook*, (pp. 158–186). Mouton de Gruyter.

Hidden Structure and Function in the Lexicon

Olivier Picard¹, Mélanie Lord¹, Alexandre Blondin-Massé², Odile Marcotte^{3,6},
Marcos Lopes⁴ and Stevan Harnad^{1,5}

¹ Département de psychologie, Université du Québec à Montréal
picard.olivier.2@courrier.uqam.ca
{lord.melanie,harnad}@uqam.ca
<http://wwd.crcsc.uqam.ca>

² Département de mathématique, Université du Québec à Chicoutimi
ablondin@uqac.ca
<http://thales.math.uqam.ca/~blondin/>

³ Département d'informatique, Université du Québec à Montréal

⁴ Department of Linguistics, Universidade de São Paulo, Brazil
marcoslopes@usp.br

⁵ Department of Electronics and Computer Science, University of Southampton, UK
harnad@ecs.soton.ac.uk

⁶ Centre de recherches mathématiques, Université de Montréal
Odile.Marcotte@gerad.ca

Abstract. How many words are needed to define all the words in a dictionary? Graph-theoretic analysis reveals that about 10% of a dictionary is a unique Kernel of words that define one another and all the rest, but this is not the smallest such subset. The Kernel consists of one huge strongly connected component (SCC), about half its size, the Core, surrounded by many small SCCs, the Satellites. Core words can define one another but not the rest of the dictionary. The Kernel also contains many overlapping Minimal Grounding Sets (MGSs), each about the same size as the Core, each part-Core, part-Satellite. MGS words can define all the rest of the dictionary. They are learned earlier, more concrete and more frequent than the rest of the dictionary. Satellite words, not correlated with age or frequency, are less concrete (more abstract) words that are also needed for full lexical power.

1 Introduction

Dictionaries catalogue and define the words of a language.¹ In principle, since every word in a dictionary is defined, it should be possible to learn the meaning of any word through verbal definitions alone (Blondin-Massé et al. 2013). However, in order to understand the meaning of the word that is being defined, one has to understand the meaning of the words used to define it. If not, one has to look up the definition of those words too. But if one has to keep looking up the definition of each of the words used to define a word, and then the definition of each of the words that define the words that define the words, and so on, one will eventually come full circle, never having learned a meaning at all.

This is the *symbol grounding problem*: The meanings of all words cannot be learned through definitions alone (Harnad 1990). The meanings of some words, at

least, have to be “grounded” by some other means than verbal definitions. That other means is probably direct sensorimotor experience (Harnad 2010), but the learning of categories from sensorimotor experience is not the subject of this paper. Here we ask only *how many words* need to be known by some other means such that all the rest can be learned via definitions composed only of those already known words, and *how do those words differ from the rest?*

2 Dictionary Graphs

To answer this question dictionaries can be analyzed using graph theory. These analyses have begun to reveal a hidden structure in dictionaries that was not previously known (see Fig. 1). By recursively removing all the words that are defined but do not define any further word, every dictionary can be reduced by about 90% to a unique set of words (which we have called the *Kernel*) from which all the words in the dictionary can be defined (Blondin-Massé et al. 2008). There is only one such Kernel in any dictionary, but *the Kernel is not the smallest number of words out of which the whole dictionary can be defined*. We call such a smallest subset of words a *Minimal Grounding Set* (MGS). (In graph theory it is called a “minimum feedback vertex set”; Karp 1972; Lapointe et al. 2012.) The MGS is about half the size of the Kernel (Table 2), but, unlike the Kernel, it is not unique: There are a huge number of (overlapping) MGSs in every dictionary, each of the same minimal size; each is a subset of the Kernel and any one of the MGSs grounds the entire dictionary.

The Kernel, however, is not just a large number of overlapping MGSs. It has structure too. It consists of a large number of strongly connected components (SCCs). (A directed graph -- in which a directional link indicates that word A belongs to the definition of word B -- is “strongly connected” if every word in the graph can be reached by a chain of definitional links from any other word in the graph.) Most of the SCCs of the Dictionary’s Kernel are small, but in every dictionary we have analyzed so far there also turns out to be one very large SCC, about half the size of the Kernel. We call this the Kernel’s *Core*².

The Kernel itself is a self-contained dictionary, just like the dictionary as a whole: every word in the Kernel can be fully defined using only words in the Kernel. The Core is likewise a self-contained dictionary; but the Core is also an SCC (at least in all the full-size dictionaries of natural languages that we have so far examined³), whereas the Kernel is not: Every word within the Core can be reached by a chain of definitions from any other word in the Core. In what follows, our statements about the Core will assume that we are discussing full-size dictionaries of natural languages (unless stated otherwise).

The Kernel is a Grounding Set for the dictionary as a whole, but it is not a *Minimal Grounding Set* (MGS) for the dictionary as a whole. The Core, in contrast, is not only *not* an MGS for the dictionary as a whole: it is not even a Grounding Set at all. The words in the Core alone are not enough to define all the rest of the words in the dictionary, outside the Core.

In contrast, the MGSs -- which, like the Core, are each about half the size of the Kernel -- are each contained within the Kernel, but none is completely contained within the Core: Each MGS straddles the Core and the surrounding “Satellite” layer

of smaller SCCs. Each MGS can define all the rest of the words in the dictionary, but no MGS is an SCC (see Fig. 1 & Table 1).

The MGSs of Dictionaries hence turn out empirically⁴ to consist of words that are partly in the Core (which is entirely within the Kernel) and partly in the remainder of the Kernel (K) -- the part outside the Core (C), which we call the *Satellites* (K minus C) because they consist of much smaller SCCs, encircling the huge Core. The MGSs, the smallest subsets capable of defining all the rest of the dictionary, are hence part-Core and part-Satellite. The natural question, then, is: Is there any difference between the *kinds* of words that are in the various components of this hidden structure of the dictionary: the MGSs, the Core, the Satellites, the Kernel, and the rest of the dictionary outside the Kernel?

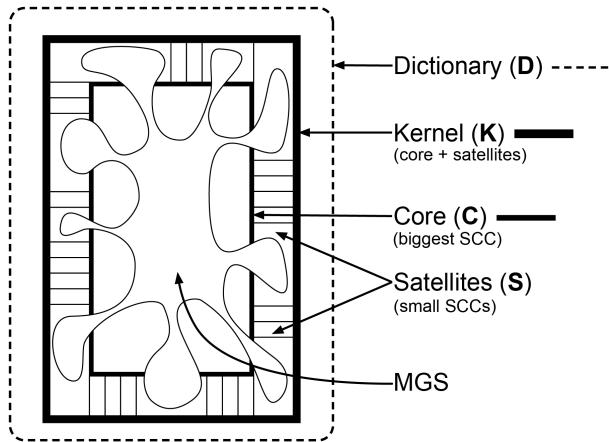


Fig. 1. Diagram of the hidden structure of dictionaries. Every dictionary (D) tested so far (see Table 2) contains a *Kernel* (K) of words (fewer than 10% of the dictionary) from which all the words in the dictionary can be defined. But the Kernel is not the smallest number of words that can define all the rest. In the graph of the Kernel, a directional link means that one word defines another word. The Kernel consists of many subsets in which every word is connected to every other word via a chain of definitional links. Each such subset is called a *Strongly Connected Component* (SCC). About half of the Kernel consists of one big SCC, the *Core* (C). The rest of the Kernel is small SCCs (*Satellites*) (S) surrounding the Core. The Core alone can define all of its own words, but not all the rest of the words in the Kernel (hence it cannot define the dictionary as a whole either). Solving a graph-theoretic problem for the Kernel of a dictionary (finding its “minimal feedback vertex set”) reveals the smallest number of words from which all the rest of its words can be defined: the *Minimal Grounding Set* (MGS). The MGS is also about half the size of the Kernel, but there are a huge number of overlapping MGSs in the Kernel, each of which includes words from both the Core and its Satellites (but only one of the MGSs is illustrated here). The words in these different structural components of the Dictionary Graph turn out to have different psycholinguistic properties (Figs. 2 & 3). (Note that the diagram is not drawn to scale, as K is really only 10% of D.)

Associated with this hidden graph-theoretic structure of the dictionary some evidence of hidden psycholinguistic function is beginning to emerge. It turns out that the words in the Kernel are learned at a significantly younger age, and are more

concrete and frequent than the words in the rest of the Dictionary. The same is true, but moreso, comparing the Core with the rest of the dictionary, and still moreso comparing the MGSSs with the rest of the dictionary (Fig. 2, left). There are hints, however, that something more subtle is also going on: All five psycholinguistic variables are themselves highly inter-correlated. If we factor out their inter-correlations and look only at their independent effects, the relationships with the hidden structure of the dictionary change subtly: The words in the Kernel remain younger and more frequent than the rest of the dictionary, but once the variance correlated with age is removed, for the residual variance the Kernel is *more abstract* than the rest of the Dictionary. In contrast, this reversal does not happen for either the Core or the MGSSs. Hence the locus of the reversal is the Satellite layer (Fig. 2, right.). We will now describe the analyses that generated this pattern of results.

Table 1. Practically speaking, a *Dictionary* (D) is a set of words and their definitions in which all the defined and defining words are defined. By recursively removing words that are not used to define further words and that can be reached by definition from the remaining words, a Dictionary can be reduced by about 90% to a *Kernel* (K) of words from which all the other words can be defined. The Kernel is hence a *Grounding Set* (GS) of a Dictionary: a subset that is itself a Dictionary, and that can also define all the words in the rest of the Dictionary. A *Strongly Connected Component* (SCC) of a Dictionary graph is a subset in which there is a definitional path from every word to every other word in the subset. The Kernel's Core (C) is the union of all the strongly connected components (SCCs) of the Kernel that do not receive any incoming definitional links from outside themselves. *Minimal Grounding Sets* (MGSSs) are the smallest-sized subsets of words that can define all the words in the rest of the Dictionary. (Any Dictionary has only one Kernel but many MGSSs. In all full dictionaries analyzed so far, the Core has always been an SCC, but in some mini-dictionaries generated by the online Dictionary game the Core was not an SCC, but a disjoint union of SCCs).

a \Rightarrow is necessarily a ↓	Dict	Kern	GS	SCC	Core	MGS
Dictionary (D)	x	x	-	-	x	-
Grounding Set (GS)	x	x	x	-	-	x
Strongly Connected Component (SCC)	-	-	-	x	x	-
Minimal Grounding Set (MGS)	-	-	-	-	-	x

3 Psycholinguistic Properties of Hidden Structures

The MRC database (Wilson 1987) provides psycholinguistic data for words of the English language, including (1) average age of acquisition, (2) degree of concreteness (vs. abstractness), (3) written frequency, (4) oral frequency and (5) degree of (visual) imageability. Analyses of Variance (ANOVAs) reveal that the words in the Kernel (K) differ significantly ($p < .001$) from words in the rest of the Dictionary (D) for all five variables: Kernel words are learned significantly younger, more concrete, more frequent, both orally and in writing, and more imageable than words in the rest of the dictionary. The same effect was found for all five variables in comparing Core (C) words with the rest of the dictionary as well as in comparing MGS words with the rest

of the dictionary. The effect was likewise found in comparing Core words with the rest of the Kernel (the Satellites, S) rather than the rest of the dictionary, as well as in comparing MGS words with the rest of the Kernel rather than the rest of the dictionary. Hence the conclusion is that the effects get stronger as one moves from Dictionary to Kernel to Core to MGS for each of the five psycholinguistic variables, as schematized on the left part of Fig. 2. (Three different MGSs were tested, with much the same result.)

One can summarize these findings as the relation MGS > C > S > K > D, meaning that words in an MGS (for instance) are learned younger and more concrete, frequent, and imageable than words in the Core, the Satellites, the Kernel, or the whole Dictionary minus the Kernel (D – K).

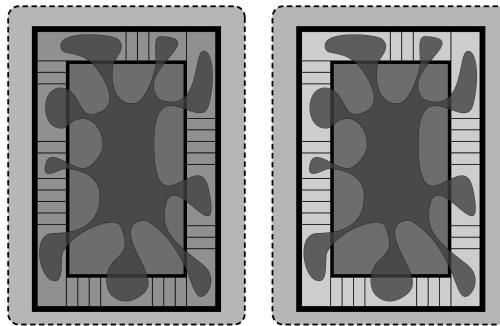


Fig. 2. Moving inward toward the Core, words are more concrete, more frequent, and learned younger. *Left:* Based on data from the MRC Psycholinguistic database (Wilson 1987), the general pattern observed is that compared to the words in the remaining 90% of the dictionary (LDOCE, Proctor 1981), the words in the Kernel tend to be learned at a significantly younger age, more concrete, more imageable, and more frequent, both orally and in writing. The darkness level in the figure indicates the size and direction of the difference, which is about the same for all five variables: MGS > C > S > K > D. (Fig. 1 has arrows pointing to each of these structures). *Right:* All 5 variables are intercorrelated, however, so when the comparison is done with a multiple regression analysis that measures each variable's contribution independently, the pattern is similar, but the difference in imageability and oral frequency becomes insignificant and a significant reversal in one of the variables (concreteness) emerges: Those Satellite words that are uncorrelated with age of acquisition or frequency tend to be significantly more *abstract* than the rest of the dictionary. This figure illustrates the pattern schematically; the quantitative data are shown in Fig. 3. Only one MGS is shown; the pattern is similar for all three MGSs tested.

The five psycholinguistic variables are all highly inter-correlated, however, so in order to test their effects independently of one another, we performed a step-wise multiple regression analysis, introducing one variable at a time according to its strength in accounting for the variance (Fig. 3). In the comparison of the Kernel vs. the rest of the dictionary with all 5 variables, 83% of the variance was accounted for, but two of the variables (imageability and oral frequency) no longer made a significant contribution -- nor did they do so in any of the other stepwise regressions; so we drop them from our analysis and interpretation. Age made the biggest contribution, in the same direction as in the ANOVAs, the Kernel words being

(acquired) younger than the rest of the dictionary. The second strongest variable was written frequency, likewise in the same direction as the ANOVAs; and the third variable was concreteness. The independent predictive contribution of all three variables was significant ($p < .001$). However, the direction of the concreteness variable was reversed: For the component of the variance *not* correlated with age or frequency, the Kernel words turn out to be *more abstract* than the rest of the dictionary (Fig. 3a).

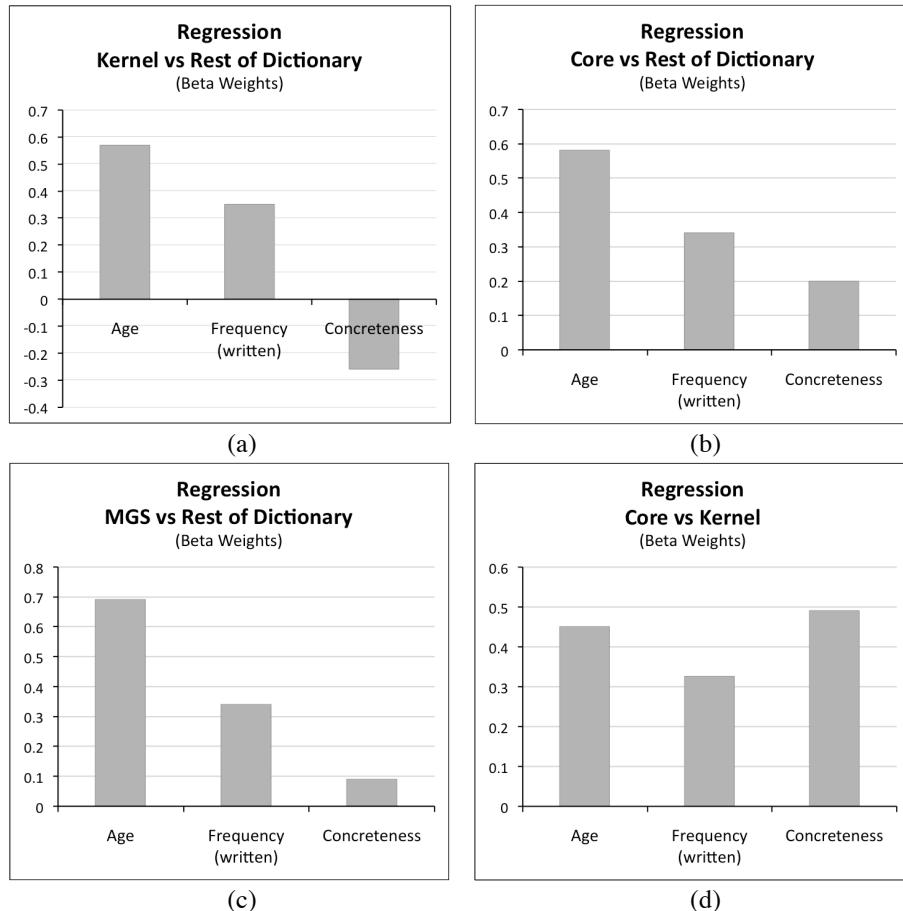


Fig. 3. Independent regression analysis of psycholinguistic differences. Stepwise linear regression reveals that the words in the Kernel (a), Core (b) and MGS (c) of the dictionary are learned significantly younger and are more frequent in writing than words in the rest of the dictionary. Pairwise ANOVAs show that the Kernel, Core and MGS words are also more concrete than the rest of the Dictionary, but the regression analysis shows a reversal for concreteness in the Kernel (a). Since the Core is more concrete than the Kernel (d), the likely cause of the reversal for concreteness is that the Satellite layer of SCCs that is located between the Core and the Kernel is more abstract (see text for discussion).

This significant reversal in the Satellite layer was the only one observed in the stepwise regressions. For the Core versus the rest of the dictionary (see Fig. 3b), the directions of the independent effects in the stepwise regression were the same as they were in the ANOVAs for age, concreteness and written frequency. The same was true for the MGS versus the rest of the dictionary, except that the effect of concreteness was very weak (see Fig. 3c).

The regression results for the Core versus the rest of the Kernel were also in the same direction as the ANOVAs, but in this comparison, the biggest effect of the three variables was for concreteness. In all the other comparisons the biggest effect had been that of age. The regressions comparing MGSs to the rest of the Kernel were inconclusive.

We accordingly conclude that in the Satellite layer of the Kernel, the words whose acquisition is uncorrelated with age or frequency are more abstract. The Core words may be the more concrete and frequent words that must be learned early, whereas the Satellite words that are not learned early may be more abstract because they are the kinds of words needed, in addition to Core words, in order to form MGSs that can generate the meanings of all other words -- and these Satellite words continue to grow throughout the life cycle.

4 Discussion

Our findings suggest that in addition to the overall tendency for words to be younger, more concrete and more frequent as one moves from the outer 90% of the dictionary to the Kernel to the Core to the MGSs, something importantly different may be happening at the Satellite layer, which, unlike the deeper layers (Core and MGS) is more abstract than the rest of the dictionary, rather than more concrete, like the rest of the Kernel (for words not learned at any particular age). It is almost certain that the Core is the most concrete of all, and that the MGSs are somewhat less concrete because, besides containing Core words, they also contain some Satellite words, which are more abstract.

These results have implications for the understanding of symbol grounding and the learning and mental representation of meanings. In order for language users to learn and understand the meaning of words from verbal definitions, they have to have the vocabulary to understand the words in the definitions, or at least to understand the definitions of the words in the definitions, and so on. They need an already grounded set of word meanings sufficient to carry them on, verbally, to the meaning of any other word in the language, if they are to learn its meaning through words alone. A grounding set clearly has to be acquired before it is true that all possible further words can be acquired verbally; hence it makes sense if the grounding set needs to be acquired earlier. It also makes sense that the words in the grounding set are more frequently used, possibly because they are used more often to define other words, especially initially (and perhaps even moreso when used formally, in writing, rather than orally).

That the grounding words are more concrete is also to be expected, because word meanings that do not come from verbal definitions have to be acquired by nonverbal means, and those nonverbal means are likely to be the learning of categories through

direct sensorimotor experience: learning what to do and not do with what kind of thing. It is easy, then, to associate a category that one has already learned nonverbally with the (arbitrary) name that a language community agrees to call it (Blondin-Massé et al. 2013). The words denoting sensorimotor categories are hence likely to be more concrete (although they may not necessarily be visually imageable, as there are other concrete sensorimotor modalities too, such as hearing, touch and movement).

Categorization itself, however, is by its nature also *abstraction*: To abstract is to single out some properties of a thing, and ignore others. The way we learn what kinds of things there are, and what to do and not do with them, is not by simply memorizing raw sensorimotor experiences by rote. We learn through trial and error sensorimotor interactions to abstract the invariant properties of sensorimotor experiences that determine whether or not an instance is a member of a category, and we learn to ignore the rest of the sensorimotor variation as irrelevant. The process of abstraction in the service of categorization leads in turn to higher-order categories, which are hence more likely to be verbal ones rather than purely sensorimotor ones. For example, we can have a preverbal category for “bananas” and “apples,” based on the differing sensorimotor actions needed to eat them; but the higher-order category “fruit” is not as evident at a nonverbal level, being more abstract. It is also likely that having abstracted the sensorimotor properties that distinguish the members and nonmembers of a concrete category nonverbally, we will not just give the members of the category a name, but we may go on to abstract and name their properties (yellow, red, round, elongated) too. It may be that some of these higher-order category names for more abstract categories are as essential in forming a grounding set as the more concrete categories and their names.

Finally, the lexicon of the language – our repertoire of categories – is open-ended and always growing. To understand the grounding of meaning it will be necessary not only to look at the growth across time of the vocabulary (both receptive and productive) of the child, adolescent and adult, but also the growth across time of the vocabulary of the language itself (diachronic linguistics), to understand which words are necessary, and when, in order to have the full lexical power to define all the rest (Levary et al. 2012). We have discussed Minimal Grounding Sets (MGSs), and it is clear that there are potentially very many of these; but it is not clear that anyone uses just one MGS, or could actually manage to learn everything verbally knowing just an MGS. Perhaps we need some redundancy in our Grounding Sets. The Kernel, after all, is only twice as big as an MGS. And perhaps we don’t even need a full Grounding Set in order to get by, verbally; maybe we can manage with gaps. Certainly the child must, at least initially. Nor is it clear -- even if we have full mastery of enough MGSs or a Kernel -- that the best way to learn the meaning of all subsequent words is from verbal definitions alone. Although language may well have evolved in order to make something like that possible in principle -- acquiring new categories purely by verbal “telling,” without sensorimotor “showing” (Blondin-Massé et al. 2013) -- in practice the learning of new word meanings may still draw on some hybrid show-and-telling.

Limitations. Many approximations and simplifications have to be taken into account in interpreting these findings. We are treating a definition as an unordered string of words, excluding functional (stop) words and not making use of any syntactic structure. Many words have multiple meanings, and we are using only the first meaning of each word. The MRC psycholinguistic database only provides data

for about a quarter of all the words in the dictionary. The problem of extracting MGSs is NP-hard. In the special case of dictionary graphs -- and thanks also to the empirical fact that the Core turns out to be so big, and surrounded by small Satellites -- we have been able, using the algorithm of Lin & Jou (2000) and techniques from integer linear programming (e.g., Nemhauser & Wolsey 1999), to extract a number of MGSs for the small dictionary whose results we are reporting here (LDOCE). This analysis needs to be extended to a larger number of independent MGSs, to other, bigger dictionaries, such as Merriam-Webster and WordNet (Fellbaum 2010), as well as to other languages. These further studies are underway (Table 2). Note that to compute MGSs of dictionaries as large as Merriam-Webster and WordNet, one will need sophisticated techniques from integer linear programming and combinatorial optimization: we will report on these in subsequent articles.

Table 2. For all four full dictionaries of natural languages analyzed to date, the Kernel is less than 10% (5-9%) of the dictionary as a whole, the Core (biggest SCC) and its Satellites (small SCCs) are each about half the size of the Kernel (39-61%), and each MGS (part-Core, part-Satellites) is also about half the size of the Kernel. (LDOCE: Longman Dictionary of Contemporary English; CIDE: Cambridge Dictionary of Contemporary English; MWC: Merriam-Webster Dictionary; WN: WordNet)

	Dictionary Name			
	LDOCE	CIDE	MWC	WN
Whole Dictionary (D) Number of words	70545	47988	249739	132477
Kernel (K) Word count %D	4656 7%	4169 9%	13181 5%	12015 9%
Satellites (S) - all small SCCs Word count %D %K	2762 4% 59%	2042 5% 49%	5028 2% 38%	5623 4% 47%
Core (C) - largest SCC Wordcount %D %K	1894 3% 41%	2127 4% 51%	8153 3% 62%	6392 5% 53%
MGS - Minimal Grounding Set Word count %D %K	2254 3% 48%	1973 4% 47%	future work	future work

5 Future Work

In order to compare the emerging hidden structure of dictionaries with the way word meaning is represented in the mind (the “mental lexicon”) we have also created an online dictionary game in which the player is given a word to define; they must then define the words they used to define the word, and so on, until they have defined all the words they have used. This generates a mini-dictionary of a much more tractable size (usually less than 500 words; Figs. 4 & 5).⁵

We are currently performing the same analyses on these much smaller mini-dictionaries, to derive the Kernel, Core, Satellites and MGSs and their psycholinguistic correlates (age, concreteness, imageability, oral/written frequency), to determine whether these inner “mental” dictionaries share the hidden structure and function that we are discovering in the formal external lexicon (see Figs. 4 & 5). These mini-dictionaries will also allow us to analyze the difference in functional role among the words in the various components of the hidden structures by examining all the individual words, which is impossible with full-size dictionaries.

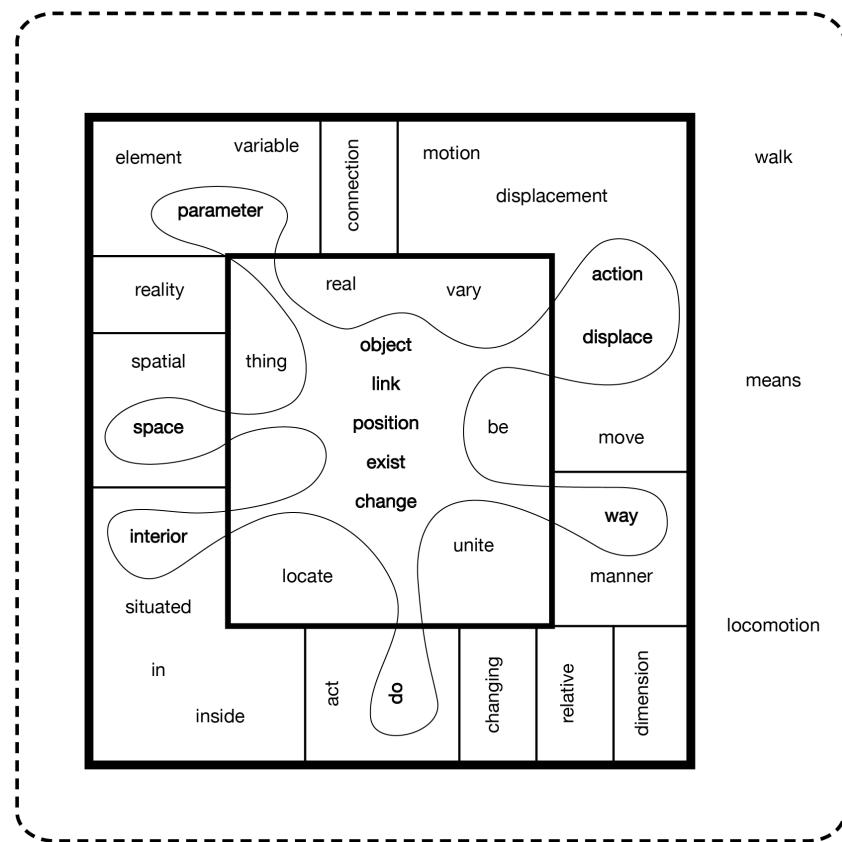


Fig. 4. Mini-dictionary Diagram. The diagram is the same as Fig. 1, but with real words to provide a concrete example. This 37-word mini-dictionary was generated by a player of an online dictionary game. The player is given a word and must define that word, as well as all the words used to define it, and so on, until all the words used are defined. The smallest resulting dictionary so far (37 words) is used here to illustrate the mini-dictionary's Kernel and Core plus one of its MGSs. Note that all the words in this mini-dictionary are in the Kernel except the start word, "walk," plus "locomotion" and "means." Fig. 5 displays the graph for this mini-dictionary.

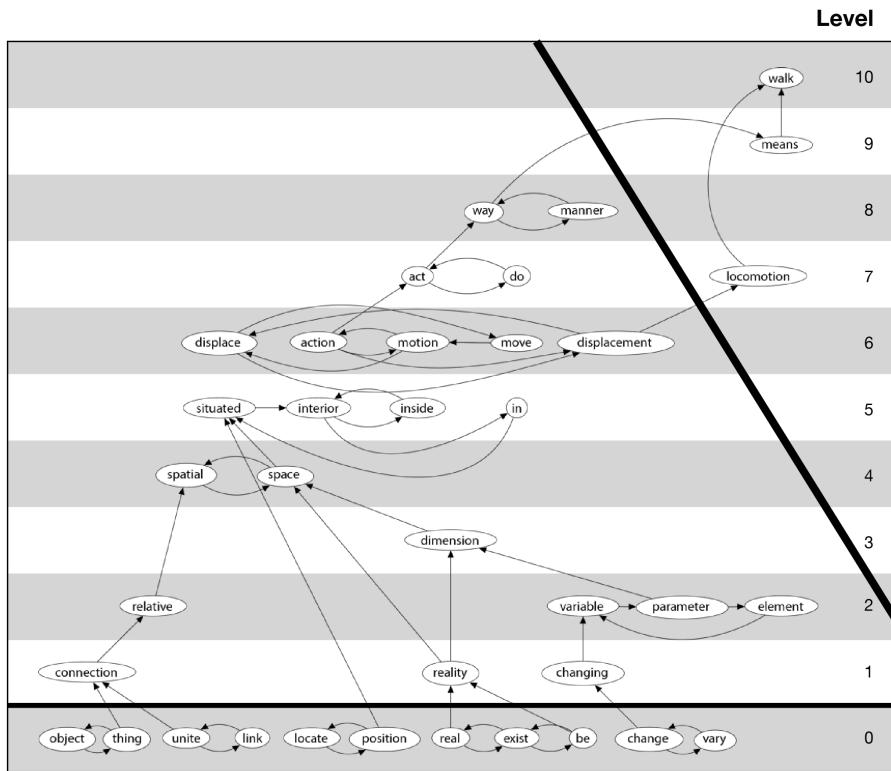


Fig. 5. Mini-dictionary Graph. Graph of mini-dictionary in Fig. 4, showing the definitional links. Note that in this especially tiny mini-dictionary, unlike in the full dictionaries and many of the other mini-dictionaries, the words in the Core (level 0), rather than being the single largest SCC, are the union of multiple SCCs. The oblique boldface line separates the Kernel from the (three) words in the rest of this mini-dictionary.

References

Blondin-Massé, A., Harnad, S., Picard, O. & St-Louis, B. (2013) Symbol Grounding and the Origin of Language: From Show to Tell. In: Lefebvre C., Comrie B. & Cohen H. (Eds.) *Current Perspective on the Origins of Language*. Benjamin.
<http://eprints.ecs.soton.ac.uk/21438/>

Blondin-Massé, A., Chicoisne G., Gargouri, Y., Harnad S., Picard O., & Marcotte, O. (2008). How Is Meaning Grounded in Dictionary Definitions? In *TextGraphs-3 Workshop - 22nd International Conference on Computational Linguistics*.
<http://www.archipel.uqam.ca/657/>

Fellbaum, C. (2010). *WordNet*. Springer: Netherlands. <http://wordnet.princeton.edu>

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42: 335-346.
<http://cogprints.org/0615/>

Harnad, S. (2005) *To Cognize is to Categorize: Cognition is Categorization*. In Lefebvre, C. & Cohen, H., Eds., *Handbook of Categorization*. Elsevier.
<http://eprints.ecs.soton.ac.uk/11725/>

Harnad, S. (2010) *From Sensorimotor Categories and Pantomime to Grounded Symbols and Propositions*. In: Tallerman, M. & Gibson, K.R.: *The Oxford Handbook of Language Evolution*, Oxford University Press.
<http://eprints.ecs.soton.ac.uk/21439/>

Karp, R.M. (1972) Reducibility Among Combinatorial Problems. In: *Complexity of Computer Computations*, Proc. Sympos. IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y.. New York: Plenum, pp. 85-103.

Lapointe, M., A. Blondin-Massé, P. Galinier, M. Lord & O. Marcotte. (2012) Enumerating minimum feedback vertex sets in directed graphs. *Bordeaux Graph Workshop* 101-102.
<http://bgw2012.labri.fr/booklet.pdf>

Levary, D., Eckmann, J. P., Moses, E., & Tlusty, T. (2012). Loops and self-reference in the construction of dictionaries. *Physical Review X*, 2(3) 031018.

Lin, H.M., & Jou, J.Y. (2000) On computing the minimum feedback vertex set of a directed graph by contraction operations. *IEEE Transactions on CAD of Integrated Circuits and Systems* 19(3) 295-307.

Nemhauser, G., and Wolsey, L. (1999) *Integer and Combinatorial Optimization*, Wiley.

Procter, P. (1981) *Longman dictionary of contemporary English*.

Wilson, M. D. (1987) *MRC Psycholinguistic Database: Machine Usable Dictionary; Version 2.00*. Informatics Division, Science and Engineering Research Council, Rutherford Appleton Laboratory.

¹ Almost all the words in a dictionary (whether nouns, verbs, adjectives or adverbs) are “content” words, i.e., they are the names of categories (Harnad 2005). Categories are kinds of things, both concrete and abstract (objects, properties, actions, events, states). The only words that are not the names of categories are logical and grammatical “function” words such as *if, is, the, and, not*. Our analysis is based solely on the content words; function (“stop”) words are omitted.

² Formally, the Core is defined as the union of all the strongly connected components (SCCs) of the Kernel that do not receive any incoming definitional links from outside themselves. (In graph-theoretical language: there is no incoming arc into the Core, i.e., there is no definitional link from a word not in the

Core to a word in the Core.) It turns out to be an empirical fact about all the full-sized dictionaries we have analyzed so far, however, that their Core is itself always an SCC, and also by far the largest of the SCCs in the Kernel, the rest of which look like many small satellites surrounding one big planet (Fig. 1).

³ In some of the mini-dictionaries generated in our online dictionary game, however, the Core is not an SCC, but a disjoint union of SCCs (Figs. 4 & 5).

⁴ Most of the properties described here are empirically observed properties of Dictionary graphs, not necessary properties of directed graphs in general.

⁵ The 37-word mini-dictionary in Figs. 4 & 5 is displayed because it is small enough to illustrate the hidden dictionary graph structure at a glance (and the referees asked for a real example). It was generated before we had added a new rule that a definition is not allowed to be just a synonym: In the more recent version of the game a definition has to be at least two content words (and we may eventually also rule out second-order circularity [$A = B + C$, $B = C + \text{not}A$, $C = A + \text{not}B$]). But it has to be borne in mind that (because of the symbol grounding problem) every dictionary is necessarily *approximate* and (at some level) *circular* (much the way all SCCs are circular). This is true whether it is a full dictionary or a game mini-dictionary generated by one player. Definitions can only convey new meanings if the mind already has enough old meanings, grounded by some means other than definition.

From Stimulus to Associations and Back

Reinhard Rapp

Aix-Marseille Université, Laboratoire d'Informatique Fondamentale
163 Avenue de Luminy, F-13288 Marseille, France
reinhardrapp@gmx.de

Abstract. Free word associations are the words human subjects spontaneously come up with upon presentation of a stimulus word. In experiments comprising thousands of test persons large collections of associative responses have been compiled. In previous publications it was shown that these human associations can be resembled by statistically analyzing the co-occurrences of words in large text corpora. In the current paper for the first time we consider the reverse question, namely whether the stimulus can be predicted from the responses. By presenting an algorithm which produces surprisingly good results our answer is clearly affirmative.

1 Introduction

Word associations have always played an important role in psychological learning theory, and have been investigated not only in theory, but also in experimental work where e.g. such associations were collected from human subjects. Typically, the subjects obtained questionnaires with lists of stimulus words, and were asked to write down for each stimulus word the spontaneous association which first came to mind. This led to collections of associations, the so-called association norms, as exemplified in Table 1.

Association theory, which can be traced back to Aristotle in ancient Greece, has often stated that our associations are governed by our experiences. For example, more than a century ago William James (1890) formulated this in his book "The principles of Psychology" as follows:

"Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity."

This citation is talking of *objects*, but the question arose whether for words the same principles might apply, and with the advent of corpus linguistics it was possible to verify this experimentally by looking at the distribution of words in texts. Among the first to do so were Church & Hanks (1990), Schvaneveldt et al. (1989), and Wettler & Rapp (1989).

Their underlying assumption was that strongly associated words should often occur in close proximity in text corpora. This is actually confirmed by corpus evidence:

Figure 1 assigns to each stimulus word position 0, and displays the occurrence frequencies of its primary associative response (most frequent response as produced by the test persons) at relative distances between -50 and +50 words. However, to give a general picture and to abstract away from idiosyncrasies, the figure is not based on a single stimulus/response pair, but instead represents the average of 100 German stimulus/response pairs as used by Russell & Meseck (1959). The effect is in line with expectations: The closer we get to the stimulus word, the higher the chances that the primary associative response occurs. Only for distances plus and minus one there is an exception, but this is an artefact because content words are typically separated by function words, and among our 100 primary responses there are no function words. Also, test persons typically select content words only.

CIRCUS	FUNNY	NOSE
clown (24)	laugh (23)	face (16)
ring (10)	girl (11)	eyes (12)
elephant (6)	joke (8)	mouth (11)
tent (6)	laughter (6)	ear (10)
animals (5)	amusing (4)	eye (6)
top (5)	hilarious (4)	throat (4)
boy (4)	comic (3)	smell (3)
clowns (3)	ha ha (3)	bag (2)
horse (2)	ha-ha (3)	big (2)
horses (2)	sad (3)	handkerchief (2)

Table 1: Top ten sample associations to three stimulus words as taken from the Edinburgh Associative Thesaurus. The numbers of subjects responding with the respective word are given in brackets.

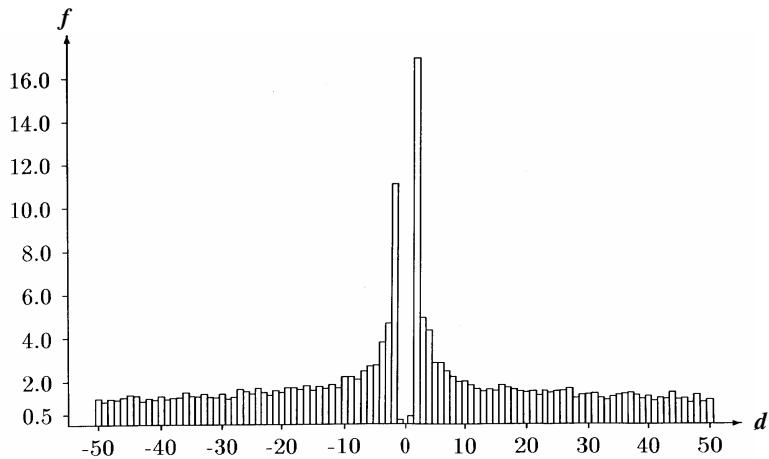


Figure 1: Occurrence frequency f of a primary response at distance d from a stimulus word, averaged over 100 stimulus/response pairs (Rapp, 1996).

Whereas such considerations are the basis underlying our work, in the current paper the focus is on whether it is possible to not only compute the responses from the stimulus, but also to compute the stimulus from the responses. To the best of our knowledge, this has not been attempted before in a comparable (distributional semantics) framework, so we are not aware of any directly related literature.

However, this task is somewhat related to the computation of associations when given several stimulus words simultaneously, which is sometimes referred to using the term *multiword associations* (Rapp, 2008) or the term *remote association test* (RAT). A recent notable publication on the RAT, which gives pointers to other related work, is Smith et al. (2013) who apply this for problems that require consideration of multiple constraints, such as choosing a job based on salary, location, and work description. Another one is Griffiths et al. (2007) who assume that concept retrieval from memory can be facilitated by inferring the gist of a sentence, and using that gist to predict related concepts and disambiguate words. They implement this by using a topic model.

Our approach differs from this previous work in that it focuses on a related but different and particularly well defined task. In our approach, we have eliminated all (for this particular task) unnecessary sophistication, such as *Latent Semantic Analysis* (which we used extensively in previous work) or *Topic Modeling*, resulting in a simple yet effective algorithm. For example, Griffiths et al. (2007) report 11.54% correctly predicted first associates. Rapp (2008) presents a number of evaluations using various corpora and data sets, but with all results below 10%. The above mentioned paper by Smith et al., 2013, gives no such figures at all. In comparison, the best results presented here are at 54%, see section 3.3. It should be emphasized, however, that all comparisons have to be taken with caution as there is no commonly used gold standard for this, so all authors used different test data, and also they used different corpora. Note also that, in contrast to the related work, our focus is on the novel reverse association task, which gives us test data of unprecedented quality and quantity (as any word association norm can be used), but for which the previous test data is unsuitable as it relates to a somewhat different task.

The paper is structured as follows: We first look at how we compute associations to single stimulus words. This lays the basis for the second part where we reverse our viewpoint and compute the stimulus word from its associations. For both tasks we present results and conclude with a discussion of our findings.

2 Computing forward associations

2.1 Procedure

As discussed in the introduction, we assume that there is a relationship between word associations as collected from human subjects and word co-occurrences as observed in a corpus. As our source of human data we use the *Edinburgh Associative Thesaurus* (EAT; Kiss et al. 1973; Kiss 1975) which is the largest classical collection of its kind.¹ The EAT comprises about 100 associative responses as requested from British students for each of altogether 8400 stimulus terms. As some of these stimulus terms are multiword units which we did not want to include here, we removed these from the thesaurus, so that 8210 items remained.

To obtain the required co-occurrence counts we aimed for a corpus which is as representative as possible for the language environment of the EAT's British test subjects. We therefore chose the *British National Corpus* (BNC), a 100-million-word corpus of written and spoken language which was compiled with the intention of providing a balanced sample of British English (Burnard & Aston, 1998). For our purpose it is also an advantage that the texts in the BNC are not very recent (from 1960 to 1993), thereby including the time period when the EAT data was collected (between June 1968 and May 1971).

Since function words were not considered important for our analysis of word semantics, to save memory requirements and processing time we decided to remove them from the text. This was done on the basis of a list of approximately 200 English function words.

We also decided to lemmatize the corpus using the lexicon of full forms provided by Karp et al. (1992). This not only improves the problem of data sparseness, but also significantly reduces the size of the co-occurrence matrix to be computed. Since most word forms are unambiguous concerning their possible lemmas, we only conducted a partial lemmatization that does not take the context of a word into account and thus leaves the relatively few words with several possible lemmas unchanged. For consistency reasons, we applied the same lemmatization procedure to the whole EAT. Note that, as the EAT contains only isolated words, in this case a lemmatization procedure that takes the context of a word into account would not be possible.

For counting word co-occurrences, as in most other studies a fixed window size is chosen and it is determined how often each pair of words occurs within a text window of this size. Choosing a window size usually means a trade-off between two parameters: specificity versus the sparse-data problem. The smaller the window, the more salient the associative relations between the words inside the window, but the more severe the problem of data sparseness. In our case, with ± 2 words, the window size looks rather small. However, this can be justified since we have reduced the effects of data sparseness by using a large corpus and by lemmatizing the corpus. It also should

¹ An even larger, though possibly more noisy, association database has been collected via online gaming at www.wordassociation.org.

be noted that a window size of ± 2 applied after elimination of the function words is comparable to a window size of ± 4 applied to the original texts (assuming that roughly every second word is a function word).

Based on the window size of ± 2 , we computed the co-occurrence matrix for the corpus. By storing it as a sparse matrix, it was feasible to include all of the approximately 375,000 lemmas occurring in the BNC.

Although word associations can be successfully computed based on raw word co-occurrence counts, the results can be improved when the observed co-occurrence-frequencies are transformed by some function that reduces the effects of absolute word frequency. As it is well established, we decided to use the log-likelihood ratio (Dunning, 1993) as our association measure. It compares the observed co-occurrence counts with the expected co-occurrence counts, thus strengthening significant word pairs and weakening incidental word pairs. In the remainder of this paper, we refer to co-occurrence vectors and matrices that have been transformed this way as association vectors and matrices.

2.2 Results and evaluation

To compute the associations for a given stimulus word, we look at its association vector as computed in the way described above, and rank the words in the vocabulary according to association strength. Table 2 (right two columns) exemplifies the results for the stimulus word *cold*.² For comparison, the left two columns list the responses from the EAT, and words occurring in both lists are printed in bold. It can be seen that especially the test persons' most frequent responses are predicted rather well in the simulation: Among the top eight experimental responses six can be found among the computed responses.

Surprisingly, although the system solely relies on word co-occurrences, it predicts not only syntagmatic but also paradigmatic associations (e.g. not only *cold* → *ice* but also *cold* → *hot*; cf. de Saussure, 1916; Rapp 2002).

We conducted a straightforward evaluation of the results. It is based on lemmatized versions of both the British National Corpus and, as this is the quasi standard for evaluation in related work, the Kent & Rosanoff (1910) subset of the Edinburgh Associative Thesaurus which comprises 100 words.

For altogether 17% of the stimulus words, the system produced the primary associative response, which is the most frequent response as produced by the human subjects.³ In comparison, the average participant in the Edinburgh Associative Thesaurus (Kiss et al. 1973) produced 23.7% primary responses to these stimulus words. This means that the system performs reasonably but not quite as well as the test persons.

² In order not to lose information in contrast to all other results presented in this paper this table is based on an unlemmatized corpus and an unlemmatized association norm.

³ Wettler et al. (2005) report somewhat better results by additionally taking advantage of the observation that test persons typically answer with words from the mid frequency range. As it is not clear how this effects the results when computing associations for several given words, we did not do this in the current paper.

Observed responses	# of subjects	Computed responses	# of subjects
hot	34	water	5
ice	10	hot	34
warm	7	weather	0
water	5	wet	3
freeze	3	blooded	0
wet	3	ice	10
feet	2	air	0
freezing	2	winter	2
nose	2	freezing	2
room	2	bitterly	0
sneeze	2	damp	0
sore	2	wind	0
winter	2	warm	7
arctic	1	felt	0
bad	1	war	1
beef	1	night	0
blanket	1	icy	0
blow	1	heat	1
cool	1	shivering	0
dark	1	cistern	0
drink	1	feel	0
flu	1	windy	0
flue	1	stone	0
frozen	1	morning	0
hay fever	1	shivered	0
head	1	eyes	0
heat	1	clammy	0
hell	1	sweat	0
ill	1	blood	0
north	1	shower	0
often	1	rain	0
shock	1	winds	0
shoulder	1	tap	0
snow	1	dry	0
store	1	dark	1
uncomfy	1	grey	0
war	1	hungry	0

Table 2: Comparison between observed and computed associative responses to the stimulus word *cold* (matching words in bold; no lemmatization; capitalized words transferred to lower case).

3 Computing reverse associations

3.1 Problem

Having seen that word associations to single stimulus words can be computed with a quality similar to that achieved by human subjects, let us now turn to the main question of this paper, namely whether it is also possible to reverse the task, that is to compute a stimulus word from its associations.

Let us look at an example: According to the Edinburgh Associative Thesaurus, the top three most frequent responses to *clown* are *circus* (produced by 26 out of 93, i.e. 28% of the test persons), *funny* (9% of the test persons) and *nose* (8% of the test persons). The question is now: Given only the three words *circus*, *funny* and *nose*, is it possible to determine that their common stimulus word is *clown*? And if it is possible, what would be the quality of the results?

The above is an illustrative example, but other cases are often more difficult. To give a feeling for the difficulty of the task, let us provide a few more examples involving varying numbers of given words, with the solutions provided in Table 4:

apple, juice → ?
water, tub, clean → ?
grass, blue, red, yellow → ?
drink, gin, bottle, soda, Scotch → ?

3.2 Procedure

Our first idea on how to compute the stimulus given the responses was to look at the associations of the responses, and to determine their intersection. But in preliminary experiments we found out that this does not work well. The reason appears to be asymmetry in word association. But what do we mean by asymmetry in this context?

The co-occurrence counts which we extract from the corpus are symmetric, because whenever word A co-occurs with word B, word B also co-occurs with word A. Whether an association matrix computed from the co-occurrence matrix is also symmetric depends on the association measure used. But even in the case of symmetric weights, associations can still be asymmetric. Let us illustrate this using Figure 2. This is the graphical equivalent of a symmetric association matrix.⁴ As can be seen, the strongest association to *blue* is *black*. But the opposite is not true: The strongest association to *black* is not *blue* as *black* has an even stronger association to *white*.

⁴ In the asymmetric case we would require two directed connections between each pair of nodes.

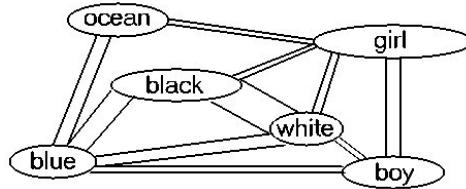


Figure 2: Associative lexical network with symmetric weights.

To give an idea about the situation in the EAT: Not considering multiword units, the EAT comprises 8210 stimulus words and likewise 8210 primary responses. However, there is not a complete overlap between these two vocabularies: Only 7387 words occur in both, which means that only for these words symmetry considerations are possible. Of these 7387 cases, 63% of the responses were symmetric, and 37% were asymmetric. Table 3 shows some examples from the EAT for both types of associations.

Symmetric Associations			Asymmetric Associations		
Stimulus	PR on Stimulus	PR on Response	Stimulus	PR on Stimulus	PR on Response
bed	sleep	bed	baby	boy	girl
black	white	black	bitter	sweet	sour
boy	girl	boy	comfort	chair	table
bread	butter	bread	cottage	house	home
butter	bread	butter	dream	sleep	bed
chair	table	chair	hand	foot	shoe
dark	light	dark	heavy	light	dark
girl	boy	girl	lamp	light	dark
hard	soft	hard	red	blue	sky
light	dark	light	sickness	health	wealth

Table 3: Examples for symmetric and asymmetric associations (PR = primary response).

Let us now return to our above example, namely *circus*, *nose*, *funny* → *clown*. Here, *circus* and *clown* are an example for the symmetric case. Both are each others primary associative responses in the EAT, so *circus* is the strongest association to *clown*, and likewise *clown* is the strongest association to *circus*. If this were always true, things would be straightforward. But this is not the case. For example, *clown* is strongly associated to *nose*, but *nose* is not strongly associated to *clown*. In the EAT, among 97 test person, given the stimulus word *nose*, none responded with *clown*. Likewise, given the word *funny*, among 98 test persons, again nobody answered with *clown*. So if we take the intersection of the associations to *circus*, *nose*, and *funny*,

clown would be out. This is why an approach based on intersecting associations does not work well.

Instead, like in word sense disambiguation and like in multiword semantics, it appears that we have to take contextual information into account.⁵ For example, in the context of *circus*, *nose* is clearly related to *clown*, but in the context of *doctor* it is not.

Such considerations resulted in the following approach: We utilize the observation that a stimulus word must have strong weights to all of its top associations, and that a strong association to only some of them does not suffice. Such a behavior can usually be put into practice by using a multiplication.

However, we do not multiply the association strengths, as the log-likelihood ratio has an inappropriate (exponential) value characteristic. This value characteristic has the effect that a weak association to one of the stimuli can easily be overcompensated by a strong association to another stimulus, which is not desirable. Instead of multiplying the association strengths, we therefore multiply their ranks. This improves the results considerably.

These considerations lead us to the following procedure (cf. Rapp, 2008): Given an association matrix of vocabulary V containing the log-likelihood ratios between all possible pairs of words, to compute the stimulus word causing the responses a, b, c, \dots the following steps are conducted:

- 1) For each word in V (by considering its association vector) look up the ranks of words a, b, c, \dots in its association vector, and compute the product of these ranks ("Product-of-Ranks algorithm").
- 2) Sort the words in V according to these products, with the sort order such that the lowest value obtains the top rank (i.e. conduct a reverse sort).

Note that this procedure is somewhat time consuming as these computations are required for each word in a large vocabulary.⁶ On the plus side, the procedure is in principle applicable to any number of given words, and with increasing number of given words there is only a slight increase in computational load.

A minor issue is the assignment of ranks to words which have identical log-likelihood scores, especially in the frequent case of zero co-occurrence counts. In such cases, the assignment of almost arbitrary ranks within such a group of words could adversely affect the results. We therefore suggest assigning corrected ranks, which are to be chosen as the average ranks of all words with identical log-likelihood scores.

In principle the algorithm can also be used if there is only a single given word. However, this does not make much sense as the algorithm is computationally far more expensive than what we described in Section 2, and the results are typically worse for the reason that ranks do not allow as fine-grained distinctions as do association strengths. For example, given the word *white*, the algorithm might find several words in the vocabulary where *white* is on rank 1 (e.g. *black* and *snow*). But as (without further sophistication) no distinction is made between these, they will end up in arbi-

⁵ Further reflections on this may lead to the fundamental question whether asymmetry of word associations is the consequence of word ambiguity, or whether word ambiguity is the consequence of asymmetry of word associations.

⁶ Considerable time savings are possible by using an index of the non-zero co-occurrences.

trary order, without taking into account that the top rank of *black* is more salient than that of *snow*.⁷

On the other hand, if the number of given words gets large, depending on the application it can be helpful to introduce a limit to the maximum rank, thereby reducing the effects of statistical variation which is especially severe for the lower ranks. Note that for the current work we used a rank limit of 10,000. But the exact value is not critical because this usually has little impact if the focus is mainly on the top ranks, as is the case here.

3.3 Results and evaluation

To give an impression of the results when applying the above algorithm on various numbers of responses from the EAT, Table 4 lists some results. For example, the EAT lists *apple* and *juice* as the top responses when given the stimulus word *fruit*, but our algorithm, when provided with *apple* and *juice*, computes that *orange* would be the best stimulus. This is not as expected, but also has some plausibility. The expected stimulus *fruit* at least shows up on the 8th position of the computed list of words.

For a quantitative evaluation, like for the forward associations we consider only the Kent & Rosanoff (1910) subset of the EAT.⁸ We count in how many cases the expected word is ranked first in the list of computed words. This leads to conservative numbers as only exact matches are taken into account. For example, the last item in Table 4, where *whisky* instead of *whiskey* is on rank 1, would count as incorrect.

When predicting the stimuli from the associative responses, the question is how many of the responses should be taken into account, and in how far the quality of the results depends on the number of responses. To answer this question, we conducted the evaluation several times, each time with another number of given words (= EAT responses). There are three expectations:

- The more subjects have given a response, the more salient it is and the more helpful it should be for predicting the stimulus.
- Responses given by only one or very few subjects might be arbitrary and therefore not helpful for predicting the stimulus.
- Considering a larger number of salient responses should improve the results.

These expectations are confirmed by the results. Figure 3 shows the percentage of correctly computed stimuli depending on the number of top responses (from the EAT) that are taken into account. As can be seen, the quality of the results improves up to seven given words where it reaches 54% accuracy, and from then on degrades. This means that, on average, already the eighth response word is not helpful for determining the respective stimulus word.

⁷ In the EAT 57 and 40 subjects, respectively, responded with *white* for these two stimulus words; another example is *lily* where the primary associative response is also *white*, but is produced by only 19 subjects. In this case the next frequent response, namely *flower*, is very close as it is produced by 17 subjects.

⁸ Results for the full EAT are in preparation. It should be noted that the Kent & Rosanoff subset typically leads to relatively high accuracies as it mostly comprises familiar words with high corpus frequencies.

TOP 2 RESPONSES FROM EAT: apple (1385) juice (1613)
STIMULUS WORD FROM EAT: fruit (3978)
COMPUTED STIMULI: orange (2333), grape (273), lemon (1019), lime (612), pineapple (220), grated (423), apples (792), fruit (3978), grapefruit (113), carrot (359)
TOP 3 RESPONSES FROM EAT: water (33449), tub (332), clean (6599)
STIMULUS WORD FROM EAT: bath (415)
COMPUTED STIMULI: rinsed (177), bath (2819), soak (315), rinse (288), wash (2449), refill (138), rainwater (160), polluted (393), towels (421), sanitation (156)
TOP 4 RESPONSES FROM EAT: grass (4295), blue (9986), red (13528), yellow (4432)
STIMULUS WORD FROM EAT: green (10606)
COMPUTED STIMULI: green (10606), jersey (359), ochre (124), bright (5313), pale (3583), violet (396), purple (1262), greenish (136), stripe (191), veined (103)
TOP 5 RESPONSES FROM EAT: drink (7894), gin (507), bottle (4299), soda (356), Scotch (621)
STIMULUS WORD FROM EAT: whiskey (129)
COMPUTED STIMULI: whisky (1451), whiskey (129), tonic (511), vodka (303), brandy (848), Whisky (276), scotch (151), lemonade (229), poured (1793), gulp (196)

Table 4: Top ten computed stimuli for various numbers of given responses. Numbers in brackets refer to corpus frequencies in the BNC.

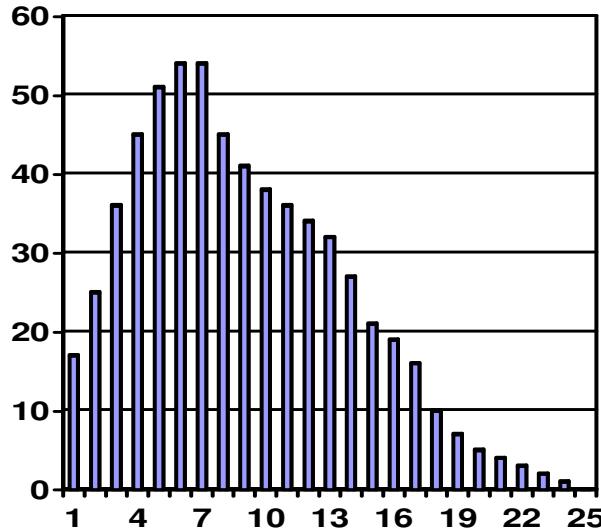


Figure 3: Percentage of correctly predicted stimuli (vertical axis) depending on the number of given words (horizontal axis).

Let us mention that we were positively surprised by the 54% performance figure, which is about three times as good as for forward association (first column in Figure 3). On the one hand, in the reverse association task there are several clues pointing to the same stimulus word. But on the other hand, the task seems non-trivial for humans, and typically there are several plausible options how the given words can disambiguate each other. For example, given *apple* and *juice* (see Table 4), the solution our system came up with, namely *orange*, seems quite as plausible as the expected solution *fruit*. However, in our evaluation *orange* is counted as wrong, and this is true for many others of the 46% incorrect results.

4 Summary, conclusions, and prospects

We introduced the product-of-ranks algorithm and showed that it can be successfully applied to the problem of computing associations if several words are given. To evaluate the algorithm, we used the EAT as our gold standard, but assumed that it makes sense to look at this data in the reverse direction, i.e. to predict the EAT stimuli from the EAT responses.

Although this is a task even difficult for humans, and although we applied a conservative evaluation measure which insists on an exact string matching between a predicted and gold standard association, our algorithm was able to do so with a success rate of up to 54%. We also showed that, up to a certain limit, with increasing number of given words the performance of the algorithm improves, and only thereafter degrades. This behavior was in line with our expectations because associative responses produced by only one or very few persons are often of almost arbitrary nature and therefore not helpful for predicting the stimulus word.

Given the notorious difficulty to predict experimental human data, we think that the performance of 54% is quite good, especially in comparison to the related work mentioned in the introduction (11,54%), and to the results on single stimuli (17%). But there is of course still room for improvement, without moving to more sophisticated (but also more controversial) evaluation methods which allow alternative solutions. We intend to advance from the product-of-rank algorithm to a product-of-weights algorithm. But this requires that we have a high quality association measure with an appropriate value characteristic. One idea is to replace the log-likelihood scores by their significance levels. Another is to abandon conventional association measures and move on to empirical association measures as described in Tamir & Rapp (2003). These do not make any presuppositions on the distribution of words, but determine this distribution from the corpus. In any case the current framework is well suited for measuring and comparing the suitability of any association measure.

Concerning applications, we see a number of possibilities: One is the tip-of-the-tongue problem, where a person cannot recall a particular word but can nevertheless think of some of its properties and associations. In this case, descriptors for the properties and associations could be fed into the system in the hope that the target word comes up as one of the top associations, from which the person can choose.

Another application is in information retrieval, where the system can help to sensibly expand a given list of search words, which is in turn used to conduct a search. A more ambitious (but computationally expensive) approach would be to consider the

(salient words in the) documents to be retrieved as our lists of given words, and to predict the search words from these using the product-of-ranks algorithm.

A further application is in multiword semantics. Here a fundamental question is whether a particular multiword expression is of compositional or of contextual nature. The current system can help us to provide a number of quantitative measures relevant for answering the following questions:

- 1) Can the components of a multiword unit predict each other?
- 2) Can each component of a multiword unit be predicted from its surrounding content words?
- 3) Can the full multiword unit be predicted from its surrounding content words?

The results on these questions might help us to answer the question regarding a multiword unit's compositional or contextual nature, and to classify various types of multiword units.

The last application we would like to propose here is natural language generation (or any application that requires this, e.g. machine translation or speech recognition). If in a sentence one word is missing or uncertain, we can try to predict this word by considering all other content words in the sentence (or a somewhat wider context) as our input to the product-of-ranks algorithm.

From a cognitive perspective, the hope is that such experiments might lead to some progress in finding an answer concerning a fundamental question: Is human language generation governed by associations, i.e. can the next content word of an utterance be considered as an association to the representations of the content words already activated in the speaker's memory?

Acknowledgments

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. I would like to thank Manfred Wettler for the previously published joint work which provides the basis for the current paper.

References

- Burnard, L.; Aston, G. (1998). *The BNC Handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.
- de Saussure, F. (1916/1996). *Cours de linguistique générale*. Paris: Payot.
- Church, K.W.; Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1), 22–29.
- Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1), 61–74.
- Griffiths, Thomas L.; Steyvers, Mark, Tenenbaum, Joshua B. (2007). Topics in semantic representation. *Psychological Review*, Vol. 114, No. 2, 211-244.

- Karp, D., Schabes, Y., Zaidel, M., Egedi, D. (1992). A freely available wide coverage morphological analyzer for English. In: *Proceedings of the 14th International Conference on Computational Linguistics* (COLING), Nantes, 950–955
- Kent, G.H., Rosanoff, A.J. (1910). A study of association in insanity. *American Journal of Psychiatry*, 67, 317–390.
- Kiss, G. R. (1975). An associative thesaurus of English: Structural analysis of a large relevance network. In: A. Kennedy, A. Wilkes (eds.): *Studies in Long Term Memory*. London: Wiley, 103–121.
- Kiss, G.R.; Armstrong, C.; Milroy, R.; Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Bailey, N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh University Press.
- James, W. (1890). *The Principles of Psychology*. New York: Holt. Reprinted New York: Dover Publications, 1950.
- Rapp, R. (1996). *Die Berechnung von Assoziationen*. Hildesheim: Olms.
- Rapp, R. (2008). The computation of associative responses to multiword stimuli. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (COGALEX at Coling 2008, Manchester). 102–109.
- Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. *Proceedings of the 19th International Conference on Computational Linguistics*, Taipeh, ROC, Vol. 2, 821–827.
- Russell, W.A., Meseck, O.R. (1959). Der Einfluß der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für experimentelle und angewandte Psychologie*, 6, 191–211.
- Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (ed.): *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 24. New York: Academic Press, 249–284.
- Smith, Kevin A.; Huber, David E.; Vul, Edward (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition* 128, 64–75.
- Tamir, R.; Rapp, R. (2003). Mining the web to discover the meanings of an ambiguous word. In: *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, Florida, 645–648.
- Wettler, M., Rapp, R. (1989). A connectionist system to simulate lexical decisions in information retrieval. In: R. Pfeifer, Z. Schreter, F. Fogelman, L. Steels (eds.): *Connectionism in Perspective*. Amsterdam: Elsevier, 463–469.
- Wettler, M.; Rapp, R.; Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics* 12(2), 111–122.

Can Human Association Norm Evaluate Latent Semantic Analysis?

Izabela Gatkowska², Michał Korzycki¹, Wiesław Lubaszewski¹

¹ AGH University of Science and Technology, Kraków, Poland
{lubaszew,korzycki}@agh.edu.pl

² Jagiellonian University, Kraków Poland
izabela.gatkowska@uj.edu.pl

Abstract. This paper presents a comparison of a word association norm created by a psycholinguistic experiment to association lists generated by algorithms operating on text corpora. We compare lists generated by the Church and Hanks algorithm and lists generated by LSA algorithm. An argument is presented on how those automatically generated lists reflect semantic dependencies present in human association norm, and that the future comparisons should take into account a deeper analysis of those human association mechanisms observed in the association list.

1 Introduction

Originally the LSA is a word/document matrix rank reduction algorithm, which extracts word co-occurrences in the frame of a text. As a result each word in the corpus is related to all co-occurring words and all texts in which it occurs. This makes a base for an associative text comparison. The applicability of the LSA algorithm is the subject of various types of research. From a text content comparison (Deerwester, Dumais, 1990) to an analysis of human association norm (Ortega-Pacheco, Arias-Trejo, Barron Martinez, 2012). But there is still little interest in studying the linguistic significance of LSA-made associations.

It seems to be obvious that a comparison of human association norm and LSA-made association list should be the base of the study. And we can find some preliminary studies based on such a comparison: (Wandmacher, 2005), (Wettler, Rapp, Sedlmeier, 2005), (Wandmacher, Ovchinnikova, Aleksandrov, 2008), the results of which show that the problem needs further investigation. It is worth noticing that all the types of research referred to, used a stimulus – response association strength to make a comparison. The point is that, if we compare association strength computed for a particular stimulus – response pair in association norms for different languages, we can find that association strength differs, e.g. *butter* is the strongest (0.54) response to stimulus *bread* in the Edinburgh Associative Thesaurus (EAT), but in the Polish association norm described below the association *chleb* ‘bread’ – *masło* ‘butter’ is not the strongest one (0.075). In addition one can observe that association strength may not distinguish a semantic and non-semantic association, e.g. *roof* 0.04, *Jack* 0.02, and *wall* 0.01, which are responses to the stimulus *house* in EAT. Therefore

we decided to test the LSA-made association lists against human association norms excluding association strength. We use for comparison the norm made by Polish speakers during a free word association experiment (Gatkowska 2012), hereinafter referred to as the author's experiment. Because the LSA uses a whole text to generate word associations, we also tested human associations against the association list generated by Church and Hanks algorithm (Church, Hanks, 1990), which operates on a sentence-like text window. We also used three different text corpora.

First, the paper describes a comparison of Polish and English human association lists. Then we describe the comparison of human association lists to lists that were generated automatically by the LSA and the Church and Hanks algorithm. Finally we shall discuss the results referring to earlier research.

2 Human semantic associations

2.1 Word Association Test

Rather early on, it was noted that words in the human mind are linked. The American clinical psychologists G. Kent and A. J. Rosanoff, perceived the diagnostic usefulness of an analysis of the links between words. In 1910, the duo created and conducted a test of the free association of words. They conducted research on 1000 people of varied educational backgrounds and professions, asking their research subjects to give the first thought that came into their minds as a result from a stimulus-words. Those research was supplied with 100 word-stimuli, (principally nouns and adjectives). The Kent-Rosanoff list of words was translated into several languages, in which this experiment was repeated, thereby enabling comparative research to be carried out. Word association research was continued by Palermo, Jenkins (1964), Postman, Keppel (1970), Kiss, Armstrong, Milroy, Piper (1973), Moss, Older (1996), Nelson, McEvoy, Schreiber (1998), and the repeatability of results allowed the number of research subjects to be reduced, while at the same time increasing the number of word-stimuli to be employed, for example 500 research subjects and 200 words (Palermo, Jenkins, 1964), or 100 research subjects and 8400 words (Kiss, Armstrong, Milroy, Piper, 1973). Research on the free association of words has also been conducted in Poland (Kurcz, 1967) and the results makes a basis for the experiment described below.

Computational linguistics also became involved in research on the free association of words, though at times these experiments didn't employ the rigors used by psychologists when conducting experiments, for example, those that permitted the possibility of providing several responses to an individual stimulus-word. (Schulte in Walde S., Borgwaldt S., Jauch R., 2012), or those that used word pairs as a stimulus (Rapp, 2008).

There exist some algorithms, which generate an association list on the basis of text corpora. But automatically generated associations were rather reluctantly compared with the results of psycho-linguistic experiments. The situation is changing, Rapp's results (2002) were really encouraging.

Finally, association norms start serving for different tasks, as for example information extraction (Borge-Holthoefer, Arenas, 2009) or dictionary expansion (Sinopalnikova, Smrz, 2004), (Budanitsky, Hirst, 2006).

2.2 The Author's Experiment

Some 540 students of the Department of Management and Social Communication at the Jagiellonian University participated in the free word association test as described in this article. A Polish version of the Kent-Rosanoff list of stimulus words, which was previously used by I. Kurcz was employed (Kurcz, 1967). After an initial analysis it was determined that we would employ as a stimulus, each word from the Kent-Rosanoff list, which grammatically speaking is a noun, as well as the five most frequent word associations for each of those nouns obtained in Kurcz's experiment (Kurcz, 1967). If given associations appeared for various words, for example, *white* for *doctor*, *cheese*, *sheep*, that word as a stimulus appeared only once in our experiment. The resulting stimulus list contained 60 words from the Kent-Rosanoff list, in its Polish version, as well as 260 words representing those associations (responses) which most frequently appeared in Kurcz's research. It therefore, is not an exact repetition of the experiment conducted 45 years ago.

The conditions of the experiment conducted, as well as the method of analyzing the results, have been modified. The experiment was conducted in a computer lab, with the aid of a computer system, which has been created specifically for the requirements of this experiment. This system presents a list of stimuli and then writes down associations in a data base. Instructions appeared on the computer screens of each participant, which in addition were read aloud by the person conducting the experiment. After the instructions were read, the experiment commenced, whereby a stimulus word appeared on the computer screen of each participant, and he wrote the first free association word which came to his mind - only one response was possible. When the participant wrote down his association, (or the time ran out for him to write down his association), the next stimulus word appeared on his screen, until the experiment was concluded. The number of stimulus-words as well as their order, was the same for all participants.

As a result we obtained 260 association lists, which consist of more than 16,000 associated words.

Association list derived from the experiment will be used to evaluate algorithm derived association lists. But first, we have to show how the human associations are comparable.

2.3 Comparison of Human Association Lists

We shall compare a Polish list derived from our experiment to a semantically equivalent English list derived from the Edinburgh Associative Thesaurus. To illustrate the problem we selected an ambiguous Polish word *dom*, which refers to the

English words *home* and *house*. Those lists will present words associated with their basic stimulus, and ordered in accordance to their strength of association. Due to the varied number responses (95 for *home* and *house* and 540 for *dom*) we will be using a more qualitative measure of similarity based on the rank of occurring words on them, rather than on a direct comparison of association strength. That list measure $LM_w(l_1, l_2)$, given two word lists l_1 and l_2 and a comparison window, which will be equivalent to the amount of words matching in l_1 and l_2 in a window of w words taken from the beginning of the lists.

In order to establish some basic expected levels of similarity, we will compare the list obtained in our experiment for the stimulus word *dom*, which meaning covers both English word *home* and *house*. First, each Polish association-word was carefully translated into English, then the lists automatically looked for identical words. Because words may differ in rank on the compared lists, the table includes the window size needed to match a word on both lists.

Table 1. Top 10 elements of the experiment lists for *dom* (author's experiment) and the EAT lists for *home* and *house*

<i>dom</i>	<i>home</i>	<i>house</i>
rodzinny (<i>adv. family</i>)	house	home
mieszkanie (<i>flat</i>)	family	garden
rodzina (<i>n.family</i>)	mother	door
spokój (<i>peace</i>)	away	boat
ciepło (<i>warmth</i>)	life	chimney
ogród (<i>garden</i>)	parents	roof
mój (<i>my</i>)	help	flat
bezpieczeństwo (<i>security</i>)	range	brick
mama (<i>mother</i>)	rest	building
pokój (<i>room</i>)	stead	bungalow

Those lists can be compared separately, but considering the ambiguity of *dom*, we can compare the list of association of *dom* with a list of interspersed (i.e. a list composed of the 1st word related to *home*, next to the 1st word associated with *house*, then the 2nd word related to *home* etc.) associations of both *home* and *house* lists coming from EAT.

Table 2. Comparison of the experiment list and the EAT lists. Matching words are shown for their corresponding window sizes w for the $LM_w(l_1, l_2)$ measure

<i>w</i>	<i>home+house</i> vs <i>dom</i>	<i>w</i>	<i>home</i> vs <i>dom</i>	<i>w</i>	<i>House</i> vs <i>dom</i>
3	family	3	family	3	family
6	garden	9	mother	6	flat
9	mother	18	cottage	6	garden

<i>w</i>	home+house vs <i>dom</i>	<i>w</i>	home vs <i>dom</i>	<i>w</i>	<i>House</i> vs <i>dom</i>
12	roof	24	garden	11	roof
14	flat	26	parents	14	room
18	building	35	peace	15	building
19	chimney	41	security	19	chimney
26	parents			21	cottage
30	room			30	mother
32	brick			32	brick
35	cottage			34	security
64	security			40	warm
65	peace			41	warmth
74	warm				
75	warmth				

The original, i.e. used for comparison human association list, is a list of words associated to a stimulus-word ordered by frequency of responses. Unfortunately, we can not distinguish automatically that words, which enter into semantic relation to the stimulus-word by frequency or by computed association strength, for example in the list associated to the word *table* a semantically unrelated *cloth* is substantially more frequent than *legs* and *leg*, which enter into ‘part of’ relation to the *table*. (Palermo, Jenkins, 1964). The described observation is language independent. The proposed method of comparison truncates from the resulting list language specific semantic associations, e.g. *home – house* and *house – home* the most frequent on EAT as well as all non-semantic associations, e.g. *home – office* or *house – Jack*. Each resulting list consists of words, each of which is semantically related to a stimulus-word. In other words, the comparison of the human association list will automatically extract a sub-list of semantic associations.

3 Algorithm Efficiency Comparison

3.1 The Corpora

In order to compare the association lists with the ones with a Latent Semantic Analysis, we have prepared three distinct corpora to train the algorithm. The first consists of 51.574 press notes of the Polish Press Agency and contains over 2.900.000 words. That corpus represents a very broad description of reality, but can be somehow seen as restricted to only a more formal subset of the language. This corpus will be referred to as PAP.

The second corpus is a fragment of the National Corpus of Polish (Przepiórkowski et al., 2011) with a size of 3363 separate documents spanning over 860.000 words. That

corpus is representative in the terms of the dictionary of the language, however the texts occurring in it are relatively random, in the sense that they are not thematically grouped or following some deeper semantic structure. This corpus will be referred to as the NCP.

The last corpus is composed of 10 short stories and one novel *Lalka* ("The Doll") by Bolesław Prus – a late XIX century novelist using a modern version of Polish similar to the one used nowadays. The texts are split into 10.346 paragraphs of over 300.000 words. The rationale behind this corpus was to try to model some historically deeply rooted semantic associations with such basic notions as *dom*. This corpus will be referred to in as PRUS.

All corpora were lemmatized using a dictionary based approach (Korzycki, 2012).

3.2 LSA Sourced Association Lists

Latent Semantic Analysis is a classical tool for extracting automatically similarities between documents, through dimensionality reduction. A term-document matrix is filled with weights corresponding to the importance of the term in the specific document (term-frequency/inverted document frequency in our case) and reduce via Singular Value Decomposition to a lower dimensional space called the concept space.

Formally, the term-document matrix X of dimension $n \times m$ (n terms and m documents), can be decomposed into U and V orthogonal matrices and Σ a diagonal matrix through singular value decomposition:

$$X = U \Sigma V^T \quad (0)$$

That, in turn can be represented through a rank k approximation of X in a smaller dimensionally space (Σ becomes a $k \times k$ matrix). We used an arbitrary rank value of 150 in our experiment.

$$X_k = U_k \Sigma_k V_k^T \quad (1)$$

This representation is often used to compare documents in this new space, but as the problem is symmetrical it can be used to compare words. The U_k matrix of dimensions $n \times k$ represents the model of words in the new k -dimensional concept space. We can thus compare the relative similarity of each word by taking the cosine distance between their representations.

The LSA sourced lists of associations is composed of the ordered list (by cosine distance) from the given word in a model build on each of the tree corpora as described above.

A crucial element in the application of Latent Semantic Analysis (Landauer et al, 2008), is determining k , the number of concepts that are used to project the data to the reduced k -dimensional concept space. As this parameter is a characteristic of the corpus, and in some degree of the specific application, in this case it has been determined experimentally. For each corpus (PRUS, NCP and PAP), an LSA model has been built

for a range of dimensions between 25 and 400 with an increment of 25. For each corpus, the dimension has been chosen as the one that gave the highest sum of matching words from 10 association lists in a window of 1000 words. The final results as presented in 3.4 correspond to a dimension of 75 for PRUS and NCP and 300 for PAP. The calculations were made using the gensim topic modeling library.

3.3 Association Ratio Based Lists

In order to evaluate the quality of the relatively advanced mechanism of Latent Semantic Analysis, we will compare its efficiency to the *association ratio* as presented in (Church, Hanks, 1989), with some minor changes related to the nature of the processed data. For two words x and y , their association ratio $f_w(x,y)$ will be defined as the number of times y follows or precedes x in a window of w words. The original association ratio was asymmetric, considering only words y following the parameter x . This approach will however fail in the case of texts that are written in languages with no strict word ordering in sentences (Polish in our case) where syntactic information is represented through rich inflection rather than through word ordering. We will use the same value for w as is in Church and Hanks (1989) that suggested a value of 5. This measure can be seen as simplistic in comparison with LSA, but as the results will show, useful nonetheless.

3.4 Lists Comparison

First we have to compare the list obtained automatically from the three corpora for the word *dom* (*home/hose*) with the reference list, i.e. human association list obtained from human subjects in the author's experiment. The comparison will be presented in terms of $LM_w(l_1, l_2)$ for l_1 being the human association list and l_2 being the lists obtained through LSA similarities and the *association ratio* f_s as described above. In the comparison we shall apply to the reference list, the three different window sizes.

To begin, we shall compare the full human association list that is 151 words long, to the lists generated by the algorithms described above. We restrict arbitrarily, the length of automatically generated lists to 1000 words.

Table 3. $LM_w(l_1, l_2)$ values for different w , for different l_2 from various list sources , l_1 being the human experiment result list

W	PRUS f_s	PAP f_s	NCP f_s	PRUS LSA	PAP LSA	NCP LSA
10	0	0	0	0	0	0
25	0	0	0	0	0	0
50	2	1	2	0	0	0
75	2	4	3	1	0	1
100	4	7	9	2	0	2

<i>W</i>	PRUS f_5	PAP f_5	NCP f_5	PRUS LSA	PAP LSA	NCP LSA
150	11	14	17	2	2	2
300	19	24	30	2	6	3
600	34	25	41	4	11	12
1000	36	43	49	7	13	18

That can be seen as excessive as it contains also a random association of low interest to us - the lists obtained through EAT and the author's list comparison contain only 15 words.

Then we will restrict the human association list to only the first 75 words – that was also the length needed to obtain the combined list for *home* and *house* from the EAT.

Table 4. $LM_w(l_1, l_2)$ values for different w , for different l_2 from various list sources , l_1 being the human experiment result list restricted to 75 entries

<i>W</i>	PRUS f_5	PAP f_5	NCP f_5	PRUS LSA	PAP LSA	NCP LSA
10	0	0	0	0	0	0
25	0	0	0	0	0	0
50	2	0	2	0	0	0
75	2	4	3	1	0	1
100	3	5	8	2	0	1
150	8	9	10	2	1	1
300	11	15	21	2	5	1
600	21	23	30	4	7	5
1000	22	28	33	5	9	6

As can be seen, automatically generated association lists match some part of the human association list only if we use a large window size. Secondly, we can observe that Church and Hanks algorithm seems to generate a list that is more comparable to a human derived list.

The shorter word list in the EAT (*house*) contains 42 words. The 40 words is the window size, which applied to the author's list, allow us to find all the elements common to the EAT *home/house* combined list and author's experiment list for *dom*. Therefore we shall use a 40-word window for comparison.

Table 5. $LM_w(l_1, l_2)$ values for different w , for different l_2 from various list sources , l_1 being the human experiment result list restricted to first 40 entries

<i>W</i>	PRUS f_5	PAP f_5	NCP f_5	PRUS LSA	PAP LSA	NCP LSA
10	0	0	0	0	0	0
25	0	0	0	0	0	0
50	2	0	2	0	0	0
75	2	4	3	1	0	1

<i>W</i>	PRUS f_3	PAP f_3	NCP f_3	PRUS LSA	PAP LSA	NCP LSA
100	3	5	7	1	0	1
150	7	9	9	1	0	1
300	8	9	17	1	4	1
600	15	16	22	2	6	5
1000	16	20	22	3	6	6

As we can see this window size seems to be optimal, because it reduces substantially – if compared to the full list – the non-semantic associations for both algorithms.

Finally we have to test automatically generated lists against the combined human association list, i.e. list which consists of words, which are present both in the author's list and the EAT lists, presented in Table 2.

Table 6. $LM_w(l_1, l_2)$ values for different w , for different l_2 from various list sources , l_1 being the human experiment result list restricted to words that are present in both the authors and the EAT experiment, see Table 2

<i>W</i>	PRUS f_3	PAP f_3	NCP f_3	PRUS LSA	PAP LSA	NCP LSA
10	0	0	0	0	0	1
25	0	0	0	0	0	1
50	0	0	1	0	0	1
75	0	1	3	0	0	1
100	0	3	3	0	0	2
150	3	4	5	0	0	2
300	4	8	5	0	1	2
600	8	12	9	0	2	3
1000	10	12	12	2	2	3

Those results show a tendency similar to that observed during the test of human association list in full length. First, the window size influences the matching number. The second observation is also similar: the list generated by the Church and Hanks algorithm matches better the human association list - it matches 10 or 12 out of 15 words semantically related to the stimulus.

To learn more, we repeated a comparison over a wider range of words. We selected 8 words: *chleb* (bread), *choroba* (disease), *światło* (light), *głowa* (head), *księżyca* (moon), *ptak* (beard), *woda* (water), *żołnierz* (soldier). Then we used the described method to obtain a combined list for the author's experiment and the EAT.

Table 7. $LM_w(l_1, l_2)$ values for different word stimuli, different w , for different l_2 from various list sources , l_1 being the human experiment result list restricted to entries in both the authors and the EAT experiment

<i>Word</i>	<i>w</i>	PRUS f_3	PAP f_3	NCP f_3	PRUS LSA	PAP LSA	NCP LSA
bread	25	0	1	0	0	1	1
	100	0	4	2	0	1	1

<i>Word</i>	<i>w</i>	PRUS f_s	PAP f_s	NCP f_s	PRUS LSA	PAP LSA	NCP LSA
	1000	1	8	3	0	2	2
disease	25	0	1	0	0	0	0
	100	1	3	5	0	0	0
	1000	1	9	8	1	7	2
light	25	1	1	0	0	1	0
	100	3	4	3	1	1	0
	1000	3	5	3	4	5	2
head	25	1	0	2	0	1	1
	100	1	2	4	0	1	1
	1000	3	6	6	1	2	3
moon	25	0	3	3	1	0	2
	100	3	4	5	1	0	3
	1000	3	4	6	4	2	5
bird	25	1	2	1	1	0	1
	100	2	4	2	1	0	2
	1000	2	5	7	4	3	3
water	25	0	1	2	1	1	0
	100	0	4	6	2	3	2
	1000	4	8	10	3	5	6
soldier	25	2	2	2	2	1	3
	100	2	5	5	2	6	3
	1000	2	12	9	3	10	4

The table below contains similar comparison, but without restricting the association list to words contained in both experiments.

Table 8. $LM_w(l_1, l_2)$ values for different word stimuli, different w , for different l_2 from various list sources , l_1 being the unrestricted human experiment result list

<i>Word</i>	<i>w</i>	PRUS f_s	PAP f_s	NCP f_s	PRUS LSA	PAP LSA	NCP LSA
bread	25	1	1	2	0	1	2
	100	2	5	6	1	2	5
	1000	4	19	12	3	4	9
disease	25	0	1	1	0	1	0
	100	1	3	7	0	2	0
	1000	3	13	14	1	13	8
light	25	2	1	1	1	1	0
	100	6	6	4	3	1	0
	1000	11	15	9	10	9	3
head	25	3	1	3	0	3	1
	100	6	6	7	0	5	1
	1000	17	17	12	7	9	7
moon	25	1	4	6	1	0	2
	100	5	5	11	1	1	4

<i>Word</i>	<i>w</i>	PRUS f_s	PAP f_s	NCP f_s	PRUS LSA	PAP LSA	NCP LSA
bird	1000	5	9	15	7	5	12
	25	1	8	2	2	0	2
	100	3	9	5	3	2	2
water	1000	5	13	19	8	9	9
	25	1	2	3	1	1	1
	100	3	7	8	2	4	3
soldier	1000	9	20	21	10	9	15
	25	1	5	4	1	2	3
	100	2	11	9	4	7	6
	1000	3	25	22	9	20	11

As can be seen, the values in the columns corresponding to the f_s algorithm are clearly better than the corresponding LSA values, regardless of the size of the human lists.

4 Conclusion

If we look at our results, we may find that in general they are comparable with the results of related research of Wandmacher (Wandmacher, 2005) and (Wandmacher, Ovchinnikova, Aleksandrov, 2008). Generally speaking the LSA algorithm generates an association list, which contains only a small fraction of the semantic relations, which are present in the human association norm. Surprisingly, the Church and Hanks algorithm does much better, which suggests that the problem of how the LSA-made associations relate to the human association norm should be investigated more carefully. The first suggestion may be derived from (Wettler, Rapp, Sedlmeier, 2005) – we have to learn more about the relation between the human association norm and the text to look for a method more appropriate than a simple list comparison. A second suggestion may be derived from an analysis of the human association list. It is well known that such a list consists of responses, which are semantically related to the stimulus, responses which reflect pragmatic dependencies and so-called ‘clang responses’. But within this set of semantically related responses one can find more frequent direct associations, i.e. such as those which follow a single semantic relation, e.g. ‘whole – part’: *house – wall* and not so frequent indirect associations like: *mutton* (baranina) – *horns* (rogi), which must be explained by a chain of relations, in our example: ‘source’ relation *mutton* (baranina) – *ram* (baran), followed by ‘whole – part’ relation *ram* (baran) – *horns* (rogi) or the association: *mutton* (baranina) – *wool* (wełna), explained by a ‘source’ relation *mutton* (baranina) – *ram* (baran), followed by ‘whole – part’ *ram* (baran) – *fleece* (runo), which is followed by a ‘source’ relation’ *fleece* (runo) – *wool* (wełna). These association chains suggest that some associations are based on a semantic network, and it would be very interesting to test the LSA associating mechanism against these indirect associations.

Acknowledgement

This research was partially supported by EC grant FP7-218086, the INDECT project.

References

- Borge-Holthoefer J., Arenas A., 2009, *Navigating word association norms to extract semantic information*, in: Taatgen N., van Rijn H., Proceedings of the 31st Annual Conference of the Cognitive Science Society, Groningen.
- Budanitsky A., Hirst G., 2006, *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics 32.1: 13-47.
- Church K. W., Hanks P., 1990, „Word Association Norms”, *Mutual Information, and Lexicography. Computational Linguistics*, t. 16, 1, p.22-29.
- Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R., 1990, *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science 41 (6): 391–407.
- Gatkowska I., 2012, *Jak słowa łączą się z sobą w umyśle użytkowników*, Tertium Conference, Krakow, 2012.
- Kent G., Rosanoff A. J., 1910, *A study of association in insanity*, American Journal of Insanity 67 (37-96), p. 317-390.
- Kess J. F., 1992, *Psycholinguistics: Psychology, linguistics and the study of natural language*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Kiss G. R., Armstrong C., Milroy R., Piper J., 1973, *An associative thesaurus of English and its computer analysis*. in: The Computer and Literary Studies red. Aitken, A.J., Bailey, R.W. Hamilton-Smith, N., Edinburgh University Press.
- Korzycki, M., 2012, *A dictionary based stemming mechanism for Polish* NLPCS 2012, p. 143–150
- Kurcz I., 1967, *Polskie normy powszechności skojarzeń swobodnych na 100 słów z listy Kent-Rosanoffa*, Studia Psychologiczne, t.VIII, red. T. Tomaszewski, Wrocław-Warszawa-Kraków, p.122- 255.
- Landauer T. K., Dumais S. T., Latent Semantic Analysis, Scholarpedia, 3(11):4356, 2008.
- Moss H., Older L., 1996, *Birkbeck word association norms*, Psychology Press.
- Nelson D. L., McEvoy C. L., Schreiber T. A., 1998, The University of South Florida word association, rhyme, and word fragment norms.
- Ortega-Pacheco D., Arias-Trejo N., Barron Martinez, J. B., 2012, *Latent Semantic Analysis Model as a Representation of Free-Association Word Norms*, 11th Mexican International Conference on Artificial Intelligence, MICAI 2012 , Puebla, p. 21-25
- Palermo D. S., Jenkins J. J., 1964, *Word Associations Norms: Grade School through College*, Minneapolis.
- Postman L. J., Keppel G., 1970, *Norms of word association*, Academic Press.
- Przepiórkowski A., Bańko M., Górska R., Lewandowska-Tomaszczyk B., Łaziński M., Pęzik P., 2011, *National Corpus of Polish*. In: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 259–263, Poznań, Poland.

- Rapp R., 2002, *The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches*, Proceedings of the 19th International Conference on Computational Linguistics, Taipei.
- Rapp R., 2008, *The Computation of Associative Responses to Multiword Stimuli*, Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008): Coling 2008, p. 102–109. Manchester,
- Sinopalnikova A., Smrz P., 2004, *Word Association Thesaurus as a Resource for extending Semantic Networks*, Proceedings of the International Conference on Communications in Computing, CIC '04, Las Vegas, Nevada, USA, p. 267-273.
- Schulte im Walde S., Borgwaldt S., Jauch R., 2012, *Association Norms of German Noun Compounds*, in: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul.
- Wandmacher, T., 2005, *How semantic is Latent Semantic Analysis*, Proceedings of TALN/RECITAL 5 .
- Wandmacher T., Ovchinnikova E., Alexandrov T., 2008 *Does Latent Semantic Analysis reflect human associations* , In Proceedings of the ESSLLI Workshop on Distributational Lexical Semantics
- Wettler M., Rapp R., Sedlmeier P., 2005, *Free word associations correspond to contiguisties between words in text*, Journal of Quantitative Linguistics, 12(2/), p. 111 – 122.

A Cognition-Oriented Approach to Fundamental Frequency Estimation

Ulrike Glavitsch and Klaus Simon

EMPA, Swiss Federal Laboratories for Materials Science and Technology
Ueberlandstrasse 129, 8600 Dübendorf, Switzerland
ulrike.glavitsch@empa.ch, klaus.simon@empa.ch

Abstract. This paper presents an efficient, two-phase fundamental frequency detection algorithm in the time-domain. In accordance with the human cognitive process it first computes base fundamental frequency estimates, which are verified and corrected in a second step. The verification step proceeds from high-energy stable segments, where reliable estimates are expected, to lower-energy regions. Irregular cases are handled by computing a series of fundamental frequency variants that are evaluated for highest plausibility, in analogy with the hypothesis testing principle of human thinking. As a proof of concept, the algorithm was evaluated on a clean speech database where it shows significantly lower error rates than a comparable reference method.

1 Introduction

The fundamental frequency F_0 plays an important role in human speech perception and is used in all fields of speech research. For instance, the human brain is supposed to evaluate the positions of formants with respect to F_0 [1], and accurate estimates of F_0 are a prerequisite for prosody control in concatenative speech synthesis [2].

Fundamental frequency detection has been an active field of research for more than forty years. Early methods used the autocorrelation function [3], cepstral analysis [4] and inverse filtering techniques [5] for F_0 detection. In most of these approaches, threshold values are used to decide whether a frame is assumed to be voiced or unvoiced. More advanced algorithms incorporate a dynamic programming stage to calculate the F_0 contour based on frame-level F_0 estimates gained from either a conditioned linear prediction residual [6] or a normalized cross correlation function (NCCF) [7]. In the last decade, techniques like pitch-scaled harmonic filtering (PSHF) [8], Nonnegative Matrix Factorization (NMF) [9, 10] and time-domain probabilistic approaches [11] have been proposed. These achieve low error rates and high accuracies but at a high computational cost - either at run-time or during model training. These calculative approaches generally disregard the principles of human cognition and the question is whether F_0 estimation can be performed equally well or better by considering these.

In this paper, we propose an F_0 estimation algorithm based on the elementary appearance and inherent structure of the human speech signal. A period, i.e. the

inverse of F_0 , is primarily defined as the distance between two maximum or two minimum peaks, and we use the same term to refer to the speech section between two such peaks. The speech signal can be divided into *stable* and *unstable* segments. Stable segments are those regions with a quasi-constant energy or quasi-flat envelope whereas unstable segments exhibit significant energy rises or decays. On stable segments, the F_0 periods are mostly regular, i.e. the sequence of maximum or minimum peaks is more or less equidistant, whereas the F_0 periods in unstable regions are often shortened, elongated, doubled, or may show little similarity with their neighboring periods. Speech signals are highly variable and such special cases occur relatively often. Thus, it makes sense to first compute F_0 estimates in stable segments and use this knowledge to find those of unstable segments in a second step. The F_0 estimation method for stable segments is straight-forward as regular F_0 periods are expected. The F_0 estimation approach for unstable segments, instead computes variants of possible F_0 continuation sequences and evaluates them for highest plausibility. The variants reflect the regular and all the irregular period cases and are calculated using a peak look-ahead strategy. We denote this F_0 estimation method for unstable segments as F_0 propagation since it computes and verifies F_0 estimates by considering previously computed values.

We regard the proposed algorithm as cognition-oriented inasmuch as it incorporates several principles of human cognition. First, human hearing is also two-stage process. The inner ear performs a spectral analysis of a speech section, i.e. different frequencies excite different locations along the basilar membrane and as a result different neurons with characteristic frequencies [12]. This spectral analysis delivers the fundamental frequency and the harmonics. The brain, however, then checks the information delivered by the neurons, interpolating and correcting it where necessary. Our proposed F_0 estimation algorithm performs in a similar way, in that the F_0 propagation step proceeds from regions with reliable F_0 estimates to ones where F_0 is not clearly known yet. We have observed that F_0 is very reliably estimated on high-energy stable segments, which typically represent vowels. Thus, we always compute F_0 for unstable segments by propagation from high-energy stable segments to lower-energy regions. Next, we have adopted the hypothesis testing principle of human thinking [13] for generating variants of possible F_0 sequences and testing them for the detection of F_0 in unstable segments. Lastly, human cognition uses context to decide a situation. For instance, in speech perception humans bear the left and right context of a word in mind if its meaning is ambiguous. In an analogous way, our algorithm looks two or more peaks ahead to find the next valid maximum or minimum peak for a given F_0 hypothesis. Special cases in unstable segments can often not be disambiguated by just looking a single peak ahead.

The resulting algorithm is very efficient, thoroughly extensible, easy to understand and has been evaluated on a clean speech database as a proof of concept. Recognition rates are clearly better than those of a reference method that uses cross-correlation functions and dynamic programming. In addition, it delivers a

segmentation of the speech signal into stable and unstable segments that may be useful for an automatic speech recognition component.

The outline of the paper is as follows. Section 2 describes the preprocessing steps of peak detection and energy computation. Section 3 presents the F_0 computation on stable segments. Section 4 describes the F_0 propagation stage. Section 5 outlines the post-processing step of computing F_0 on entirely unstable voiced segments. The evaluation of the algorithm is presented in Section 6. Finally, we draw conclusions and give an outlook for future work in Section 7.

2 Preprocessing

The first preprocessing step is the extraction of signal peaks. A peak is defined as either a local minimum or a local maximum in the sequence of signal samples. For each peak p , we maintain a triple of values $\langle x, y, c \rangle$ where x and y are the peak coordinates and c is the peak classification - either a minimum or a maximum peak. In a second step, we compute the mean energies for all signal frames. A frame is a small section of the signal where successive frames overlap by some extent. We selected a frame length of 20 ms and an overlap length of 10 ms. For periodic signal parts, the mean energy must be computed on an integer multiple of a period to be meaningful. However, as the period of a frame is not known at this point in time, we therefore compute the mean energy on a scale of window lengths each of which corresponds to a different period length. An optimization step then finds the best window length for each frame. This procedure is similar to pitch-scaled harmonic filtering (PSHF) [8] where an optimal window length is calculated for finding harmonic and non-harmonic spectra. The window lengths are selected such that periods of F_0 between 50 and 500 Hz roughly fit a small number of times into at least one of these lengths. The selected window lengths correspond to fundamental frequencies of 50, 55, 60, ..., 95 Hz. Each window length is centered around the frames middle position. The optimal window length is the one where the mean energies of a small number of frames around the frames middle position show the least variation.

3 F_0 Estimation on Stable Segments

The estimation of the fundamental frequency is first performed on stable voiced segments, i.e. voiced speech sections with a quasi-constant energy. For this purpose, the speech signal is segmented into voiced and unvoiced parts and into stable and unstable sections within the voiced regions. The method to estimate F_0 on stable segments is relatively straight-forward as we mainly expect regular periods. The F_0 estimates of stable segments are grouped into sequences of roughly equal F_0 values in order to provide anchor points for the F_0 propagation described in Section 4.

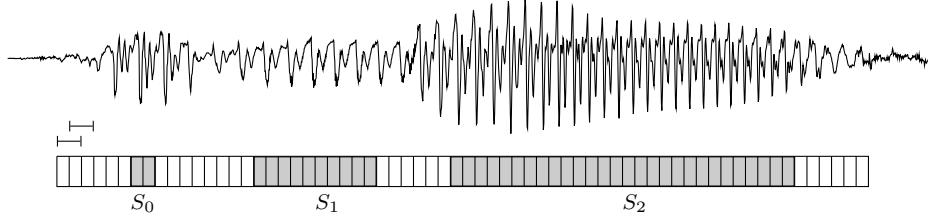


Fig. 1. Overlapping frames of voiced segment of speech signal "the north" uttered by a male speaker which contains three stable segments S_0 , S_1 and S_2 of lengths 1, 9 and 27.

3.1 Segmentation of Speech Signal

A signal frame is voiced if its mean energy exceeds a certain threshold, the absolute height of the frames maximum or minimum peak is above a given level, and the number of zero crossings is greater or equal to some configurable number. Consecutive voiced frames represent a so-called voiced segment. A voiced frame is classified as stable if its mean energy, as computed in Section 2, does not deviate more than a given percentage from the mean energy of both the previous and the next frame. The current value is 50 %, i.e. we allow some energy deviation between neighboring frames but not too much. Consecutive sequences of stable frames form stable segments. In Fig. 1, a voiced segment of a speech signal with stable segments S_0 , S_1 and S_2 is depicted.

3.2 F_0 Estimation Method

The F_0 estimation method for stable segments finds a quadruple of peaks $P = \langle p_L, p_0, p_1, p_R \rangle$ of either maximum or minimum peaks p_i , $i = L, 0, 1, R$, such that the center position of the frame is between p_0 and p_1 . The F_0 estimate is the inverse of the mean of the period lengths found in P , i.e. the mean of the distances between peaks p_L and p_0 , p_0 and p_1 as well as p_1 and p_R . The tuple P is selected among a series of possible candidate peak tuples according to some similarity score. Furthermore, it is checked whether the peak tuple is not a multiple of the supposedly true F_0 period, otherwise, a different peak tuple is selected. In the following, we describe the algorithm to find such a peak tuple P for each stable frame.

We start by finding the peak in the frame that has the highest absolute value. We then look for candidate peaks that have a similar absolute height and whose distance from the highest peak is within the permissible range of period lengths. The search for candidate peaks is performed in the direction of the center position of the frame. Given a peak pair of the highest absolute peak and a candidate peak, the algorithm looks for the peaks to the left and right of the given pair to complete the quadruple. We select the peak with the highest absolute value above some threshold and within a tolerance range on the time axis to the left and the right of the candidate peak pair. The peak tuple P may

reduce to a triple peak sequence if such a peak at one side cannot be found. Each such candidate peak quadruple or peak triple is scored and the tuple with the highest score is selected as the tentatively best candidate.

The proposed score measures the uniformness of the peaks in the peak tuple with respect to their distances and their absolute heights. The score s for peak tuple $P = \langle p_L, p_0, p_1, p_R \rangle$ is the product of partial scores s_x and s_y . The value s_x measures the equality of the peak intervals whereas s_y is a measure of the sameness of the absolute peak heights. The partial score s_x is defined as $1 - a$, where a is the root of the mean square difference between the peak distances at the tuple edges from the peak distance of the two middle peaks p_0 and p_1 . Similarly, partial score s_y is given as $1 - b$, where b is the root of the mean square difference of the absolute peak heights from the maximum absolute peak height. The equations below show how the score s is computed for a peak quadruple as defined above. The formulas are easily adapted for tuples with only three peaks.

$$s = s_x s_y \quad (1)$$

The partial score s_x is defined as follows:

$$s_x = 1 - a \quad (2)$$

$$a = \sqrt{(b_0^2 + b_1^2)/2} \quad (3)$$

$$b_0 = \frac{d_1 - d_0}{d_1}, \quad b_1 = \frac{d_1 - d_2}{d_1} \quad (4)$$

$$d_0 = x_0 - x_L, \quad d_1 = x_1 - x_0, \quad d_2 = x_R - x_1. \quad (5)$$

The value x_i , $i = L, 0, 1, R$ refers to the x-coordinate of peak p_i as mentioned in Section 2.

The partial score s_y is given by:

$$s_y = 1 - b \quad (6)$$

$$b = \sqrt{\frac{1}{4}(g_L^2 + g_0^2 + g_1^2 + g_R^2)} \quad (7)$$

$$g_i = (|y_i| - y_{max})/y_{max}, \quad i = L, 0, 1, R \quad (8)$$

$$y_{max} = max(|y_i| \mid i = L, 0, 1, R). \quad (9)$$

Similarly, y_i denotes the peak height of peak p_i in tuple P , $i = L, 0, 1, R$. The score s delivers exactly 1 if the peak heights and peak intervals are equal and less than 1 if they differ.

The peak tuple with the highest score may be a multiple of the true period or it may also be half of a period if the signal has a strong first harmonic. The case of a multiple period candidate is checked by testing the existence of equidistant partial peaks within the peak pair. If such partial peaks are found, we look for a candidate peak tuple with the partial peak distance and install it as the currently best candidate. We then check whether the best candidate tuple is only half of

a true period, by comparing the normalized cross correlation function (NCCF) [7] of the currently best candidate tuple with the NCCF of the tuple with the double period. If the NCCF of the former is significantly smaller than that of the latter, we install the peak tuple with the double period as the best candidate.

The final step in the F_0 estimation of a stable frame finds the peak tuple in the center of the frame that has the same period length as the best candidate tuple. This is achieved by looking for peaks to either the left or right side of the best candidate in the distance of the period length until a peak tuple is found where the frame's center position is between the two middle peaks.

3.3 Equal Sections

The last step of this stage is the detection of sequences of roughly equal F_0 estimates within a stable segment. These sequences are referred to as *equal sections*. The F_0 estimates of the frames in an equal section must not deviate by more than a given threshold from the mean F_0 of the equal section. The longest such equal section with a minimum length of 3 is stored as the equal section of the stable segment. The remaining equal sections of the stable segment are maintained in a list for eventualities.

4 F_0 Propagation

The F_0 propagation is the second major stage of the proposed F_0 detection algorithm. Its purpose is to calculate and check F_0 estimates in regions where no reliable F_0 estimates exist. This mainly affects unstable regions, but also portions of stable regions where e.g. the F_0 estimates do not belong to an equal section. The main idea is that the F_0 propagation starts at the stable segment with the highest energy from where it proceeds to the regions to both its left and right side. It always progresses from higher-energy to lower-energy regions. Once a local energy minimum is reached, it continues starting from the next stable segment in propagation direction that is a local energy maximum. For the verification and correction of calculated F_0 estimates we have developed a peak propagation procedure that computes the most plausible peak continuation sequence given the peak tuple of a previous frame. The most plausible peak sequence is found by considering several variants of peak sequences that reflect the regular and irregular period cases. In the following, we describe the control flow of the F_0 propagation and explain the particular peak propagation procedure.

4.1 Control Flow

The propagation of F_0 estimates is performed separately for each voiced segment. The first step in this procedure is the definition of the propagation order and the propagation end points. The propagation starts with the stable segment that contains the frame with the highest mean energy in its equal section. From this equal section the propagation flows first to the left and then to the right side.

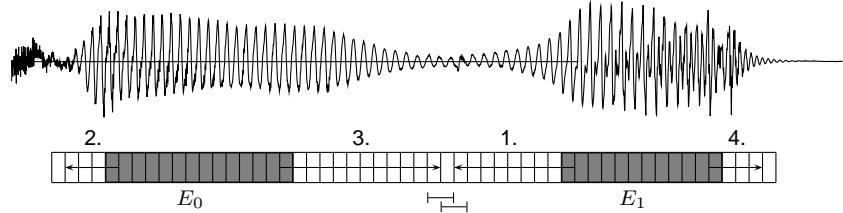


Fig. 2. Propagation order, directions and end points of a voiced segment "disputing which" (bold part) uttered by a female speaker. Propagation start and end points are marked at the center of the corresponding frames. Propagation starts from equal section E_1 as it has frames with higher energies than E_0 .

For each stable segment containing an equal section we define the right and left propagation end points. They are the start and end frame of the voiced segment if there is only one stable segment in the voiced segment. The propagation end point is a local energy minimum frame, or its direct neighbor frame if there is a local energy minimum region between two stable segments. Fig. 2 shows the propagation directions, order and end points of a voiced segment that contains two stable segments with equal sections E_0 and E_1 .

After identifying the propagation anchor points, directions and end points we compute candidate F_0 values for the unstable frames following the method presented in Section 3.2 but with a restricted allowable range for F_0 . We allow an F_0 range of more than an octave lower and two thirds of an octave higher than the mean F_0 of the equal section where the propagation starts. In contrast with the F_0 estimation method of Section 3.2, the check for multiple periods and strong harmonics is omitted, since it would hardly work in unstable regions with potentially strongly varying peak heights.

The core part of this stage is to check whether the F_0 estimate of a frame is in accordance with the F_0 of its previous frame and if not, to perform the peak propagation step (see Section 4.2) to find the most plausible peak continuation sequence from which the frame's actual F_0 estimate is derived. The F_0 estimate of a frame may deviate from the F_0 of its predecessor by a given percentage. As soon as the propagation end point is reached, we check whether the mean F_0 of the equal section of the next stable segment is similar to the mean F_0 of the most recently calculated values. Propagation continues normally from the next stable segment if this condition holds, otherwise the list of equal sections for eventualities (see Section 3.3) is searched for a better fitting equal section and the algorithm uses this as the new propagation starting point.

4.2 Peak Propagation

The peak propagation step computes a set of peak sequence variants that may follow the peak tuple of the previous frame and evaluates them for plausibility. Each peak sequence is computed by a look-ahead strategy for the next peak. In general, we look two peaks ahead before deciding on the next one.

The following peak sequence variants are considered:

- V1 (regular case): The peaks continue at about the same distance as the peaks in the previous frame.
- V2 (elongated periods): The periods are elongated and the peak distances become larger.
- V3 (octave jump down): The peaks follow at double distance as in the previous frame.
- V4 (octave jump up): The peaks follow at half the distance as in the previous frame.

These peak sequence variants are computed depending on the octave jump state of the previous frame. The octave jump state is maintained for each frame and its default value is 'none'. There are two additional values 'down' and 'up' for the state of F_0 that is an octave higher than normally, and the state of an F_0 estimate that has fallen by an octave. Variant V2 is used to detect extended periods that may not be captured by V1. However, V2 may easily deliver too large periods, e.g. in the case of spurious peaks, requiring additional checks. V3 is necessary to test the case of a sudden octave jump down but is only calculated in the case of an octave jump state of 'none'. V4 is considered only in an octave jump down state to check whether such a phase ends. Currently, for simplicity we forbid sequences of repeated octave jumps down and also sudden octave jumps up.

For each of these variants V1 to V4, we define interval ranges where subsequent peaks are expected. These ranges are defined relative to the last peak distance D , i.e. D is the distance between the last two peaks in propagation direction of the previously computed peak sequence. The peak sequence starts with the peak tuple of the previous frame and adds peaks to the left or to the right as propagation proceeds. Each new peak in the peak sequence is searched for in the given interval, while at the same time checking whether a peak exists in the interval that follows. Each such peak pair is scored by computing their mean absolute height. The first peak in the pair which achieves the highest such score is installed as the definite next peak in the peak sequence. The peak propagation stops as soon as the center frequency of the addressed frame has been passed by two peaks or if no further peak is found. It is deliberate that the score for the peak propagation considers only the absolute peak heights. A measure that accounted also for the peak distances would deliver false peak sequences, owing to the irregular peak distances that we expect in unstable regions. Fig. 3 shows the look-ahead strategy for successor peaks in a left direction, starting at peak p_0 that is part of peak tuple $\langle p_0, p_1, p_R \rangle$. Peaks $p_{k(1)}$ and $p_{k(2)}$ are inspected in interval I_0 from p_0 , peaks $p_{k(2,1)}$ and $p_{k(2,2)}$ in interval $I_{k(2)}$ from peak $p_{k(2)}$ gained in the first round. The peak pair $p_{k(2)}$ and $p_{k(2,2)}$ achieves the highest score, i.e. highest mean absolute value, thus $p_{k(2)}$ is installed as the next valid peak.

The final step in the peak propagation stage is the evaluation of peak sequence variants. In general, the variant with the highest score, i.e. with the

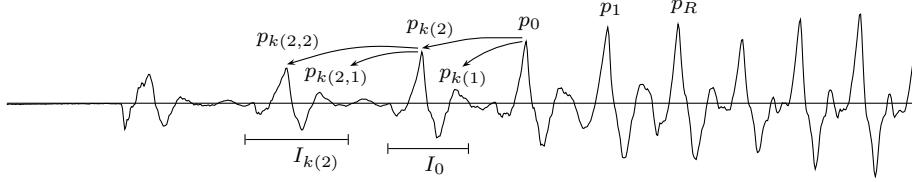


Fig. 3. Peak propagation for V2 (extended period case) with tuple $P = \langle p_0, p_1, p_R \rangle$ in propagation direction to the left. Peaks $p_{k(1)}$ and $p_{k(2)}$ are inspected from p_0 in interval I_0 , peaks $p_{k(2,1)}$ and $p_{k(2,2)}$ are found in interval $I_{k(2)}$ when starting from $p_{k(2)}$

highest mean absolute peak height, is the best peak continuation sequence. However, some checks still need to be performed to verify it. Here we describe the evaluation process for the case where the previous frame has no octave jump (regular case): a similar procedure is applied if the previous frame is in an octave jump down state. In the regular case, we first check whether V1 and V2 deliver the same peak sequence. If so, we keep V1 and discard V2. Otherwise, an additional peak propagation step for the next frame is performed to see whether V2 diverges and delivers periods which are too large. In this case, V2 is discarded and V1 is kept. In all other cases, we keep the variant with the larger score, i.e. the higher absolute mean peak height, in V1. Then, if V3 has a score greater than or equal to V1, we evaluate V1 against V3. V3 is installed and the frame's octave jump state is set to 'down' only if V3 has no middle peaks of sufficient heights, i.e. if the absolute height of the middle peak is smaller than a given percentage of the minimum of the absolute heights of the enclosing peaks. Otherwise, V1 is established.

5 Unstable Voiced Segments

Voiced segments without stable regions, or voiced segments that have no sufficiently large subsequences of equal F_0 , are treated in a separate post-processing step. Basically, the same propagation procedure is applied, but the propagation starting point or anchor is found using looser conditions and additional information.

First, we compute the mean F_0 of the last second of speech. The mean F_0 is calculated by considering only those frames with a reliable F_0 estimate, i.e. frames of voiced segments where the verification step, i.e. the F_0 propagation, has been performed. We then compute candidate F_0 values for all unstable frames of the voiced regions in the range of the mean F_0 . The permissible F_0 range is the same as described in Section 4.1. The anchor point for propagation is found by inspecting the equal section list of the stable segments in the voiced regions, or a small section around the highest-energy frame if no stable segment in the voiced region exists. The propagation starts from such a section if the mean of the F_0 estimates does not deviate too largely from the last second's mean F_0 . If

no such section can be found, we leave the F_0 estimates unchanged. In this case, no propagation takes place.

6 Experiments and Results

Our F_0 estimation algorithm was evaluated on the Keele pitch reference database for clean speech [14] as a proof of concept. We measured the voiced error rate (VE), the unvoiced error rate (UE) and the gross pitch error rate (GPE). A voiced error is present if a voiced frame is recognized as unvoiced, an unvoiced error exists if an unvoiced frame is identified as voiced and a gross pitch error is counted if the estimated F_0 differs by more than 20 % from the reference pitch. The precision is given by the root mean square error (RMSE) in Hz for all frames classified as correct, i.e. as neither voiced, unvoiced, nor gross pitch errors. Results for the proposed algorithm - denoted as HCog (human cognition based) - are given in Table 1. We also cite results of other state-of-the-art F_0 estimation methods: RAPT [7] (one of the best time-domain algorithms based on cross correlation functions and dynamic programming), and the two frequency-domain algorithms PSHF Based [8] and Nonnegative Matrix Factorization (NMF) [9].

Table 1. Results obtained on the Keele pitch reference database

	VE (%)	UE (%)	GPE (%)	RMSE (Hz)
HCog	2.53	4.46	1.49	5.09
RAPT	3.2	6.8	2.2	4.4
PSHF	4.51	5.06	0.61	2.46
NMF	7.7	4.6	0.9	4.3

The results show that the proposed algorithm performs excellently in terms of VE and UE. None other of the cited algorithms shows such low VE and UE. The GPE, at 1.49 %, is clearly lower as for RAPT but not as low as for the frequency-domain algorithms. However, a GPE of 1.49 % in the presence of a VE of only 2.53 % is very low. A higher VE may also hide several gross pitch errors.

The RMSE, at 5.09 %, is higher than with the other algorithms. We see two reasons for this. First, maximum and minimum peaks often have an inclination - either to the left or to the right - and often, there is a set of close peaks around the maximum or minimum peak so that F_0 is not as accurately calculated as with other methods. Second, it may occur that the leftmost or rightmost peak of a peak tuple is not the true period end point due to thresholds selected and the fact that propagation is started from suboptimal peaks. However, accuracy can certainly be improved by adjustment procedures and smoothing.

7 Conclusions

We have presented an F_0 estimation algorithm as an approximate model of the human cognitive process. The algorithm achieves very low error rates, outperforming the state-of-the-art correlation-based reference method in this respect. These results are achieved with little resources in terms of memory and computing power. Obviously, the strength and potential of the algorithm lie in the concepts which simulate human recognition of F_0 .

The algorithm is thoroughly extensible, as new special cases are easily implemented. In this sense, the algorithm can also be applied to other tasks, e.g. spontaneous speech, by analyzing the new cases and modeling them. In this way it will become more and more generic. This procedure closely reflects human learning, which is said to function by adopting examples and building patterns independently of the frequency or probability of their occurrence [15]. For this reason, we have refrained from using weights or probabilities to favor one or another case but look ahead and evaluate until the case is decided.

Our algorithm delivers a classification of the speech signal into stable and unstable segments additionally to the F_0 contour. Automatic speech recognition systems may profit from this classification since the recognition of phonemes should preferably be started from stable segments as well. The spectral information required for phoneme recognition is certainly more reliably computed on those segments.

Future work will focus on extending the algorithm to other tasks and improving the accuracy of the F_0 estimates.

8 Acknowledgements

The authors wish to thank Prof. Jozsef Szakos from Hong Kong Polytechnic University for valuable comments on this paper. We thank Matthias Scheller Lichtenauer and Iris Sprow from Swiss Federal Laboratories for Materials Science and Technology, as well as Prof. Guy Aston from University of Bologna for their careful proof-reading.

References

1. Traunmüller, H.: Paralinguistic Variation and Invariance in the Characteristic Frequencies of Vowels. *Phonetica* **45**(1) (1988) 1–29
2. Ewender, T., Pfister, B.: Accurate Pitch Marking for Prosodic Modification of Speech Segments. In: Proc. of Interspeech. (2010) 178–181
3. Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A.: A Comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**(5) (October 1976) 399–418
4. Noll, A.M.: Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection. *J. Acoust. Soc. Am.* **36**(2) (February 1964) 296–302
5. Markel, J.D.: The SIFT Algorithm for Fundamental Frequency Estimation. *IEEE Transactions on Audio and Electroacoustics* **20**(5) (December 1972) 367–377

6. Secrest, B.G., Doddington, G.R.: An integrated pitch tracking algorithm for speech systems. In: Proc. of ICASSP. (1983) 1352–1355
7. Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W.B., Paliwal, K.K., eds.: Speech Coding and Synthesis, Elsevier Science B. V. (1995)
8. Roa, S., Bennewitz, M., Behnke, S.: Fundamental frequency estimation based on pitch-scaled harmonic filtering. In: Proc. of ICASSP. (2007) 397–400
9. Sha, F., Saul, L.K.: Real-Time Pitch Determination of One or More Voices by Nonnegative Matrix Factorization. In: Advances in Neural Information Processing systems 17, MIT Press (2005) 1233–1240
10. Peharz, R., Wohlmayr, M., Pernkopf, F.: Gain-robust multi-pitch tracking using sparse nonnegative matrix factorization. In: Proc. of ICASSP. (2011) 5416–5419
11. Achan, K., Roweis, S., Hertzmann, A., Frey, B.: A Segment-Based Probabilistic Generative Model Of Speech. In: Proc. of ICASSP. (2005) 221–224
12. Moore, B.C.J.: An Introduction to the Psychology of Hearing. Emerald, Bingley, United Kingdom (2008)
13. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2011)
14. Plante, F., Meyer, G.F., Ainsworth, W.A.: A pitch extraction reference database. In: Proc. Eurospeech '95. (1995) 837–840
15. Kuhn, T.S.: Second Thoughts on Paradigms. In: The Essential Tension. Selected Studies in Scientific Tradition and Change, The University of Chicago Press, Chicago and London (1977) 293–319

On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness?

Eva Lasarcyk¹, Charlotte Wollermann², Bernhard Schröder²
and Ulrich Schade³

¹Institute of Phonetics, Saarland University, Germany

evaly@coli.uni-saarland.de

²German Linguistics, University of Duisburg-Essen, Germany

{charlotte.wollermann, bernhard.schroeder}@uni-due.de

³Fraunhofer Institute for Communication, Information Processing
and Ergonomics FKIE, Germany

ulrich.schade@fkie.fraunhofer.de

Abstract

In this paper we present work on the modelling of uncertainty by means of prosodic cues in an articulatory speech synthesizer. Our stimuli are embedded into short dialogues in question-answering situations in a human-machine scenario. The answers of the robot vary with respect to the intended level of (un)certainty, the independent variables are *intonation* (rising vs. falling) and *filler* (absent vs. present). We perform a perception study in order to test the relative impact of the prosodic cues of uncertainty on the perception of uncertainty and also of naturalness. Our data indicate that the cues of uncertainty are additive. If both prosodic cues of uncertainty are present, the perceived level of uncertainty is higher as opposed to the deactivation of a single cue. Regarding the relative contribution of *intonation* vs. *filler* our results do not show a significant difference between judgments. Moreover, the correlation between the judgment of uncertainty and of naturalness is not significant.

1 Introduction

The general topic of this paper is the role of uncertainty in question-answering situations. Suppose a communicative situation with two conversational partners. A asks B a question, and B is not certain with respect to her answer. Why is B uncertain in this situation? There might be several reasons: i) B only partially knows the answer,

ii) B cannot judge what the listener already knows, iii) B does not know how to formulate the message etc. For a detailed presentation of the process of language production in general and its possible troubles cf. Levelt (1989).

In addition, uncertainty can be regarded as a complex phenomenon. In some works uncertainty is categorized as emotion (Rozin, Cohen, 2003; Keltner, Shiota, 2003), in other works it is assumed to have a cognitive character (Kuhltau, 1993). In the context of question-answering situations, the following questions arise: Which prosodic cues do speakers use for encoding uncertainty in answers? Which prosodic cues contribute to the perception of uncertainty?

2 Communication of uncertainty

In this section we firstly discuss the role of uncertainty in human-human communication (Section 2.1). Afterwards we refer to previous studies on the role of uncertainty in human-machine communication (Section 2.2). In Section 2.3, we give a general motivation for investigating uncertainty as an expressive ability of machines.

2.1 Uncertainty in human-human communication

In human-human communication, conversation partners use several prosodic cues in order to signal and also to perceive uncertainty in answers. With respect to speech production in the study of Smith and Clark (1993), metacognition judgments in question-answering situation were elicited by using the *Feeling of Knowing* (FOK) paradigm. Results suggest that speakers mark uncertainty by using *rising intonation*, *pauses*, *fillers* and *lexical hedges*. For investigating the hearer's side as well, in Brennan and Williams (1995) the *Feeling of Another's Knowing* (FOAK) was defined. Results from their perception study show for the acoustic channel that the *intonation*, the *form* of answers, *pauses* and also *fillers* effect the FOAK.

Furthermore, *fillers* and *pauses* have been found as relevant cues with respect to self-repair in speech, especially to those self-repairs that do not contain lexical material (coined *c-repairs*) (Goldman-Eisler, 1967; Levelt, 1983). These repairs occur if the speaker recognises and corrects the slip of the tongue even before a speech signal is produced. A connectionist model of such a kind of repairs can be found in Schade and Eikmeyer (1991).

Swerts and Krahmer (2005) replicated the study of Smith and Clark (1993) and extended the design to the visual aspect. For the audio channel, *delay*, *pauses* and *fillers* were found as being relevant for marking uncertainty; for the visual modality, *smiles* and *funny faces*. In order to test the relevance of these cues for speech perception, audio-only, visual-only, and audiovisual stimuli were presented to subjects and had to be judged with respect to uncertainty. Results suggest that subjects were able to distinguish certain from uncertain utterances in all three conditions, but identification was easier in the bimodal condition than in the unimodal conditions.

Also with respect to audiovisual cues of uncertainty, Borràs-Comes et al. (2011) tested the relative contribution of *facial gestures*, *intonation* and *lexical choice* on uncertainty perception. Results suggest that all three cues have a significant effect on

perceived uncertainty. Furthermore, in the case of a mismatch between *gesture* and *intonation*, *gesture* has a stronger impact.

2.2 Uncertainty in human-machine communication

In the context of human-machine communication however, it is less clear if these cues contribute to the perception of uncertainty in a comparable way. Marsi and van Rooden (2007) argue that the modelling of uncertainty can improve information systems by enriching expressive abilities. With respect to acoustic speech synthesis, Adell et al. (2010) modelled *filled pauses* on the basis of a ‘synthetic disfluent speech model’. For these purposes an unit-selection synthesizer was used. In a next step a perception study was performed in order to test whether *filled pauses* can be generated without decreasing the system’s quality. The results show no significant decrease of the system’s naturalness. In the study of Andersson et al. (2010) utterances were selected from spontaneous conversational speech. The goal was to generate *fillers* without affecting the system’s naturalness in a negative way. By using a machine-learning algorithm, type and placement of *fillers* and of *filled pauses* were predicted. Again, a unit-selection voice was used. Similar to the findings of Adell et al. (2010), no significant decrease of naturalness was observed during the evaluation.

In addition, the role of uncertainty in human-machine communication has also been investigated with respect to visual speech synthesis. The results of Oh (2006) suggest that the variation of *facial expressions* and *head movements* affects the recognition of uncertainty. According to Marsi and van Rooden (2007) *head movement* alone, and also combined with *eyebrow movement*, affects the perception of uncertainty as well.

The automatic detection of uncertainty in utterances by dialogue systems is for instance useful for systems that function as tutors. The study of Pon-Barry et al. (2006) suggests that the learning process of the student can be affected positively if the system adapts to the student’s uncertainty. For training these systems, corpora consisting of natural conversations between tutors and students are often used. Uncertain utterances have been detected with an accuracy of ca. 75% by the usage of prosodic cues covering *fundamental frequency*, *intensity*, *tempo* and *duration* (Liscombe et al., 2005; Pon-Barry, Shieber, 2009).

2.3 Motivation

As already mentioned in the previous section, the modelling of uncertainty can be useful to create systems with expressive abilities (Marsi, van Rooden, 2007). Why is it useful to have systems equipped with those abilities? Natural language is characterized by a high degree of variability (Murray, Arnott, 1996). Speech does not only differ from speaker to speaker, but also within an individual speaker. This variability is caused by different factors, e.g. by speaking style and by emotion and mood (cf. Murray, Arnott, 1996). If one aims to develop speech synthesis systems with an as natural as possible speech output, this variability needs to be taken into account. We regard the expression of uncertainty as one factor which can contribute to the variability of synthetic speech.

Moreover, we are interested in simulating uncertainty as a human meta-cognitive state by an artificial system which is able to express this uncertainty in the synthetic

signal. Also, we would like to investigate whether human listeners ascribe this meta-cognitive state to the machine.

In our work we model different degrees of uncertainty by means of prosodic cues, using an articulatory speech synthesizer to generate the utterances. A motivation is given in the following section. We perform a perception study to test to what extent the intended uncertainty indeed affects speech perception.

3 Articulatory speech synthesis

To generate the highly variable speech, we use the articulatory synthesis system Vocal-TractLab (Birkholz, 2006). The system produces utterances of high acoustic quality. It processes a timeline of articulatory gestures which are translated into trajectories of speech articulators in a virtual three-dimensional vocal tract (Birkholz et al., 2011). In an aerodynamic-acoustic simulation step, the speech signals are generated. Since each utterance is created ‘from scratch’, the system is very versatile and offers large degrees of freedom for variation. The prosodic demands on the manner of speaking can be integrated at the foundation of the utterance planning, and no post-hoc signal processing needs to be applied.

4 Related work

An initial investigation on the modelling and perception of uncertainty using the articulatory speech synthesizer by Birkholz (2006) was presented in Wollermann and Lasarczyk (2007). Four different degrees of intended uncertainty were generated by varying the cues *intonation* (rising vs. falling), *delay* (present vs. absent) and the *filler* ‘hmm’ (present vs. absent). The scenario was a fictitious telephone dialogue between a weather expert system and a user. The answer of the system was marked by different degrees of uncertainty. Results show that the activation of all uncertainty cues has a stronger impact on the perceived uncertainty than *rising intonation* alone and *delay* combined with *rising intonation*. In a follow-up study (Lasarczyk, Wollermann, 2010), all eight possible combinations of the three cues were used for conveying different degrees of uncertainty. Moreover, the stimuli were presented in a modified scenario, an interaction between a robot for image recognition and a user. The user showed pictures of fruits and vegetables to the robot and asked the robot, ‘Was siehst Du?’/*What do you see?* The robot recognized the objects. Depending on a fictitious recognition confidence score, the system conveyed (un)certainty in its answer by using the cues mentioned above. Results provide evidence for additivity of all three uncertainty cues with respect to uncertainty perception. Compared to the effects of *rising intonation* and *filler*, the influence of *delay* was relatively weak.

From our findings we infer the following questions which need to be further investigated: i) Does a much longer duration of the cue *delay* contribute more strongly to the perception of uncertainty? ii) To what extent does the filler ‘uh’ affect the perception of uncertainty? iii) Does the expression of uncertainty influence the naturalness of the synthetic utterances? We address these questions in the current paper. To do this, we modify the speech material used in Lasarczyk and Wollermann (2010).

5 Material

Our stimuli consist of four different one-word phrases in German ('Melonen'/melons, 'Bananen'/bananas, 'Tomaten' tomatoes, 'Kartoffeln'/potatoes). Each one is generated in eight different levels of uncertainty by varying *intonation* (rising vs. falling), *delay* (absent vs. present) and the *filler* 'uh' (absent vs. present).

The variation of *intonation* takes place in the last syllable of each word: For *rising intonation* fundamental frequency increases to around 200 Hz, for *falling intonation* it decreases to around 70 Hz. The *delay* refers to the time between the user's question ('Was siehst Du?'/What do you see?) and the system's response ('Bananen', 'Tomaten', ...). In each case there is a default *delay* of 1000 ms. In the case of a long *delay* there are two subcases: i) when *filler* is absent the additional *delay* is 4000 ms, ii) when *filler* is present we apply the default *delay* (1000 ms) + *filler* 'uh' (duration of 370 ms) + *delay* (3630 ms). For the *filler* we choose the particle 'uh' this time, since 'uh' is the *filler* which occurs most often in the Verbmobil corpus for German (Batliner et al., 1995).

To distract the subjects from our interest we use four distractor items ('Bohnen'/beans, 'Paprika'/sweet pepper, 'Gurken'/cucumber, 'Knoblauch'/garlic). To generate the distractor items, we use *falling intonation*, default *delay*, and no *filler*. By using the distractor items it should be precluded that the subjects' linguistic awareness is focused on the tested question.

6 Experimental design

Our overall experimental design consists of three experimental blocks. In each experimental block we vary two of our three prosodic factors. In the first block we test the relative contribution of *filler* vs. *delay* on the perception of uncertainty (cf. Table 1, left side). In the second block we investigate the influence of *intonation* vs. *delay* (cf. Table 1, middle). In the last block, the relative impact of *intonation* vs. *filler* is tested (cf. Table 1, right side). Furthermore, in all three cases we calculate whether there is a correlation between the perception of uncertainty and the perception of naturalness.

The results of block I and II are described in detail in Wollermann et al. (2013) and will only briefly be summarized here. In block I, the stimuli were presented to 74 subjects. They rated the degree of uncertainty and naturalness of each stimulus on 5-point Likert scales. Results suggest an effect of additivity of the uncertainty cues. If both *filler* and *delay* are present, the level of perceived uncertainty is higher as opposed to when one of the cues is deactivated. Furthermore, there was no significant difference between the effects of *filler* and *delay* – as single cues – on the perceived level of uncertainty. Moreover, our data do not suggest evidence for a correlation of uncertainty ratings and naturalness ratings in a significant way.

In block II, the stimuli were evaluated by 79 participants. Similar to block I, a principle of additivity can be observed since *rising intonation* combined with *delay* has a stronger impact on the perceived uncertainty than *rising intonation* alone or *delay* alone. When comparing the effects of the single cues against each other, our data indicate that *rising intonation* yields a stronger level of perceived uncertainty than *delay*. Again, no significant correlation between the perception of uncertainty and naturalness is found.

Table 1: Cues of uncertainty. Left: Block I. Middle: Block II. Right: Block III.

Level	Filler	Delay	Level	Intonation	Delay	Level	Intonation	Filler
C	–	–	C	–	–	C	–	–
U3	–	+	U3	–	+	U4	–	+
U4	+	–	U8	+	–	U8	+	–
U7	+	+	U11	+	+	U12	+	+

Table 2: Ordering of the stimuli (highlighted in yellow), as presented in the perception test groups 1 to 4. Positions 2, 3, 6, and 7 are filled with distractor items.

Position	Group 1	Group 2	Group 3	Group 4
1	C-Kartoffeln	U4-Bananen	U8-Tomaten	U12-Melonen
2	C-Bohnen	C-Knobauch	C-Paprika	C-Gurken
3	C-Gurken	C-Paprika	C-Bohnen	C-Knobauch
4	U8-Bananen	U12-Kartoffeln	C-Melonen	U4-Tomaten
5	U12-Tomaten	U8-Melonen	U4-Kartoffeln	C-Bananen
6	C-Knobauch	C-Bohnen	C-Gurken	C-Paprika
7	C-Paprika	C-Gurken	C-Knobauch	C-Bohnen
8	U4-Melonen	C-Tomaten	U12-Bananen	U8-Kartoffeln

7 Perception study

In the following section we present the experimental design, the procedure and the results of block III. The goal of this study is to test the impact of *intonation* and/or *filler* on the perception of uncertainty and naturalness.

7.1 Material and hypothesis

We used the four different levels of intended (un)certainty shown in Table 1, right side.¹ To illustrate the structure of the stimuli, the simulated interactions between the human and the machine concerning *bananas* are listed below. A question mark at the end of a phrase indicates rising intonation.

C: Human: ‘Was siehst Du?’ (What do you see?) – Machine: [delay 1 s] ‘Bananen.’ (Bananas.)

U4: Human: ‘Was siehst Du?’ – Machine: [delay 1 s] [‘uh’ 370 ms] ‘Bananen.’

U8: Human: ‘Was siehst Du?’ – Machine: [delay 1 s] ‘Bananen?’

U12: Human: ‘Was siehst Du?’ – Machine: [delay 1 s] [‘uh’ 370 ms] ‘Bananen?’

The stimuli were divided into four sets, as shown in Table 2. In each group we presented eight stimuli: four items and the four distractor items. Each stimulus occurred exactly once with respect to the overall data.

We assume that prosodic indicators of uncertainty have an additive effect with respect to uncertainty perception, i.e. the more uncertainty cues are activated, the higher the level of perceived uncertainty. Our detailed assumption is as follows: C will receive,

¹We plan to test more than these four levels of uncertainty. To make the current stimuli comparable to future experiments, the coding of the levels is not done using a straight count.

relatively to the other levels, the highest rating of perceived certainty. U4, U8, and U12 are intended levels of uncertainty. We expect that U12 will lead to the highest rating of perceived *uncertainty*. We further assume that U4 and U8 will be rated between C and U12.

Our goal is to model different levels of intended uncertainty which are closely connected to a relatively high level of naturalness. We refer to naturalness as *relatively* high because we assume that the human listeners will identify the artificial nature of the synthesized speech (as opposed to the human speech) and thus will not ascribe absolute naturalness to the system's utterances. Our expectation is as follows: If we are able to model uncertainty by prosodic cues without decreasing the naturalness of the system, the prosodic cues are adequate to trigger different degrees of intended uncertainty. Therefore we expect no significant correlation between perceived naturalness and perceived uncertainty.

7.2 Procedure

108 undergraduate students (82 f, 26 m) from the University of Duisburg-Essen took part in the perception study. All of them were native speakers of German. The subjects were tested in four groups (g1: N=25, g2: N=17, g3: N=31, g4: N=35). In each group a subset of the stimuli was presented and also a different order of the items was used to neutralise the impact of learning effects.

The dialogues consisted of the question-answer pairs described in the previous section and were played back over loudspeakers. The procedure started with an example stimulus. For each dialogue, subjects were instructed to judge the answer of the system on a questionnaire, using two 5-point Likert scales to indicate how (un)certain the answer sounded and also how natural it sounded (5=certain, 1=uncertain; 5=natural, 1=unnatural).

For statistical analysis, we firstly test the overall difference between judgments with respect to uncertainty and naturalness, respectively, using the Kruskal-Wallis Rank Sum Test. Secondly, we perform the Wilcoxon Signed Rank test with Bonferroni correction to calculate single comparisons between the different levels. Finally, we use Spearman's Rho Test to test if there is a correlation between the uncertainty ratings and the naturalness ratings.²

8 Results

In the following we present the results of the perception of uncertainty (Section 8.1) and of the perception of naturalness (Section 8.2).

8.1 Uncertainty

The Kruskal-Wallis Rank Sum Test indicates that the overall difference between uncertainty judgments is highly significant ($p<0.0001$, level of significance: 5%). Figure 1

²Results of perceived uncertainty alone were presented at the Workshop of the Scandinavian Association for Language and Cognition in June 2013 in Joensuu, Finland (without publication).

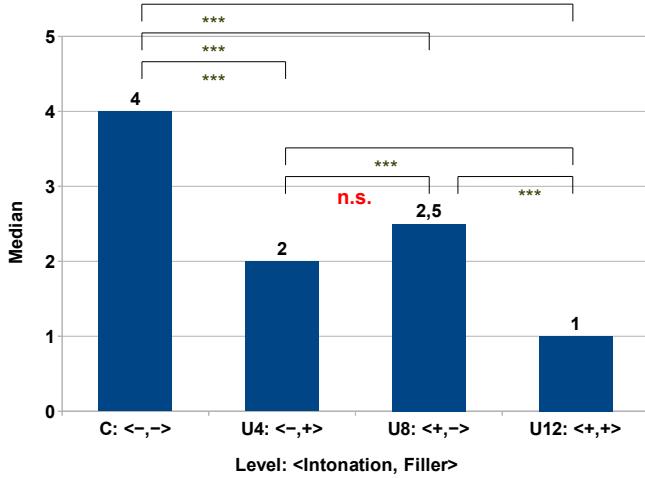


Figure 1: Clustered data – uncertainty judgments; $p<0.008$:*, $p<0.001$:**, $p<0.0001$:***

shows the results for the clustered data, i.e. aggregated for all four stimuli of each level of uncertainty. The Wilcoxon Signed Rank Test with Bonferroni correction (level of significance: $1/6 \times 5\%$) results in $p<0.0001$ for all comparisons, except for the comparison U4 vs. U8. In the latter case there is no significant difference between judgments ($p>0.008$).

In a next step, we analyse the judgments for the individual stimuli. The results are illustrated in Figure 2. For all four wordings, the following comparisons show a significant difference between judgments: C vs. U4, C vs. U8, C vs. U12, and U8 vs. U12. The levels U4 (*filler* activated individually) vs. U8 (*intonation* activated individually) are never rated significantly differently. U4 vs. U12 only shows a significant difference for *Bananen* und *Tomaten*, but not for *Kartoffeln* and *Melonen*.

8.2 Naturalness

For naturalness, the Kruskal-Wallis Rank Sum Test does not show a significant difference between judgments when we look at the data overall ($p>0.05$). It can be observed that each of the four different levels of (un)certainty is judged with a median of 4 (cf. Figure 3). The Wilcoxon Signed Rank Test with Bonferroni correction indicates for each of the six inter-level comparisons that judgments do not differ significantly from each other ($p>0.008$ in all cases). Regarding a possible correlation of the ratings of uncertainty and naturalness, the Spearman's Rho Test results in a correlation coefficient of -0.11 ($p>0.05$). Thus, as expected, our data do not suggest evidence for a correlation in a significant way.

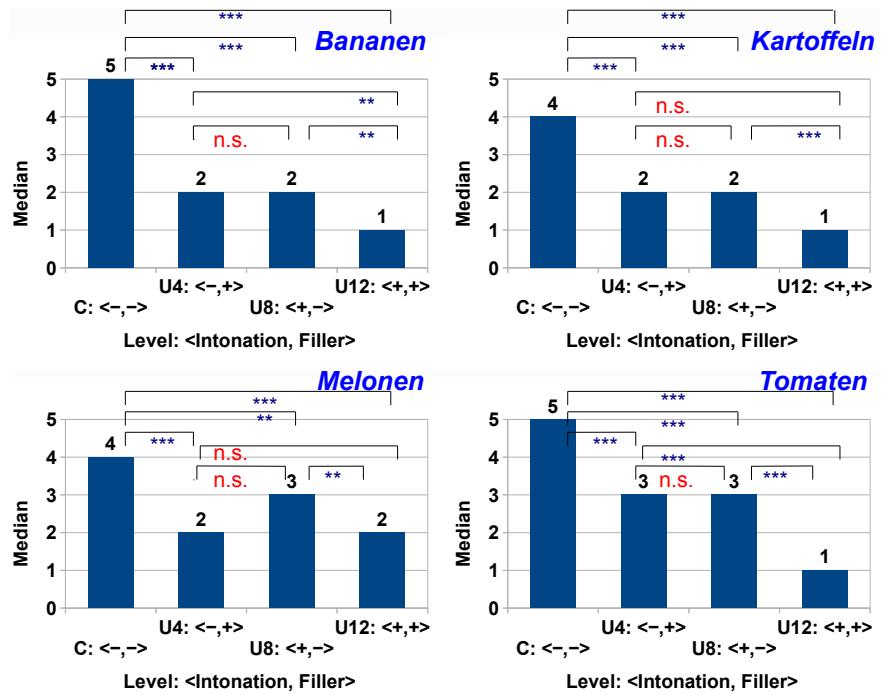


Figure 2: Individual stimuli – uncertainty judgments; $p<0.008$:*, $p<0.001$:**, $p<0.0001$:***

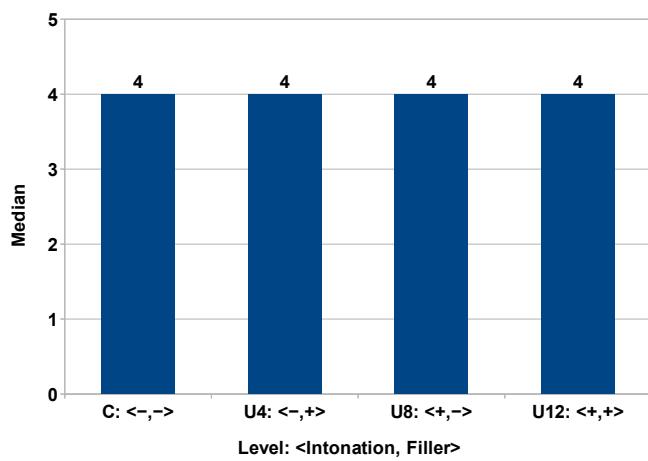


Figure 3: Clustered data – naturalness judgments

9 Discussion

In this paper we presented a study on the modelling of uncertainty by prosodic cues in articulatory speech synthesis. We varied *intonation* and *filler* and tested the relative impact of these cues on perceived uncertainty and perceived naturalness. Regarding uncertainty, the results of the experiment suggest that the cues are additive, i.e. the more uncertainty cues are activated the higher the perceived level of uncertainty. However, our data do not suggest evidence for a stronger effect of *filler* or *intonation* since the subjects' judgments do not differ significantly when these cues are activated individually (U4 vs. U8).

With respect to the perception of naturalness, we do not observe a significant effect of the prosodic cues. In a similar way, the correlation between perceived uncertainty and perceived naturalness is not significant. This result is in line with our assumptions because it indicates that *filler* and *delay* increase the perceived level of uncertainty but do not reduce naturalness. If that were the case, it would be problematic since it could indicate that listeners perceived high levels of uncertainty due to low naturalness, and not due to prosodic variation.

We conclude that different degrees of uncertainty can be expressed by the variation of prosodic cues. As modelled here, varying prosody neither increases nor decreases the naturalness of the utterances. Thus, we assume that – for our scenario – listeners decode uncertainty in the answers of the system and ascribe a meta-cognitive state to the machine.

For future work, we regard it as important to evaluate for different scenarios whether the modelling of uncertainty is a benefit for human-machine communication. Also, we would like to take into account the visual aspect of speech. In several studies, visual prosodic cues have been synthesized (e.g. Krahmer et al., 2002; Granström, House, 2007), and uncertainty in particular has been modelled by means of audiovisual prosody (Oh 2006; Marsi, van Rooden, 2007). We would like to further investigate the interplay between audio and visual prosody and its relevance for perceived uncertainty.

10 Acknowledgments

Many thanks to Denis Arnold and Bernhard Fissen for helpful comments.

References

- [1] Adell, J., Bonafonte, A., Escudero-Mancebo, D. (2010) Modelling Filled Pauses Prosody to Synthesise Disfluent Speech. In: *Proceedings of Speech Prosody 2010*, Chicago, IL, pp. 100624:1-4.
- [2] Andersson, S., Georgila, K., Traum, D., Aylett, M., Clark, R. A. J. (2010) Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech. In: *Proceedings of Speech Prosody 2010*, Chicago, IL, pp. 100116:1-4.

- [3] Batliner, A., Kieling, A., Burger, S., Nöth, E. (1995) Filled Pauses in Spontaneous Speech. In: *Proceedings of 13th International Congress of Phonetic Sciences*, 3, Stockholm, Sweden, pp. 472-475.
- [4] Birkholz, P. (2006) *3D-Artikulatorische Sprachsynthese*, Berlin: Logos.
- [5] Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C.J. (2011) Model-Based Reproduction of Articulatory Trajectories for Consonant-Vowel-Sequence. In: *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5), pp. 1422-1433.
- [6] Borràs-Comes, J., C., Roseano, P., del Mar Vanrell, M., Chen, A., Pietro, P. (2011) Perceiving uncertainty: facial gestures, intonation, and lexical choice. In: *Proceedings of the Workshop on Gesture and Speech in Interaction*. Bielefeld, Germany.
- [7] Brennan, S.E., Williams, M. (1995) The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. In: *Journal of Memory and Language*, 34, pp. 383-398.
- [8] Goldman-Eisler, F. (1967) Sequential temporal patterns and cognitive processes in speech. In: *Language and speech* 10(2), pp. 122-132.
- [9] Granström, B., House, D. (2007) Inside out – acoustic and visual aspects of verbal and non-verbal communication. In: *Proceedings of the 16th International Congress of Phonetic Sciences 2007*, Saarbrücken, Germany, pp. 11-18.
- [10] Keltner, D., Shiota, M.N. (2003) New Displays and New Emotions: A Commentary on Rozin and Cohen (2003). In: *Emotion*, 3(1), pp. 86-91.
- [11] Krahmer, E., Ruttkay, Z., Swerts, M., Wesselink, W. (2002) Pitch, Eyebrows and the Perception of Focus. In: *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France, pp. 443-446.
- [12] Kuhlthau, C.C. (1993) *Seeking Meaning: A Process Approach to Library and Information Services*, Norwood, NJ: Ablex.
- [13] Lasarczyk, E., Wollermann, C. (2010) Do prosodic cues influence uncertainty perception in articulatory speech synthesis? In: *Proceedings of the 7th ISCA Workshop on Speech Synthesis*. Kyoto, Japan, pp. 230-235.
- [14] Levelt, W.J.M. (1983) Monitoring and self-repair in speech. In: *Cognition* 14, pp. 41-104.
- [15] Levelt, W.J.M. (1989) *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- [16] Liscombe, J., Hirschberg, J., Venditti, J.J. (2005) Detecting certainness in spoken tutorial dialogues. In: *Proceedings of Interspeech 2005*. Lisboa, Portugal, pp. 1837-1840.

- [17] Marsi, E., Rooden, F. van (2007) Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System. In: *Proceedings of the Workshop on Multimodal Output Generation*. Enschede, Netherlands, pp. 105-116.
- [18] Murray, I.R., Arnott, J.L. (1996) Synthesizing Emotions in Speech: Is it Time to Get Excited? In: *Proceedings of the International Conference on Spoken Language Processing 1996*, Philadelphia, PA, pp. 1816-1819.
- [19] Oh, I. (2006) Modeling Believable Human-Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty. PhD thesis, Rutgers University, New Brunswick, NJ, USA.
- [20] Pon-Barry, H., Schultz, K., Bratt, E.O., Clark, B., Peters, S. (2006) Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. In: *International Journal of Artificial Intelligence in Education* 16(2), pp. 171-194.
- [21] Pon-Barry, H., Shieber, S. (2009). The Importance of Subutterance Prosody in Predicting Level of Certainty. In: *Proceedings of the Companion Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL) 2009*. Boulder, CO, pp. 105-108.
- [22] Rozin, P., Cohen, A.B. (2003) High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. In: *Emotion*, 3(1), pp. 68-75.
- [23] Schade, U., Eikmeyer, H.-J. (1991) "wahrscheinlich sind meine Beispiele so sprunghaft und und und eh ehm zu zu telegraph" – Konnektionistische Modellierung von "covert repairs". In: Th. Christaller (ed.): *GWAI-91 1. Fachtagung für Künstliche Intelligenz*. Berlin: Springer Verlag, pp. 264-273.
- [24] Smith, V.L., Clark, H.H. (1993) On the Course of Answering Questions. In: *Journal of Memory and Language*, 32, pp. 25-38.
- [25] Swerts, M., Krahmer, E. (2005) Audiovisual prosody and feeling of knowing. In: *Journal of Memory and Language*, 53, pp. 81-94.
- [26] Wollermann, C., Lasarcyk, E. (2007) Modeling and Perceiving of (Un)Certainty in Articulatory Speech Synthesis. In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis*. Bonn, Germany, pp. 40-45.
- [27] Wollermann, C., Lasarcyk, E., Schade, U., Schröder (2013) Disfluencies and Uncertainty Perception Evidence from a Human-Machine Scenario. In: *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*. Stockholm, Sweden, pp. 73-76.

Supervised learning model for parsing Arabic language

¹Nabil Khoufi, ²Souhir Louati, ¹Chafik Aloulou, ¹Lamia Hadrich Belguith

ANLP Research Group-MIRACL Laboratory, University of Sfax, Tunisia

¹{nabil.khoufi, chafik.aloulou, l.belguith}@fsegs.rnu.tn,

²louati.sou@gmail.com

Abstract. Parsing Arabic language is a difficult task given the specificities of this language and given the scarcity of digital resources (grammars and annotated corpora). In this paper, we suggest a method for Arabic parsing based on supervised machine learning. We used the SVMs algorithm to select the most probable syntactic labels of the sentence. Furthermore, we evaluated our parser following the cross validation method by using the Penn Arabic Treebank. The obtained results are very encouraging.

1 Introduction

Syntactic parsing represents an important step in the automatic processing of any language as it ensures the crucial task of identifying the syntactic structures of the sentences in a particular text. Several studies have been conducted in order to solve the problems of parsing. These efforts can be classified in three distinct approaches: the linguistic approach, the numerical approach, and the mixed or hybrid approach (Aloulou 2005). The linguistic approach uses lexical knowledge and language rules in order to parse sentences whereas numerical approaches are essentially based on statistics or on probabilistic models. This type of approach is mainly based on the frequencies of occurrence that are automatically calculated from the corpora. The third approach is called hybrid approach which is a mixture of the two previous ones: it integrates a linguistic analysis with a numerical one.

This paper is organized into four sections: in section 2, we present the works related to Arabic language parsing. Section 3 describes the different phases of the suggested method. Section 4 presents the principles and results of the evaluation. And section 5 presents our conclusions and suggestions for further research perspectives.

2 Related works

Many works have focused on Arabic syntactic parsing. However, the number of these papers is very limited compared to the number of works dealing with other natural languages such as English or French. To our knowledge, the majority of works around Arabic language parsing use the linguistic approach. The latter gives satisfying results, but these are not yet at the English state-of-the-art level. (Ouersighni et al. 2001) developed a morphosyntactic analyser in modular form for Arabic. The analy-

sis is based on the grammatical AGFL (Affixs Grammars over a Finite Lattice) formalism. The analyser of (Othman et al. 2003) was realised in a modular form too and is based on the rules of the UBG (Unification Based Grammar) formalism. (Zemerli et al. 2004) have established a simple morphosyntactic analyser through the development of an application for vocalic synthesis of the Arabic language based on vowelized Arabic texts. This morphosyntactic analyser consists of two parts: the lexical database and the analysis procedure. In this analysis, the processing order of the text's words is crucial since it allows minimizing labelling errors. Aloulou (Aloulou 2005) has developed a parsing system called MASPAN (Multi-Agent System for Parsing Arabic) based on a multi-agent approach. The chosen grammatical formalism is HPSG (Head-Driven Phrase Structure Grammar). It is a representation that permits to minimize the number of syntactic rules and to provide rich and well-structured lexical representations. The (Bataineh et al., 2009) analyser uses recursive transition networks. (Al-Taani et al. 2012) constructed a grammar under the CFG formalism (Context Free Grammar) and then implemented it in a parser with a top-down analysis strategy. All of these use hand-crafted grammars, which are time-consuming to produce and difficult to scale to unrestricted data. Moreover, these grammars do not fully cover all the specificities of the described language.

There are other Arabic parsers designed according to the numerical approach which are based on statistical calculations or supervised learning techniques. (Tounsi et al., 2009) have developed a parser that learns from the Penn Arabic treebank (PATB) the functional labels in order to assign the respective syntactic structures to the different phrases according to the LFG (Lexical Functional Grammar) formalism. As an example, the analyser of (Ben Fraj 2010) learns from a corpus of syntactic tree patterns how to assign the most appropriate parse tree for syntactic interpretation of a sentence. (Diab et al., 2009) present a machine learning-based method for base phrase chunking.

The study of related works shows that numerical methods for parsing Arabic language remain largely untapped. It is also difficult to compare the results of existing parsers because each one uses a different evaluation metric. But according to the overview of the results of existing parsers, numerical-based parsers give better results than knowledge-based ones and are tested on a larger scale (see Table 1). These good results depend on the use of large amounts of annotated corpora. Since we have access to the PATB corpus and assume that the numerical analysers provide better results also with other languages (Charniak et al., 2005) (Vanrullen et al., 2006), we opted for a numerical method to build our system for parsing Arabic language. More precisely, we use Machine learning techniques based on supervised learning. Table 1 presents a comparison of evaluation results of parsers for the Arabic language.

Table 1. Comparison of the evaluation results

System	Testing data	Results
Al-Taani et al. 2012	70 sentences	Accuracy 94 %
Bataineh et al. 2009	90 sentences	85.6% correct 2,2% wrong 12,2% rejected
Mona Diab 2007	PATB, 10 %	F-score 96.33%.
Ben Fraj 2010	50 sentences	Accuracy 89,85

3 The suggested method

This section is devoted to the presentation of the general architecture of our suggested method.

The suggested method for parsing the Arabic language has two phases: the learning phase and the analysis phase. The first phase requires a training corpus, extraction features and a set of rules extracted from the learning corpus. The second phase implements the learning results from the first phase to achieve parsing. The phases of our approach are illustrated in the following figure:

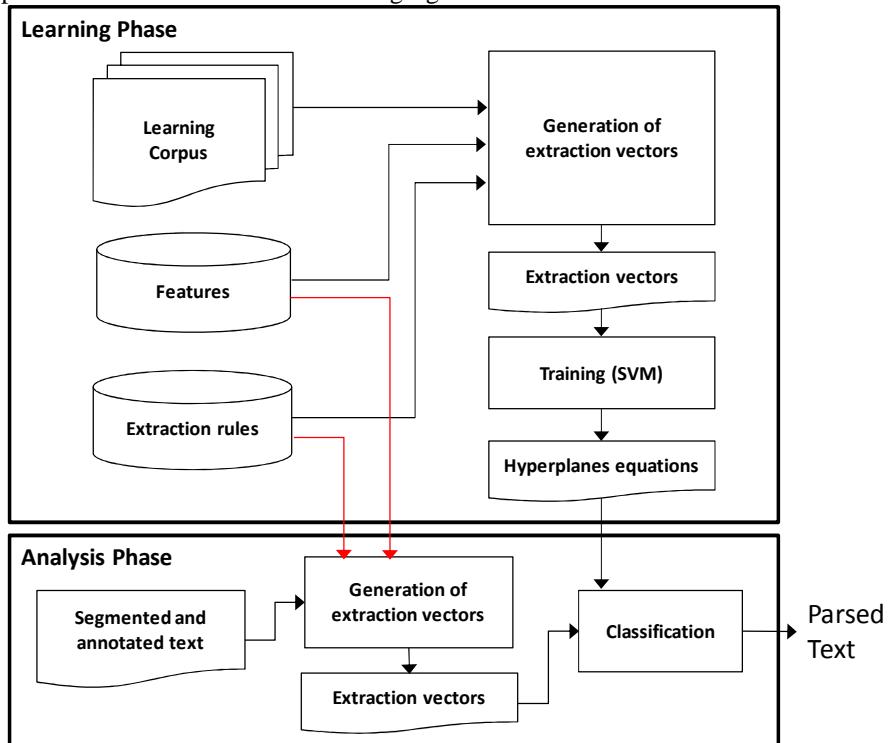


Fig. 1. The suggested method

3.1 The learning phase

The learning phase involves the use of a training corpus, a set of features and rules extracted from the learning corpus analysis in order to train the SVM (Support Vector Machine) classifier.

Learning corpus. The Penn Arabic Treebank (ATB) was developed in the laboratory of Linguistic Data Consortium (LDC) at the University of Pennsylvania (Maamouri M. et al., 2004). It is composed of data from standard and modern linguis-

tic sources written in Arabic. It comprises 599 texts of different stories and news from the Lebanese newspaper An-Nahar. The texts in the corpus do not contain any vowels as it is typically in use in most texts written in Arabic. In the learning phase, we use the version ATB v3.2 of this corpus.

Extraction features. These features indicate the information used from the annotated corpus during the training stage, which is the morphological annotation.

We classified these features into two classes namely, part of speech features and contextual features:

- A part of speech (POS) feature specifies the morphological category of the word being processed.
- A contextual feature indicates the POS of the words in the left vicinity of the word being analyzed with a maximum depth equal to four.

The following table shows the different features used and their explanations:

Table 2. List of utilised extraction features.

	Feature name	Explanation
A Part of speech feature	POS-W	Extract the POS annotation of the word which being processed.
Contextual features	POS-LEFT-i+1	Extract the POS annotation of the word in the left vicinity at position i +1.
	POS-LEFT-i+2	Extract the POS annotation of the word in the left vicinity at position i +2.
	POS-LEFT-i+3	Extract the POS annotation of the word in the left vicinity at position i +3.
	POS-LEFT-i+4	Extract the POS annotation of the word in the left vicinity at position i +4.

Extraction rules. These rules are derived from a deep analysis of the ATB. They are used to train our system in grouping the sequences of labels that may belong to the same syntactic grouping and thus better define their borders. The combination of features and rules extracted from the training corpus allows allocating each word in the sentence to its most probable syntactic group and thus training our analyser to classify them automatically. These rules have the following structure:

$$\text{Rule : } \{M1, M2, M3, M4, M5\} \rightarrow Ci$$

Where M1 through M5 represents the morphological category of the words in a given syntactic group and Ci represents the syntactic class of the group. A rule may be composed of one, two, three, four or five elements. We extracted 53 rules from the ATB. We used the same tag set of the ATBs to simplify the learning process. Here are some examples of extracted rules:

$$R1 : \text{PREP}, \text{NOUN} \rightarrow \text{PP}$$

R2 : ADJ,CONJ,ADJ → ADJP
 R3 : NOUN_PROP → NP
 R4 : PREP,NOUN,POSS_PRON → PP

Generation of the extraction vectors. This step aims to annotate each word of a sentence in the learning corpus according to the different extraction features presented above. Extraction rules are also used to identify the syntactic class of the groups of words.

Each group of words is described by a vector called extraction vector. The nominal value for a given feature corresponds to the morphological annotation of the word (adjective, noun, verb, punctuation, etc.) according to the features used. This vector is completed by the appropriate syntactic class (NP, VP, PP ...) selected from the syntactic annotations in the ATB corpus. The set of extraction vectors forms an input file for the learning stage. At the end of this process, the learning corpus is converted from its original format into a vector format and we obtain a tabular corpus which consists of a set of vectors as shown in the following example (Question mark represent unused features):

Vector1 : PREP,NOUN,?, ?, ?, PP
 Vector2 : ADJ,CONJ,ADJ,?, ?ADJP
 Vector3 : NOUN_PROP,?, ?, ?, ?, NP

Training. This stage uses the previously generated extraction vectors in order to produce equations known as hyperplanes equations. The learning algorithm used in this stage is the SVM algorithm. To our knowledge, there is no work using SVMs for parsing the Modern Arabic Standard. So we decided to use SVMs for learning to test the potential of SVMs in parsing the Arabic language.

Since SVMs are binary classifiers, we have to convert the multi-class problem into a number of binary-class problems. This algorithm generates several hyperplane equations which are used to classify the different word groups according to their appropriate syntactic class (NP, PP, VP, ADJP ...). The training step generates 25 hyperplane equations. It is noteworthy that the learning stage is done only once and is only repeated in case we increase the size of the corpus, or change the type of corpus.

This step is performed using 80 % of the ATB and the Weka library (Frank, E. & Witten, Ian H., 2005)

3.2 The analysis phase

This phase implements the results of the learning phase in order to parse a sentence. The user must provide a segmented and a morphologically annotated text as input to our system. This phase proceeds in two steps as follows:

Firstly, a pre-processing phase is applied to the input sentence. Indeed, we use features and rules to arrange words in groups following the vector format as presented in the learning stage. This pre-processing generates extraction vectors like those generated as input for the learning stage. The only difference is that these vectors do not contain the syntactic class. This information will be calculated by the SVM classifier.

Then, the extraction vectors generated in the first step and the hyperplane equations generated in the learning stage are provided as input to the classification module. Indeed, for each vector, we calculate a score using hyperplane equations. Each equation discriminates between two syntactic classes (e.g. PRT/ADV). So every vector will have 25 scores according to the number of equations. The score and its sign are used to identify the suitable syntactic class for the test vector.

At the end of this stage we obtain a parsed sentence in a tree form.

4 Results

The evaluation of our analyser is achieved following the cross-validation method using the Weka tool. To realise that, we divided the ATB corpus into two distinct parts, one for learning (80%) and one for the evaluation (20%). The results are exposed in the table 3.

Table 3. Evaluation results.

Precision	Recall	F-score
89.01%	80.24%	84.37%

The obtained results are encouraging and represent a good start for the use of supervised learning for parsing the Arabic language.

We noticed that the analysis of short sentences ($<=20$ words) presents the highest measures of recall and precision. As the sentence gets longer, there will be a more complex calculation, which reduces system's performance. This is due to the fact that our system does not handle very complex syntactic structures.

We believe that these results can be improved. In fact, we think that we can improve the learning stage by adding other features besides the POS features. As example of additional features, we can incorporate lexical data (external dictionary) to identify multi-word expressions. During the implementation of our system, we noticed that the larger the number of rules is, the higher the recall and precision are. So we believe that the enrichment of the rules database can significantly improve the results. The addition of syntactic rules is a solution to analyse long sentences.

5 Conclusion and perspectives

In this paper we presented our approach for Arabic parsing based on supervised learning. We used SVM for the learning phase and we obtained an f-score of 84.37%. As a perspective, we plan to integrate an efficient morphological analyser such as MADA in our system in order to process plain text. We also intend to add other features to the learning phase such as *group function*. Lexical data may be integrated to identify multi-word expressions.

References

- Al-Taani, Ahmad T. Msallam Mohammed, M. & Wedian Sana, A. (2012), A top-down chart parser for analyzing arabic sentences. *The International Arab Journal of Information Technology*, Volume 9: pp109-116
- Aloulou, C. (2005) Une approche multi-agent pour l'analyse de l'arabe : Modélisation de la syntaxe, Thèse de doctorat en informatique, Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisie.
- Bataineh Bilal, M. & Bataineh Emad, A. (2009) An Efficient Recursive Transition Network Parser for Arabic Language, *Proceedings of the World Congress on Engineering 2009*, Vol II, WCE 2009, London, U.K, Juillet 1 - 3, pp1307- 1311.
- Ben fraj, F. (2010) Un analyseur syntaxique pour les textes en langue Arabe à base d'un apprentissage à partir des patrons d'arbres syntaxiques. *Thèse de doctorat en informatique*, Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisie.
- Charniak, E. & Johnson, M. (2005) Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, pp 173–180, Ann Arbor, June 2005.
- M. Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking, *MEDAR 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.
- Frank, E. Witten, Ian H. (2005) Practical Machine Learning Tools and Techniques, Second Edition. (Morgan Kaufmann series in data management systems).
- Maamouri, M. Bies, A. Buckwalter, T. and Mekki, W. (2004), The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus, In *NEMLAR International Conference on Arabic Language Resources and Tools*, pp. 102 – 109.
- Othman, E. Shaalan, K. Rafea A. (2003) A Chart Parser for Analyzing Modern Standard Arabic Sentence In *proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, Louisiana, USA.
- Ouersighni, R. (2001) A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts, In *ACL 39th Annual Meeting. Workshop on Arabic Language Processing: Status and Prospect*, Toulouse, pp. 66-72.

Tounsi, L., Attia, M., et Van Genabith, J. (2009) Parsing Arabic Using Treebank-Based LFG Resources, *In the 14th Proceedings of Lexical Functional Grammar*, Cambridge, UK.

Vanrullen, T., Blache, P. et Balfourier, J.-M. (2006). Constraint-based parsing as an efficient solution : Results from the parsing evaluation campaign EASy. *In Proceedings of the 5th Language Resource and Evaluation Conference (LREC'06)*, Genes, Italie. 22, 25

Zemirli Z. et Khabet S. (2004). TAGGAR : Un analyseur syntaxique dédié à la synthèse vocale de textes arabes voyellés, *In Actes des Journées d'Etudes sur la Partie (JEP) et Traitement Automatique des Langues Naturelles (TALN)*, Fès, Maroc.

Disambiguation of the Semantics of German Prepositions: a Case Study

Simon Clematide, Manfred Klenner, and Lenz Furrer

Institute of Computational Linguistics
University of Zurich

Abstract. In this paper, we describe our experiments in preposition disambiguation based on a – compared to a previous study – revised annotation scheme and new features derived from a matrix factorization approach as used in the field of distributional semantics. We report on the annotation and Maximum Entropy modelling of the word senses of two German prepositions, *mit* ('with') and *auf* ('on'). 500 occurrences of each preposition were sampled from a treebank and annotated with syntacto-semantic classes by three annotators. Our coarse-grained classification scheme is geared towards the needs of information extraction, it relies on linguistic tests and it strives to separate semantically regular and transparent meanings from idiosyncratic meanings (i.e. of collocational constructions). We discuss our annotation scheme and the achieved inter-annotator agreement, we present descriptive statistical material e.g. on class distributions, we describe the impact of the various features on syntacto-semantic and semantic classification and focus on the contribution of semantic classes stemming from distributional semantics.

Keywords: Word Sense Disambiguation, Preposition, Distributional Semantics, German

1 Introduction

Prepositions in the sense of single word prepositions are a rather small closed lexical class with several dozen types in languages such as German, English and French. In terms of word occurrences, however, prepositions contribute a substantial amount of tokens. For instance, in the German newspaper treebank TIGER (Brants and Hansen, 2002) 12% of 768,971 word tokens (not counting punctuation tokens) are tagged as prepositions. Prepositions occurring very frequently show a high degree of ambiguity and polysemy. For 13 frequent English prepositions, Litkowski and Hargraves (2006) recorded 211 senses.

Linguistics has a long-standing tradition of sense classification of prepositional phrases used as adjuncts. Traditional dictionaries also collect detailed sense information about prepositions. In case of *mit*, the German online dictionary Duden¹ specifies 8 main senses, additionally some of them have subsenses

¹ See <http://www.duden.de>

resulting in a total of 12 senses. It is yet unclear which classification schemes should be used for applications that require semantic interpretation such as information extraction or questions answering – although there have been two preposition word sense disambiguation (PWSD) shared tasks for English in the past. In this paper, we want to gain experience for a larger attempt in classifying the semantic contributions of prepositions across different languages as German, English and French. Our main interest is to differentiate between semantically transparent contributions that prepositional phrases can provide in a general or productive manner on the one hand and the less transparent contributions in collocational constructions on the other hand. Additionally, many prepositions are subcategorized by verbs and the semantic contribution of the selected prepositions is weak or unspecific – a fact that is often revealed by cross-lingual comparisons of subcategorization frames.

In the Maximum Entropy model we propose, we exploit contextual and syntactic features that have proved most helpful in previous approaches on English PWSD. But we also focus on (German) language-specific features like e.g. morphological case, which turns out to be a strong feature for the preposition *auf* ('on'). Moreover, we have experimented with distributional semantics in order to derive semantic classes for preposition governors and for the noun phrase heads governed by the preposition. To best of our knowledge, this is the first attempt to utilize semantic knowledge derived in a corpus-driven manner in the task of PWSD.

The rest of this paper is organized as follows. Section 2 presents related work. In Section 3, we describe our syntacto-semantic classification system used in the annotation. We also present the approach borrowed from distributional semantics and used in the machine learning experiments for the automatic prediction of the classes. Section 4 contains a systematic evaluation of the different types of evidence that we have integrated in our approach.

2 Related Work

The meaning of a prepositional phrase (PP) depends – among others – on the meaning of its preposition and (the head of) the embedded noun phrase. Determining the functional role such a PP plays within a sentence can be regarded as semantic role labelling (SRL). Preposition word sense disambiguation, thus, is sometimes casted as a variant of SRL (e.g. O'Hara and Wiebe, 2009). For the English language, annotated data is available from the *Penn Treebank II* (Marcus et al., 1994), where thematic roles carried by prepositional phrases are marked, and *FrameNet* (Baker et al., 1998), which was annotated as part of the *Preposition Project* (Litkowski and Hargraves, 2006).

For German, the Salsa 2.0 project (Rehbein et al., 2012) made a substantial amount of FrameNet-like annotations available built on top of the TIGER corpus. About 20,000 verbs and 16,000 nouns are marked as frame-evoking concepts. In Salsa annotations, prepositional phrases appear as frame elements that are linked to the evoking target by named roles. Figure 1 shows the most fre-

59 (Message), 59 (Interlocutor_2), 52 (Partner_2), 48 (Cause), 39 (Phenomenon), 37 (Event), 37 (Response), 31 (Descriptor), 21 (Item2), 20 (Instrument), 18 (Means), 15 (Content), 13 (Goal), 13 (Side_2), 11 (Theme), 11 (Fact), 10 (Degree_of_involvement), 8 (Money), 7 (Goods), 7 (Co_Signatory), 7 (Funds), 7 (Creator), 7 (Contribution_salsa), 6 (Agent), 5 (Result), 5 (Manner), 5 (Party_2), 5 (Defendant), 5 (Outcome), 4 (State_of_affairs), 4 (Quantity), 4 (Medium), 4 (Action), 4 (Party2), 4 (Persistent_characteristic), 4 (Punishment), 4 (Award), 4 (Addressee), 3 (Specification), 3 (Effect), 3 (Body_part), 3 (Mode_of_transportation), 3 (Reason), 3 (Topic), 3 (Relation), 3 (Protagonist), 3 (Accused)

Fig. 1. Frequencies and names of the frame elements of the German FrameNet annotation Salsa 2.0 of PPs headed by *mit* occurring at least 3 times. In total there are 701 occurrences with 111 different frame element roles. 39 roles occur only once, 14 twice.

quent roles associated with PPs headed by *mit*. The fine-grained classification of the English FrameNet (with its larger annotation database) has been a PWS challenge for O’Hara and Wiebe (2009). The even more fine-grained and less generalized role inventory of Salsa 2.0 makes the task of utilizing such a resource demanding.

A substantial contribution on preposition classification and disambiguation for English has been carried out in the Preposition Project (Litkowsky and Hargraves, 2006) (see also the *SemEval* Task on WSD of prepositions, Litkowsky and Hargraves, 2007). A fine-grained classification scheme was derived from the *Oxford Dictionary*, e.g. the preposition *on* is specified on the basis of 25 different senses. Other elaborated classification schemes can be found as part of *VerbNet* (Kipper et al., 2004) and *PrepNet* (Saint-Dizier, 2008). As can be seen from the diversity of these approaches, there is no agreed classification scheme for preposition disambiguation. Moreover, some authors argue that preposition classes are (in part) language-specific (Müller et al., 2011). They have specified an even more fine-grained and hierarchical classification scheme (compared to the Preposition Project), where German gold standard annotations are based on the traversal of manually specified preposition-specific decision trees. As a consequence of the complexity of the annotation scheme, no attempt was made so far by the authors to learn a model for preposition classification based on their semantic classes. Their approach based on logistic regression as described in Kiss et al. (2010) focuses on determiner omission in PPs.

Preposition classification is not only crucial for applications such as information extraction (see Baldwin et al., 2009, p. 134 for an application-oriented discussion), but also supports machine translation, see e.g. Shilon et al. (2012, p. 106). Although semantic information helps to tackle the translation task, the semantic class of a preposition does not perfectly determine the correct translation. As a consequence, these approaches do not strive to carry out preposition WSD, but to use semantic features in order to more directly map source prepositions to target prepositions (Li et al., 2005; Agirre et al., 2009). Turning the tables in a previous study, we used statistical machine translation for helping with WSD (Clematide and Klenner, 2013). However, using imperfect translations

as a machine learning feature resulted in rather moderate improvements for only one of the prepositions in focus. Further research based on parallel corpora is needed here.

On the methodological side, preposition disambiguation with machine learning heavily relies on features derived from the surrounding context of the preposition, but also uses semantic resources such as *WordNet* (Fellbaum, 1998). The best system from the *SemEval* Task on preposition WSD, Kim and Baldwin (2007), combines collocational (surrounding words), syntactic (part of speech tags, chunks) and semantic features (semantic role tags, *WordNet*) in a Maximum Entropy model. They achieve an accuracy of 69.3% in the fine-grained classification task. Their conclusion is that the semantics of prepositions can be learned mostly from the surrounding context and not from syntactic or verb-related properties. O’Hara and Wiebe (2009) use an additional feature, hypernym collocations (*WordNet* hypernyms as collocation provider), to carry out disambiguation relative to either coarse-grained Penn Treebank functional roles or more sophisticated FrameNet roles. They achieve an accuracy of 89.3% given the six Penn Treebank annotated semantic classes. The results in the task of semantic role labelling based on preposition disambiguation are, due to the large number of frame roles (641), low, namely 23.3%.

Hovy et al. (2010) significantly improved on the results of O’Hara and Wiebe (2009); they achieved an accuracy of 91.8% (coarse-grained) and 84.8% (fine-grained using the *SemEval* data). The key to the success of their method seems to lie in the vast amount of different features ranging from suffix information to the holonyms of words. Not all of them are linguistically well motivated (e.g. the first and last two or three letters of each word, respectively). While their approach certainly sets a new standard, its utility to languages other than English is not guaranteed, since some features are geared towards English resources such as *WordNet* (or Roget’s Thesaurus) that are not available in the same quality in other languages. Other features like capitalization are unlikely to be useful for German.

3 Methods

3.1 Resources

As mentioned in Section 2, the Penn Treebank comprises shallow semantic annotations to PPs. There, a distinction is made between several semantic classes of PPs: locative, direction, manner, purpose, temporal, and extent. Unfortunately, none of the large German treebanks (TIGER (Brants and Hansen, 2002), Tüba-D/Z (Telljohann et al., 2004)) provide such a comparable rudimentary scheme that could be a starting point for our case study. There is no resource that we could use, although one is currently being developed by another group (Müller et al., 2011), but it is not yet released. Since we believe that treebanks could benefit from such an additional annotation layer, we decided to work with the largest German treebank, the *Tübinger Baumbank* Tüba-D/Z 7.0. It comprises about 65,000 annotated sentences. Besides phrase structure, topological fields

Table 1. Distribution of the syntacto-semantic functions of *auf* (‘on’) in relation to the syntactic dependencies from the Tüba-D/Z treebank and from the ParZu parser. For Tüba-D/Z, the syntactic function “predicative” is labelled as “p”, “–” is used if there is no governor available (e.g. syntactically not integrated PPs) or if there is another rare syntactic configuration.

Tüba-D/Z							ParZu						
sem\syn	opp	mod	vmod	?mod	– p	Σ	sem\syn	v-pp	n-pp	v-objp	– a-pp	Σ	
LOC	10	45	79	9	6 2	151	LOC	92	29	8 17	5	151	
verbal	119	2	1		2	124	verbal	63	5	47 8	1	124	
nominal		67			2	69	nominal	5	62		2	69	
coll	44	4	7		2	57	coll	20	3	31 1	2	57	
DIR	24	3	8		1	36	DIR	20	10	2 3	1	36	
MOD	3	3	8	4 1		19	MOD	8	5	1 4	1	19	
TLOC		3	12	1		16	TLOC	12	3		1	16	
?	1	4	2		2 1	10	?	5	3	1 1		10	
CAU			3	3 1		7	CAU	6		1		7	
TEM		2	4			6	TEM	4	2			6	
adjectival	1	3	1			5	adjectival	3			2	5	
Σ	202	136	125	17 15 5	500		Σ	238	122	90 37	13	500	

and grammatical functions are also specified. PPs can act as obligatory or optional (opp) complements of verbs, as NP or PP modifiers (mod), or as adjuncts (vmod).

From the ten most frequent prepositions in the Tüba-D/Z we have chosen one with a predominant local meaning (*auf* ‘on’) and one with a broader meaning spectrum (*mit* ‘with’). We randomly sampled 500 occurrences of each.

Dependency Parser Output In order to have a realistic setup for our experiments, we use syntactic evidence derived from the output of a dependency parser for German, the *ParZu* (Sennrich et al., 2009). For syntactically embedded prepositional phrases, this parser applies the following dependency labels: “objp” for verb complements (analogous to the Tüba-D/Z dependency “opp”) and “pp” for modifiers. In Table 1, we show the numbers for verbal (“v-”), nominal (“n-”), and adjectival heads (“a-”). There is quite a number of syntactically not embedded PPs (category “–”). This is mostly due to very complex and long sentences from the newspaper corpus where the parser cannot produce a fully connected dependency structure covering all tokens of a sentence and, therefore, emits forests of parse fragments instead of a parse tree.

Semantics and Annotation of *auf* and *mit* Since we envisage information extraction and question answering as an application context, a coarse-grained classification of the semantics of prepositions, tightly coupled with question words, seems appropriate.

Table 2. Distribution of the syntacto-semantic functions of *mit* ('with') in relation to the syntactic dependencies from the Tüba-D/Z treebank and from the ParZu parser. For Tüba-D/Z, the syntactic function "predicative" is not shown in the table because it appeared only once. For ParZu, the label “–” comprises syntactically not integrated PPs and the label “#” means cases that were not even integrated into a PP. Two dependency types occur only once and they are not shown in the table.

Tüba-D/Z							ParZu							
sem\syn	vmod	mod	opp	?mod	-	Σ	sem\syn	V-pp	N-pp	-	V-objp	A-pp	#	Σ
verbal	8	4	86	1		99	verbal	65	6	5	22	1		99
INS	74	4	7		1	86	INS	69	7	8	1		1	86
MOD	54	4	4	9	2	73	MOD	58	4	5		4	2	73
ORN	1	55		1	2	59	ORN	24	27	8				59
nominal	2	50				52	nominal	13	37	1			1	52
COM	31	6	12	2	1	52	COM	41	6	2		1	2	52
adjectival	1	8	9			18	adject.	13				1	4	18
IDE	7	1		7		15	IDE	15						15
coll	5	1	5	1	1	13	coll	11		2				13
SIZ	5	6	1	1		13	SIZ	8	4			1		13
?	2	3		2	4	11	?	6	3			1	10	
TEM	6		1	1		8	TEM	8						8
Σ	196	142	125	25	11	499	Σ	331	91	34	24	11	7	498

In the case of *auf* (cf. Tab. 1), we distinguish between locative (LOC *where*), directional (DIR *where to*), temporal (TEM *when, how long*), modal (MOD *how*), and causal (CAU *why*) PPs. If the noun in a temporal PP is an event (e.g. *party*), then often a locative or a temporal reading is possible (e.g. *when or where did he laugh? – at his party*). We use TLOC to refer to this usage. If the PP acts as a subcategorized modifier of an adjective or noun, it is annotated with “adjectival” or “nominal” (e.g. *decision on nuclear plants*). In case that the verb governs an otherwise semantically vacuous preposition (*warten auf* ‘to wait for’), the preposition is marked with “verbal”. Finally, any idiomatic expression comprising a PP having a non-compositional meaning like *auf den Putz hauen* ‘to kick up one’s heels’ is annotated as collocational (“coll”). The preposition does not contribute any semantics in these cases. Sometimes no decision was possible (e.g. given sentence fragments, missing global context, unclear semantics), and we used “?” to annotate these instances.

Table 1 shows the distribution of these classes and their syntactic analysis for the preposition *auf*, both relative to the treebank annotation (left-hand side) and the dependency labels of the parser (right-hand side). Local senses form the largest class (151), followed by the syntactic classes “verbal”, “nominal” and “coll”. All other senses of *auf* have lower frequencies. Syntactically, there are three groups to be distinguished: PP complements (opp, 202), NP and PP modification (mod, 136) and adjuncts (v-mod, 125).

In the case of *mit* (cf. Tab. 2), the syntactic labels “verb”, “nominal” and “coll” are used as introduced above for *auf*. The prepositions *auf* and *mit* also share two semantic classes, namely TEM (temporal) and MOD (modal). The other semantic classes of *mit* are: COM for comitative use (*to watch a movie with a friend*), ORN for ornative use (*a man with humor*), SIZ indicating size or proportion (*to demonstrate with 100 people against*), INS for the instrument reading, which is a subclass of MOD (modal) (*to break with a hammer*), and IDE for identity (*with him, hope enters the room* meaning: *he represents/is identical with hope*). Note that *mit* has a more balanced distribution of semantic classes.

Inter-Annotator Agreement For the annotations used in the previous work (Clematide and Klenner, 2013) we have measured inter-annotator agreement in two stages. There was an initial annotation round where one annotator had created the annotation strategy and initial guidelines for one preposition based on existing sense inventories from the literature. The harmonized annotation was then built after discussing the cases where the initial annotations were different. This resulted in further clarifications and refinements of the guidelines, but we also dropped some distinctions that were difficult to apply (e.g. local meaning in a physical sense of contact versus a metaphorical sense).

Table 3. Inter-Annotator agreement of the annotations. We report the percentage of agreeing decisions as well as Cohen’s κ .

Annotations	<i>auf</i>		<i>mit</i>	
	agreeing	κ	agreeing	κ
initial A vs. initial B	74	.67	85	.82
initial A vs. harmonized	85	.81	92	.90
initial B vs. harmonized	86	.83	92	.90
revised harm. A vs. majority	93	.91	96	.96
revised harm. B vs. majority	92	.90		
initial C vs. majority	82	.77	74	.70

As shown in Table 3, Cohen’s κ was high for *mit* and lower, but still substantial for *auf*. There were two problems regarding this harmonized annotation: First, *auf* was missing semantic annotations for nominal and adjectival modifiers. Second, after systematically analyzing the governor lemmas we detected some global inconsistencies regarding the distinction of syntactic classes and semantic classes. As already observed by Tseng (2000), there is no dichotomous categorial distinction between subcategorized functional prepositions and semantic (also called autosemantic) ones in all cases. It is more a difference of degree. In order to give more weight to the semantics of prepositions we revised the guidelines accordingly.

Given these circumstances a third independent annotation C was mandatory. For *auf*, annotator A and B had to revise the “nominal” cases. All annotators

again reviewed the cases with disagreement. The final version used in this paper was built by majority voting. Table 3 gives an overview of the agreement for the different steps of the annotations.

Distributional Semantics: Does it help in preposition classification?

Distributional semantics (DS) is based on the assumption that similar words appear in similar contexts and that the semantic relatedness of words can be measured by a comparison of their contexts (see Erk, 2012, for an overview). Words are represented by vectors in a high-dimensional space and their “positions” can be compared e.g. by the cosine similarity measure. In order to detect the semantic dimensions underlying this huge vector space organised as a co-occurrence matrix, factorisation methods come into play, e.g. Nonnegative Matrix Factorisation (Shashanka et al., 2008). The principle of dimension reduction, which is central to these approaches, allows to cluster words into classes (hard or soft) based on their similarity in vector space.

The general idea in our experiments was to derive, by way of matrix factorisation² and dimension reduction, separate semantic classes of the nouns a) that govern the preposition and b) are governed by the preposition (i.e. the heads of the embedded noun phrases) – called target words, henceforth. We extracted all target words of *mit* and *auf* from the Tüba-D/Z and generated vectors based on 2000 context words. A dependency-parsed version of the DeWac corpus (90 million sentences) (Baroni et al., 2009) was used in order to detect good context words of the target words. Those context words that co-occurred most frequently with as many target words (NP heads) as possible were selected. The vectors were combined into a matrix, where rows represent target words and columns are context words, a single cell records the frequency of a context words co-occurring with the target word.

We then decomposed this matrix with Nonnegative Matrix Factorisation according to different ranks, namely 10, 20 and 50, in two different matrices, a base matrix and a coefficient matrix. The base matrix can be used to determine the class membership of the target words, the classes are produced (soft clustered) by dimension reduction according to the given ranks. We determined for each target word (governor and governed NP head, respectively) the three classes with the highest numerical impact (which determines class membership strength) and used these highest ranked classes as features. The hypothesis was that there is a correlation between these classes and the semantic classes underlying our gold standard.

3.2 Supervised Machine Learning Approach

In order to measure the difficulty of an automatic classification of the syntacto-semantic classes expressed by *auf* and *mit* we conducted several experiments with the Maximum Entropy Modeling tool *MegaM* (Daumé III, 2008). The *Maximum Entropy* approach for classification is also known as *Logistic Regression*

² We worked with the Python implementation NIMFA (Zitnik and Zupan, 2012).

and has been reported to perform very well for PWSD in Tratz and Hovy (2009). For this case study, we focused on simple features gained from the output of the ParZu dependency parser, textual data from the context, and distributional semantics. Some prepositions such as *auf* can govern two different grammatical cases depending on the semantics expressed by the PP. For instance, *auf* with dative is topological whereas *auf* with accusative case is directional. The ParZu parser does not enforce the disambiguation of grammatical case in PPs. In order to have disambiguated grammatical case for each occurrence of *auf*, we used the statistical case tagger based on Conditional Random Fields from Clematide (2013). As for the distributional semantics features, information about the governor of the PP could be provided in 74% (*auf*) and 71% (*mit*) of the samples. Information about the governed head in 78% (*auf*) and 74% (*mit*) of the samples.

In Section 4 we present and analyze the results and performance contribution of the following feature sets:

- **case** Case governed by the preposition (accusative/dative). Only for *auf*.
- **syntax** The syntactic function of the PP taken from ParZu parser output.
- **neighbor** Word, POS (part of speech), and lemma of the preceding and following token.
- **context** Word, POS, and lemma in a window of 5 preceding and following tokens (taken as a bag of words, lemmas or POS).
- **head** Word, POS, and lemma of the head word (typically a noun) of the dependent phrase of the preposition, for instance, the head of *mit Sorgfalt* is *Sorgfalt* ‘care’. In case of coordinated PPs and multi-word heads, the first token was selected.
- **head n** The first 3 classes of a distributional semantics model of rank n of the head.
- **governor** The lemma of the word governing the PP.
- **governor n** The first 3 classes of a distributional semantics model of rank n of the governor.

4 Results and Discussion

The evaluations assess the performance improvement for the multi-class predictions of our 500 annotated prepositions by using different feature sets as evidence. We evaluate against a baseline system which basically predicts the majority class given the lack of any additional evidence. All results are reported as mean accuracy computed by cross-validation (stratified by classes).

4.1 Syntacto-Semantic Classification

We performed a 10-fold cross-validation evaluation for the scenario of predicting the full set of all syntactic and semantic class labels (cf. Tab. 1 and 2). The results of *auf* are shown in Tab. 4a. The best system uses almost all feature sets,

Table 4. Performance of feature sets for syntacto-semantic classification accuracy. The column “Mean” contains the average accuracy computed from the cross-validation sets. $\Delta_{rel_{bs}}$ expresses the relative performance gain. The last row contains the feature set with the best performance. Only systems beating the baseline are shown.

(a) <i>auf</i> ($N = 500$)				(b) <i>mit</i> ($N = 500$)			
System	Mean	SD	$\Delta_{rel_{bs}}$	System	Mean	SD	$\Delta_{rel_{bs}}$
baseline	30.2	0.6		baseline	19.8	0.6	
h(ead)	33.6	3.4	+11.3	g20	20.8	2.7	+5.1
h10	34.6	3.1	+14.6	h20	21.4	7.1	+8.1
g(overnor)	35.2	5.3	+16.6	g10	22.0	3.3	+11.1
h50	37.0	6.9	+22.5	g50	23.4	5.4	+18.2
h20	37.0	6.7	+22.5	h50	25.4	4.8	+28.3
g10	38.8	2.1	+28.5	s(yntax)	26.2	4.5	+32.3
g20	39.8	7.4	+31.8	h(ead)	26.2	4.3	+32.3
g50	43.4	4.9	+43.7	g(overnor)	26.8	5.3	+35.4
s(yntax)	44.4	4.6	+47.0	c(context)	33.0	6.1	+66.7
c(context)	45.4	8.5	+50.3	n(neighbor)	35.6	5.2	+79.8
ca(se)	52.8	2.5	+74.8	s/n/h/g	42.0	5.7	+112.1
n(neighbor)	53.8	5.8	+78.1	s/c/h/h20/g/g20	43.6	6.8	+120.2
s/n/ca/h/g	67.4	4.6	+123.2	s/n/c/h/h50/g/g50	43.6	5.7	+120.2
s/n/ca/h20/g/g50	71.0	4.3	+135.1	s/n/c/h/h20/g/g50	43.6	5.3	+120.2
				s/n/c/h/h20/g/g10	43.6	4.1	+120.2

“case” and “neighbor” are especially strong. The head and governor features are relatively weak, and so are their distributional equivalents. However, the distributional feature sets head 20 and governor 50 contribute to the best system. The best system without any distributional semantics shows a substantially reduced performance.

Table 4b gives the results for *mit*. The overall performance is lower. Head and governor are much stronger for *mit* compared to *auf*. The best performance is reached by rather different feature sets. The rank size, i.e. the number of distributional classes, does not have a strong influence on the results. The best system without distributional semantics performs noticeably worse.

4.2 Semantic Classification

In a further evaluation, we measured how well the purely semantic classes (i.e. the classes without “nominal”, “verbal”, “adjectival”, and “coll”) can be predicted. For *auf* we only have 235 cases with a defined semantic classification, for *mit* we have 306. Due to the smaller data sets we performed 5-fold cross-validation. Table 5a illustrates the problems from the skewed distribution of semantic classes in the case of *auf*: Just guessing the largest class LOC represents a strong baseline decision. Case information adds most of the improvement. However, distributional semantics of the head improves further. The best system

Table 5. Performance of features sets for semantic classification accuracy. The classes are LOC, DIR, MOD, TLOC, CAU, and TEM for *auf*; TEM, MOD, INS, ORN, COM, IDE, and SIZ for *mit*.

(a) <i>auf</i> ($N = 235$)				(b) <i>mit</i> ($N = 306$)			
System	Mean	SD	Δrel_{bs}	System	Mean	SD	Δrel_{bs}
baseline	64.3	1.0		baseline	28.1	0.5	
h20	66.0	5.2	+2.6	g(overnor)	30.7	3.2	+9.3
c(ontext)	66.0	3.4	+2.6	h10	32.0	3.1	+13.9
n(eighbor)	67.2	7.0	+4.6	s(yntax)	33.7	4.8	+19.9
h(ead)	67.2	4.7	+4.6	h20	35.3	5.6	+25.6
ca(se)	77.0	2.3	+19.9	h50	35.6	4.7	+26.7
ca/s/h	80.4	2.3	+25.2	h(ead)	37.6	2.4	+33.7
ca/h20	81.3	2.8	+26.5	n(eighbor)	42.8	4.2	+52.3
				c(ontext)	43.5	3.3	+54.7
				h/s/n/c/g	46.7	7.5	+66.3
				s/n/c/h/h20/g20	51.0	4.3	+81.4

without distributional semantics also includes syntax and performs only slightly worse than the best system.

The less skewed distribution of semantic classes in the case of *mit* allows for a significant improvement over the baseline system. Tab. 5b shows that all feature sets have a beneficial effect. For *mit*, distributional semantics with a rank of 20 increases the results considerably. It is interesting to note that for *auf* the effect of distributional semantics is strong for the syntacto-semantic classification and weak for the semantic classification. For *mit*, we have the opposite situation.

5 Conclusion

We have introduced a coarse-grained annotation scheme for, currently, two German prepositions, *auf* and *mit*. In our experiments with 500 annotated instances of each preposition, we did not only systematically explore the contribution of various contextual and syntactic features commonly used in the field, we also tried to work out the impact semantic information derived from distributional semantics could have on our classification tasks, the syntacto-semantic and semantic disambiguation of the two prepositions. We found that semantic classes derived by matrix factorisation do have an impact although its magnitude is not overwhelming in all cases. Further work is needed to systematically explore the contribution of these approaches. We also intent to carry out experiments with GermaNet (Kunze and Lemnitzer, 2002), the German counterpart of WordNet, in order to find out whether these distinct semantic resources interfere or rather are complementary.

The application of *Active Learning* techniques (Settles, 2012) might help to overcome another problem: the skewed distribution of semantic classes, here of

auf. In order to reliably detect small semantic classes, more training material is needed. Active learning could be used to efficiently gather interesting new instances of such classes.

We also intend to integrate further language resources, e.g. collocation information as provided by services such as *Wortschatz Leipzig*³ or *Digitales Wörterbuch der Deutschen Sprache*.⁴ Bilingual lexicons such as dict.cc⁵ might as well prove fruitful. They contain information about semantically void subcategorized prepositions, for instance *auf jdn warten* is linked to *to wait for sb*. Finally, we will continue to investigate the benefits of cross-lingual information as described in a recent paper (Clematide and Klenner, 2013).

References

- Agirre, E., Atutxa, A., Labak, G., Lersundi, M., Mayor, A., and Sarasola, K. (2009). Use of rich linguistic information to translate prepositions and grammar cases to Basque. In *Proceedings of the XIII Conference of the European Association for Machine Translation (EAMT)*, pages 58–65.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In Boitet, C. and Whitelock, P., editors, *COLING-ACL*, pages 86–90. Morgan Kaufmann Publishers / ACL.
- Baldwin, T., Kordoni, V., and Villavicencio, A. (2009). Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149.
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.
- Brants, S. and Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas.
- Clematide, S. (2013). A case study in tagging case in German: An assessment of statistical approaches. In *Systems and Frameworks for Computational Morphology: Third International Workshop (SFCM 2013)*, Communications in Computer and Information Science, pages 22–34, Berlin, Germany. Springer.
- Clematide, S. and Klenner, M. (2013). A pilot study on the semantic classification of two German prepositions: Combining monolingual and multilingual evidence. In *Proceedings of Recent Advances in Natural Language Processing*, pages 148–155, Hissar, Bulgaria.
- Daumé III, H. (2008). MegaM: Maximum entropy model optimization package. ACL Data and Code Repository, ADCR2008C003.
- Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

³ See <http://wortschatz.uni-leipzig.de>

⁴ See <http://dwds.de>

⁵ See <http://www.dict.cc>

- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- Hovy, D., Tratz, S., and Hovy, E. H. (2010). What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *COLING (Posters)*, pages 454–462.
- Kim, S. N. and Baldwin, T. (2007). MELB-KB: Nominal classification as noun compound interpretation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, page 231–236, Prague, Czech Republic.
- Kipper, K., Snyder, B., and Palmer, M. (2004). Using prepositions to extend a verb lexicon. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, pages 23–29.
- Kiss, T., Keßelmeier, K., Müller, A., Roch, C., Stadtfeld, T., and Strunk, J. (2010). A logistic regression model of determiner omission in PPs. In Huang, C.-R. and Jurafsky, D., editors, *COLING (Posters)*, pages 561–569. Chinese Information Processing Society of China.
- Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, presentation, visualization, application. In *Proceedings of LREC 2002*, pages 1485–1491.
- Li, H., Japkowicz, N., and Barrière, C. (2005). English to Chinese translation of prepositions. In *Canadian Conference on AI*, volume 3501 of *Lecture Notes in Computer Science*, pages 412–416. Springer.
- Litkowski, K. and Hargraves, O. (2007). SemEval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 24–29, Prague, Czech Republic. Association for Computational Linguistics.
- Litkowski, K. C. and Hargraves, O. (2006). Coverage and inheritance in The Preposition Project. In *Third ACL-SIGSEM Workshop on Prepositions*, pages 37–44.
- Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In *HLT*, pages 114–119. Morgan Kaufmann.
- Müller, A., Roch, C., Stadtfeld, T., and Kiss, T. (2011). Annotating spatial interpretations of German prepositions. In *ICSC*, pages 459–466. IEEE.
- O'Hara, T. and Wiebe, J. (2009). Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2):151–184.
- Rehbein, I., Ruppenhofer, J., Sporleder, C., and Pinkal, M. (2012). Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs. In *Proceedings of KONVENS 2012, Vienna, Austria*, pages 89–97.
- Saint-Dizier, P. (2008). Syntactic and semantic frames in PrepNet. In *IJCNLP*, pages 763–768. ACL.
- Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, pages 115–124, Potsdam, Germany.
- Settles, B. (2012). *Active Learning*, volume 6 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool.
- Shashanka, M., Raj, B., and Smaragdis, P. (2008). Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008. Article ID 947438.

- Shilon, R., Fadida, H., and Wintner, S. (2012). Incorporating linguistic knowledge in statistical machine translation: Translating prepositions. In *Proceedings of the EACL-2012 Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 106–114.
- Telljohann, H., Hinrichs, E. W., and Kübler, S. (2004). The Tüba-D/Z treebank: Annotating German with a context-free backbone. In *LREC*, pages 2229–2232. European Language Resources Association.
- Tratz, S. and Hovy, D. (2009). Disambiguation of preposition sense using linguistically motivated features. In *Proceedings of NAACL-HLT 2009, Companion Volume: Student Research Workshop and Doctoral Consortium, SRWS '09*, pages 96–100, Stroudsburg, PA, USA. ACL.
- Tseng, J. L. (2000). *The Representation and Selection of Prepositions*. Phd thesis, University of Edinburgh.
- Zitnik, M. and Zupan, B. (2012). NIMFA: A Python library for nonnegative matrix factorization. *Journal of Machine Learning Research*, 13:849–853.

(Fore)seeing actions in objects

Acquiring distinctive affordances from language

Irene Russo, Irene De Felice, Francesca Frontini, Fahad Khan, Monica Monachini

ILC CNR / Via G. Moruzzi 1, 56124 Pisa
{irene.russo, irene.defelice, francesca.frontini, fahad.khan,
monica.monachini}@ilc.cnr.it

Abstract. In this paper we investigate if conceptual information concerning objects' affordances as possibilities for actions anchored to an object can be at least partially acquired through language. Considering verb-noun pairs as the linguistic realizations of relations between actions performed by an agent and objects we collect this information from the ImagAct dataset, a linguistic resource obtained from manual annotation of basic action verbs, and from a web corpus (itTenTen). The notion of *affordance verb* as the most distinctive verb in ImagAct enables a comparison with distributional data: lemmas ranking based on a semantic association measure reflects that of affordances as the most distinctive actions performed with a specific object.

1 Introduction

Considerable efforts have been dedicated to the acquisition of conceptual knowledge through language, applying corpus linguistics and natural language processing methodologies to verify the results. Several lexicographic measures for collocations and semantic associations have been interpreted in terms of priming effect and distributional approaches have been used for discovering semantic norms (Burgess 1998; Riordan and Jones 2011). In general terms NLP methodologies and resources help in the acquisition of information that cognitive psychology usually collect with time-consuming experiments. Among the features that structure concept representations some of them are quite easily acquired through language while others, like perceptual aspects, are seldom explicitly mentioned in corpora. As a consequence corpora seem promising to model conceptual knowledge that concern for example functional aspects of concepts but not color or size (Bruni et al. 2012).

Affordance is defined as the quality of an object that enables an action: it concerns the relation between a perceptual property of the object and what an agent can do with it. If we interpret it as the range of possibilities for actions anchored to an object affordance plays a central role in embodied robotic approaches to action modeling. Never investigated in computational linguistics, it represents an intriguing aspect of conceptual knowledge hidden in language.

In this work we analyze if conceptual information concerning objects' affordances can be at least partially acquired through language. Our focus is on verb-noun pairs as the linguistic realizations of relations between actions performed by an agent and objects. Using the ImagAct dataset, a linguistic resource obtained from manual annotation of basic action verbs sentences extracted from Italian spoken corpora

(Monachini et al. 2012), we propose the notion of *affordance verb* as a verb that select a distinctive action for a specific object.

This paper is structured as follows: in section 2 previous works on the topic are briefly summarized, in section 3 we report on how the notion of affordances can be investigated linguistically with language resources and corpora. In section 4 a case study about a set of 100 nouns is described, while in 5 we end with conclusions.

2 Previous work

Affordances have been theorized by (Gibson 1979) as the possibilities for action that every environmental object offers. They are different and unique for every living being, in that they are not strictly related to objective properties; rather, they lie on possible ways in which living beings can interact with objects themselves, so they are necessarily related to the capabilities of the agent. For example, a handle on a door can afford grasping for adults, but not for children (“Affordances are properties taken with reference to the observer”, *ibid.*: 143). Humans’ perception is not simply directed to objects: rather, humans perceive what they can (or cannot) do with objects. From this point of view, affordances are preconditions for activity.

After Gibson, many researchers in the field of ecological psychology tried to understand the mechanisms underlying the agent’s ability to perceive affordances. A number of experiments demonstrate that humans can judge if an action is do-able or not-do-able on the basis of intrinsic optical information they receive from the environment (Warren 1984). From this point of view, in order to perceive affordances (e.g. the climbability of a stair) an agent must be able to grasp the relationship between the relevant properties of the environment (the riser height, the tread depth) and the relevant properties of his own action system (his leg length, his body mass).

In Gibson’s theory, affordances are object properties visually perceptible that interact with agent’s properties, in such a way that an activity can be supported: but they exist independently from the agent’s ability to perceive them. Furthermore, according to this line of research, affordances do not change when the agent change his/her goals and plans. Eleanor Gibson (2000; 2003) demonstrates that it is the invariability of the properties of things and events, as well as the distinctiveness of their features, that permits perceptual learning and, as a consequence, the development of knowledges about the world in children.

On the other side, the changeability of affordances and the influence of the final goals on the way they are performed is a basic tenet of robotic studies. In their model (Pastra and Aloimonos 2012) highlight how the goal as the final purpose of an action sequence of any length or complexity influences the movement even when the same tool and object complement are involved (as for *grasping a pencil in order to displace it* vs. *grasping a pencil in order to write*).

In embodied robotics affordances are often conceptualized as a quality of an object, or of an environment, which allow an artificial agent to perform an action.

Environment for artificial agents can be seen as a source of information that help the robot in performing actions, thus reducing the complexity of representation and reasoning (Horton et al. 2012), exactly in the same way an object showing such

properties that afford sitting (e.g. a chair) will help us understand how we can use it (sitting).

Even if the role of direct perception in learning affordances has been stressed in child development studies and in robotics (e.g. exploratory behaviors of agents that test actions without a goal, called babbling stage), object recognition modules can't solve all the intricacies related to the focus on potential actions and so higher level information is generally admitted to model affordances.

Several approaches implicitly mentioned the fact something is push-able or kickable as example of affordances. We think that providing affordances to robots as couples of verb-noun can be useful, if this component is related to the language understanding and visual processing module.

3 Where to find affordances in language: lexical resources and corpora

In last decades there has been a great debate between authors both in defining *what* an affordance is, and in their opinion about *where* an affordance is, whether it is to be considered on the object side, on the agent side, or somewhere in between. Still nowadays we lack a widely shared definition. The common idea is that affordances mostly lie in the (visually) perceivable and physical properties of objects (a concept that corresponds to the “perceived affordances” in Norman 1999): something is inherently graspable when it has a handle. But language is not suitable to reflect perceptual knowledge; as a consequence, we don’t expect to find in corpora data about size, shape or constituent parts of objects. Instead, linguistic resources can be most helpful if we focus on the definition of affordances as possibilities for action.

Considering affordance an inherently relational concept, different approaches can focus either on the object or on the agent. On the one hand, perceptual properties meaningful for an intentional agent can be available: from an environmental perspective, the ball says “I offer hide-ability, push-ability etc. affordances” to an agent. On the other hand, also (relational) properties of the organism-environment system actively discovered by an agent can be available: from an agent perspective, the agent says “I have push-ability affordance” when he sees a ball (Sahin et. al 2007).

Sure enough, this notion can compared with the category of action verbs that have concrete entities as their syntactical direct object. If language reveals a strong association between a specific verb and a specific object, we could consider the verb as a candidate affordance for that given object.

Analyzing corpora through verb-object patterns and measures of associations between nouns and verbs is a good start to explore the notion of affordances as possibilities for actions involving objects.

Non-concrete uses of action verbs, polysemous words and idiomatic expressions make quite difficult derive from corpora a clear picture of all the actions a noun is involved in. Moreover, due to the lack of previous computational linguistics studies on the topic, we don’t have a test set to check if what is obtained is plausible. But even if

there is not explicit encoding of affordances in lexical resources¹ for natural language processing, we find that at least one of them (the ImagAct dataset) can be potentially useful to enrich and structure the dataset of possible affordances (considering them as verb-object couples) and to better investigate how we can find out this knowledge in corpora through measures of semantic associations. Another lexical resource (SIMPLE) is used to enrich ImagAct dataset with information about semantic classes of nouns.

3.1 SIMPLE

SIMPLE (Lenci et al. 2000) is a lexical resource largely based on Pustejovsky's Generative Lexicon (GL) theory (Pustejovsky 1995). GL theory posits that the meaning of each word in a lexicon is structured into components, one of which, the qualia structure, consists of a bundle of four orthogonal dimensions.

These dimensions allow for the encoding of four separate aspects of the meaning of a word: the formal, namely that which allows the identification of an entity, i.e., what it is; the constitutive, what an entity is made of; the telic, that which specifies the function of an entity; and finally the agentive, that which specifies the origin of an entity. These qualia structures plays an important role within GL in explaining the phenomena of polysemy in natural languages. In fact SIMPLE is actually based on the notion of an extended qualia structure, which as the name suggests is an extension of the qualia structure notion found in GL. There is a hierarchy of constitutive, telic, and agentive relations that can hold between semantic units. SIMPLE contains a language independent ontology of 153 semantic types as well as 60k so called “semantic units” or USems representing the meanings of lexical entries in the lexicon. SIMPLE also contains 66 relations organized in a hierarchy of types and subtypes all subsumed by one of the four main qualia roles:

- FORMAL (is-a)
- CONSTITUTIVE, such as ACTIVITY!produced-by
- TELIC, such as INSTRUMENTAL!used-for
- AGENTIVE, such as ARTIFACTUAL!caused-by

3.2 ImagAct

The ImagAct project focuses on high frequency action verbs (approximately 600 lexical entries) of both Italian and English, which represent the basic verbal lexicon of action in the two languages. The purpose of the project is two-fold: (i) to derive information about action verbs adopting a bottom-up approach, from spoken corpora (for Italian: C-ORAL-ROM; LABLITA; LIP; CLIPS; for English: BNC-Spoken); (ii) to construct a multilingual ontology of action, anchored to videos.

¹ However, extractions of affordances in glosses through patterns can be useful. Nonetheless this content is not compulsory neither explicit in dictionary glosses.

(i) The first task mainly aims at eliciting and describing the array of pragmatic situation that each action verb can extensionally denote. Action verbs are central information elements in a sentence and they are the most frequent items in speech (Moneglia and Panunzi, 2007). For reasons of economy, humans adopt the same verbal form to denote different types of events, as emerges from synonymous choices: for example, the verb “to give” in (1) *John takes a present from a stranger* means “to receive, to accept”; but in (2) *John takes Mary the book* it means “to bring”; in (3) *John takes the pot by the handle* it simply means “to grasp”; finally, in (4) *John takes Mary to the station* it means “to conduct”. Furthermore, every language shows a different behavior in segmenting human experience into its action verbal lexicon. However we expect that, in a given language, similar events will be referred to by using the same verb: so “to take” will apply also to *John takes the children to school/his wife to the cinema*, similar to (4); we also expect these consistencies to be found in other languages. These coherent clusters of similar events, denotable by the same set of action verbs, are referred to as action types. This kind of information was derived in the following steps:

- ▲ each occurrence of an action verb is extracted from English and Italian spoken corpora;
- ▲ linguistic contexts of each occurrence are then standardized and reduced in simple sentences (3rd singular form, present tense, active voice);
- ▲ proper instances, in which action verbs are used referring to actions, are distinguished from non-proper instances, in which action verbs do not refer to concrete actions or are used metaphorically);
- ▲ proper occurrences are grouped into action types, keeping granularity to its minimal level, so that each type contains a number of instances referring to similar events (*John takes the glass/the umbrella/the pen etc.*);
- ▲ from all standardized sentences of each type, one best example is chosen (or more than one, if the verb has more than one possible syntactic structure).

(ii) The second task aims at deriving from the data extracted a language-independent ontology for action.

- ▲ the two classifications of action types, derived independently from Italian and English and represented by best examples, are compared and merged into the same ontology of action types. Each node of the ontology is represented by a video exemplifying the whole action type;
- ▲ more than one Italian or English action type can be linked to the same video, when the verbs are local synonyms (as for taking something from someone/receive something from someone).

The result of the procedure described above is a set of short videos, each one corresponding to an action type, representing simple actions (e.g. a man taking a glass on a table). By this list, an user can access to the English/Italian best examples chosen for each type (*John takes the glass/Mary grasp the pen/John prende l'accendino/Mary afferra la matita*) and to all standardized sentences extracted from corpora that have been assigned to that type, that show the actual use of the verb when

referring to a specific type of action. Also, an user can access these data by lemma: for example, searching for the verb “to take”, he will be presented with a number of scenes, showing the different action types associated to that verb, with their related information.

Scenes, and their associated best examples, represent the variation of all action verbs considered and constitute the ImagAct ontology of action. This ontology not only is inherently interlinguistic, because it is derived through an inductive process from corpora of different languages, but also takes into account the intra-linguistic and inter-linguistic variation that characterizes action verbs in human languages.

4 Invariant and distinctive affordances: a case study

Affordances, conceptualized by E. Gibson (2003) as invariant and distinctive properties of objects, can be more easily retrieved in language when considering not just the verbal lemma, but rather manually annotated basic action types that, in terms of variability of the objects involved, can contribute to the understanding of when an affordance verb is distinctive. This is exactly what we find in the ImagAct dataset: in this study, data extracted from this resource are used to derive, for each object lemma considered, not simply a list of co-occurring action verbs, but rather a list of co-occurring verb types.

We consider action verb-noun couples where nouns are direct object in the Italian ImagAct dataset. To help in semantic generalizations across action types, each noun in the theme position has been annotated with its semantic class in SIMPLE. As above mentioned, SIMPLE includes an ontology of 153 semantic types, more fine grained with respect to WordNet supersenses.

This allows us to rank action verbs according to their specificity, i.e. according to the semantic cohesion that depends on the number of semantic classes they belong to. The highest specificity is reached when a verb denotes only one type of action, and when that particular action involves only one semantic type of object. For example, *cogliere* (which roughly corresponds to “to pick, to pick up”) has only one type in ImagAct, and it occurs (in C-ORAL-ROM; LABLITA; LIP; CLIPS spoken corpora) with object lemmas as *fiore*, *fico*, *pomodoro*, *frutto*, *fiorellino* (“flower, fig, tomato, fruit, little flower”). All the object lemmas of *cogliere* pertain to a highly cohesive semantic classes, those of plant and vegetables (SIMPLE types: Vegetable; Plant; Fruit; Vegetal Entity).

The more a verb is specific, the higher is the probability that it will be a suitable candidate as affordance for at least one of its objects. We produce a ranking of action types as the ratio between number of semantic classes the nouns belong to and the number of instances of that action type in the ImagAct dataset.

Two working hypotheses that promote a comparison with corpus data:

H1: More generic action verbs (i.e. verbs with more than one type in ImagAct) express less distinctive affordances for objects in theme position. As a consequence,

objects seldom occur in corpora with generic verbs, or they are less strongly associated with a specific action types of those verbs.

H2: Verbs' types that display a lower noun semantic classes/sentences ratio are the one less distinctive and cohesive. It means that if the type x of the verb y occurs in the Imagact dataset with nouns belonging to two semantic classes in ten sentences, it could be potentially a distinctive affordance for that noun while if the semantic classes of the nouns are 7 it's probably not an affordance.

When verbs in ImagAct have more than one type they are defined as generic verbs; in terms of affordances it means that they are potentially less distinctive because a verb like *prendere* ("to take") can be predicated almost with every object (given physical limits relative to strength, size etc.) while a verb like *stirare* ("to iron") involves a small set of nouns.

We want to test if the notion of affordance that has been discussed above can be also applied looking at the strength of distributional association between words in a corpus.

From ImagAct annotated dataset of 271 verb-noun couples, we select randomly 100 nouns we want to focus on. Looking at affordances in corpora means to verify if what emerges from the ImagAct dataset can be explained in terms semantic associations and ranking of results based on them.

Even if web corpora are not necessarily useful or better than smaller, more cohesive corpora for the extraction of cognitively plausible semantic associations (Lindsey et al. 2007,) we choose to look at word sketches for 100 nouns in the itTenTen, a web corpus of 3.1 billion tokens, accessible through APIs provided by sketchengine.co.uk, because other corpora available for Italian tend not to mention concrete actions, including mainly newspapers' and books' extracts mentioning more frequently abstract activities.

Word sketches are one-page corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff et al. 2004) ordered as list of lemmas on the basis of a measure of salience estimated as the product of Mutual Information (Church and Hanks, 1989) and log frequency.

We extract the 25 most strongly associated words in the preV_N (verbs preceding the noun) pattern since itTenTen is not parsed and this pattern represents a rough approximation of V-OBJ pattern.

For each of the 100 nouns analyzed we report (i) recall as the percentage of verbs in ImagAct retrieved in the 25 words word sketch for that noun; (ii) a score of similarity between rankings in the two resources computed as similarity between sequences of strings with the SequenceMatcher class of the difflib Python module (<http://docs.python.org/2/library/difflib.html>). We report in Table 1 examples relative to three nouns, with rankings of the verbs in ImagAct that have been retrieved in itTenTen:

Noun	recall	rank#ImagAct	rank#itTenTen	SimilarityScore
<i>mela</i> ("apple")	0.307692307692	<i>sbucciare</i> ("to peel") <i>mangiare</i> ("to eat") <i>assaggiare</i> ("to taste") <i>tagliare</i> ("to cut") <i>sbucciare</i> ("to peel")	<i>mangiare</i> ("to eat") <i>assaggiare</i> ("to taste") <i>tagliare</i> ("to cut") <i>sbucciare</i> ("to peel")	0.75
<i>occhiali</i> ("glasses")	0.56	<i>mettere</i> ("to put on") <i>indossare</i> ("to wear") <i>togliere</i> ("to take off") <i>portare</i> ("to wear") <i>mettere</i> ("to put on")	<i>indossare</i> ("to wear") <i>togliere</i> ("to take off") <i>portare</i> ("to wear") <i>mettere</i> ("to put on")	0.6
<i>fiammifero</i> ("match")	0.75	<i>accendere</i> ("to light") <i>spegnere</i> ("to blow out") <i>gettare</i> ("to throw")	<i>accendere</i> ("to light") <i>spegnere</i> ("to blow out") <i>gettare</i> ("to throw")	1

Table 1 – Ordered potential affordance verbs for three nouns

A specific type of a generic verb can be highly distinctive for a small set of nouns. We expect that comparing the ranking of affordance verbs for each nouns with corpus data should enable the discovery of this evidence.

For *orologio* ("watch") *mettere* ("to wear") is the first in the ImagAct ranking, while in itTenTen on the top there are *sincronizzare* ("synchronize") and *guardare* ("to look at"). For *borsetta* ("purse") *prendere* ("to take") is relevant in ImagAct and it is included in the work sketch but at the end of the list. This implies that several usages of generic verbs don't emerge easily in distributional approaches because measures applied for them tend to find the most distinctive items; *to take* is a very frequent verb and it displays a wide set of objects as complement.

We can propose an operative definition of affordance analyzed through linguistic realizations:

General affordances: they concern the most generic verbs in ImagAct, the one with more types and nouns belonging to a wide range of semantic classes. It's a matter of variability: they can be displayed potentially by every objects so they are not distinctive of any in particular.

Specific affordances: they are the canonical/peculiar activities a specific object is involved in, like *open* for *bottle*. Our hypothesis is that measures of word association help to find them in corpora.

The qualitative evaluation of the results highlights that data about affordances extracted from the ImagAct dataset, if compared with distributional data extracted

from corpora, offer a methodology to extract in the future plausible affordances for nouns referring to objects not mentioned in the ImagAct dataset, with some caveats:

- extracting activity verbs in the first positions when the rank is based on semantic association measures is useful, but the order should not be interpreted in terms of plausibility of items as affordance verbs. However the mean similarity score between the two resources is 0.82. It means that manually annotated data about basic action types mirror rankings based on semantic association between words;
- the fact that the mean recall is not high (0.36) should be further investigated. Since we don't have a test set of ordered affordances, at the moment we cannot set a threshold, both in ImagAct and in word sketches, to exclude some verbs as not affordance verbs.

Since ImagAct does not contain all the activity verbs that could be affordances, the low recall value is not so indicative of the quality of the resource or of the notion of affordance verb based on it. More suggestive is the value about the similarity between ranking, because it implements practically an idea of affordance as distinctive actions an object can be involved in and opens to other interesting work hypotheses on the interplay between cognitive psychology and computational linguistics.

5 Conclusions

In ecological psychology organisms can perceive whether a specific action is do-able or not-do-able; we don't perceive just objects but the action possibilities (e.g. that something is climbable, passable, etc.) offered by the environment. In this paper we investigate if and how these action possibilities can be extracted from a corpus as affordance verbs.

We propose a corpus based concept of affordances focused on verb-noun couples and the notion of affordance verb as the most distinctive verb for a noun in a dataset of manually annotated action types extracted from spoken corpora. This guiding notion enables a comparison with distributional data from a web corpus that reveal how words ranking based on semantic associations mirror affordance as the most distinctive action an object can be involved in.

We don't forget that ImagAct does not comprise all the usages existing in a language and so, even if a verb is preferable with a noun, in theory a lot of other objects can be the theme of that action. Also, concerning possibilities for actions, the extraction of affordances from a corpus poses the issue of unattested instances that can be cognitively plausible but, just as a matter of chance, are not available (or not strongly associated with a word).

As future work we propose to test different association measures in order to see how they may be helpful in extracting different types of affordances from corpora. We also want to elicit speakers' judgments on verb-noun couples in order to have a test set for affordances.

References

- Bruni, E., Uijlings, J., Baroni, M. and Sebe, N. (2012), Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. Brave New Idea paper. In: *Proceedings of MM 12* (20th ACM International Conference on Multimedia), New York NY: ACM, pp. 1219-1228.
- Burgess, C. (1998) From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30: 188-198.
- Church, K., and Hanks, P. (1989) Word Association Norms, Mutual Information and Lexicography. In: Proceedings, 27th Meeting of the ACL, pp. 76-83. Gibson, E. J. (2000) Perceptual Learning in Development: Some Basic Concepts. *Ecological Psychology*, 12(4): 295-302.
- Gibson, E. J. (2003) The World Is So Full of a Number of Things: On Specification and Perceptual Learning. *Ecological Psychology*, 15(4): 283-287.
- Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Horton T. E., Chakraborty A., St. Amant R. (2012) Affordances for robots: a brief survey. *Avant*, 3(2): 70-84.
- Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004) The Sketch Engine. In: Williams G. and S. Vessier (eds.), *Proceedings of the XI Euralex International Congress*, July 6-10, 2004, Lorient, France, pp. 105-111.
- Lenci, A. Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000) SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4): 249-263.
- Lindsey, R., Veksler, V. D., Grintsvayg, A., and Gray, W. D. (2007) Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness. In: *Proceedings of the 8th International Conference on Cognitive Modeling*, Ann Arbor, MI, pp. 279-284.
- Moneglia, M. and Panunzi, A. (2007) Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project. In: Alcántara Plá, M. and Declerk, T. (eds.) *Proceeding of the International Workshop on the Semantic Representation of Spoken Language (SRSL7)*, Salamanca, 12 November - 16 November 2007, pp. 51-58.
- Moneglia, M. and Panunzi, A. (2011) Specification for the annotation of verb occurrences in the ImagAct project. Technical Report Draft, <http://lablita.dit.unifi.it/projects/IMAGACT/folder.2010-11-25.7365875310/>. Accessed 06/06/2013.
- Monachini, M., Frontini, F., De Felice, I., Russo, I., Khan, F., Gagliardi, G., and Panunzi A. (2012) Verb interpretation for basic action types: annotation, ontology, induction and creation of prototypical scenes. In: *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon - CogALex-III - COLING 2012*, Mumbai, India, 15 December 2012, pp. 69 - 80.
- Norman, D. A. (1999) Affordance, conventions, and design. *Interactions*, 6: 38-42.

- Pastra, K. and Yiannis Aloimonos (2012) The Minimalist Grammar of Action. *Philosophical Transactions of the Royal Society B*, 367(1585):103-117.
- Pustejovsky, J. (1995) *The Generative Lexicon*. Cambridge, Massachusetts: The MIT Press.
- Riordan, B., and Jones, M. N. (2011) Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2): 1-43.
- Sahin, E., Cakmak, M., Dogar, M.R., Ugur, E. and Ucoluk, G. (2007) To afford or not to afford: A new formalization of affordances towards affordance-based robot control. *Adaptive Behavior*, 15(4): 447-472.
- Warren,W. H. (1984) Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, 10: 683–703.

Extracting Opinion and Factivity from Italian political discourse

Rodolfo Delmonte¹, Daniela Gîfu², Rocco Tripodi¹

¹Ca' Foscari University, Department Language Science,

Ca' Bembo, dd. 1075, 30123, Venice

delmont@unive.it, rocco.trip@gmail.com

²"Alexandru Ioan Cuza" University, Faculty of Computer Science,

16, General Berthelot St., 700483, Iași

daniela.gifu@info.uaic.ro

Abstract. The success of a newspaper article for the public opinion can be measured by the degree in which the journalist is able to report and modify (if needed) attitudes, opinions, feelings and political beliefs. We present a symbolic system for Italian, derived from GETARUNS, which integrates a range of natural language processing tools (also available in the public domain) with the intent to characterise the print press discourse from a semantic and pragmatic point of view. This has been done on some 500K words of text, extracted from three Italian newspapers in order to highlight their stance on the deep political crisis situation which brought to the change of government that took place at the end of 2011. We tried two different approaches: a lexicon-based approach for semantic polarity using off-the-shelf dictionaries with the addition of manually supervised domain related concepts. Another one is a feature-based semantic and pragmatic approach, which computes propositional level analysis on the basis of the verbal complex and other semantic markers to process factuality and subjectivity. Results are quite revealing and contradict the otherwise common knowledge about the political stance of each newspaper on such topic.

Keywords: journal opinion, sentiment analysis, political discourse, lexico-pragmatic analysis, syntactic analysis, semantic mapping.

1 Introduction

In this paper we discuss paradigms for using linguistic interpretations of discourses in a method for opinion mining and sentiment analysis, and we present a concrete implementation in the analysis of newspaper text, by a light scaled version of a system for text understanding called GETARUNS [3]. The system thus realized is comparable to a pipeline of modules which are also freely available on the web. We focus on three aspects critical to a successful analysis: creation of large quantities of reasonably good training data, lexical-semantic and syntactic analysis. Measuring the polarity of a text is usually done by text categorization methods which rely on freely available resources. However, we assume that in order to properly capture opinion

and sentiment [6,10,11,17] expressed in a text or dialog, - that we also assume to denote the same field of research, and is strictly related to “subjectivity” analysis - any system needs a linguistic text processing approach that aims at producing semantically viable representation at propositional level. In particular, the idea that the task may be solved by the use of Information Retrieval tools like Bag of Words Approaches (BOWs) is insufficient. BOWs approaches are sometimes also camouflaged by a keyword based Ontology matching and Concept search [10], based on SentiWordNet (*Sentiment Analysis and Opinion Mining with WordNet*) [2]– more on this resource below -, by simply stemming a text and using content words to match its entries and produce some result [16]. Any search based on keywords and BOWs is fatally flawed by the impossibility to cope with such fundamental issues as the following ones, which Polanyi and Zaenen [12] named contextual valence shifters:

- presence of negation at different levels of syntactic constituency;
- presence of lexicalized negation in the verb or in adverbs;
- presence of conditional, counterfactual subordinators;
- double negations with copulative verbs;
- presence of modals and other modality operators.

It is important to remember that both Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) [16] systematically omit function or stop words from their classification set of words and only consider content words. In order to cope with these linguistic elements we propose to build a propositional level analysis directly from a syntactic constituency or chunk-based representation. We implemented these additions on our system thus trying to come as close as possible to the configuration which has been used for semantic evaluation purposes in challenges like Recognizing Textual Entailment (RTE) and other semantically heavy tasks [1,4]. The output of the system is an xml representation where each sentence of a text or dialog is a list of attribute-value pairs. In order to produce this output, the system makes use of a flat syntactic structure and a vector of semantic attributes associated to the verb compound at propositional level and memorized. An important notion required by the extraction of opinion and sentiment is also the distinction of the semantic content of each proposition into two separate categories: objective vs. subjective.

This is obtained by searching for factivity markers again at propositional level [14]. In particular we take into account the following markers: modality operators such as intensifiers and diminishers, modal verbs, modifiers and attributes adjuncts at sentence level, lexical type of the verb (from ItalWordNet classification, and our own), subject’s person (if 3rd or not), and so on.

As will become clear below, we are using a lexicon-based [9,15] rather than a classifier-based approach, i.e. we make a fully supervised analysis where semantic features are manually associated to lemma and concept of the domain by creating a lexicon out of frequency lists. In this way the semantically labelled lexicon is produced in an empirical manner and fits perfectly the classification needs. This was needed in particular after we realized that available lexica were totally insufficient to cover the domain of political discourse. Of course we are aware of the intrinsic deficiencies of any such approach whenever irony, humour and figurative language is the target to be discovered, but see [18;19] on the topic.

The paper is structured as follows. Section 2 describes the system for multi-dimensional political discourse analysis. Section 3 discusses an example of comparative analysis of print press discourses collected before, during and after Berlusconi's resignation in favour of Monti's nomination as President of Italian Government (October 12 – December 12, 2011). Finally, section 4 highlights interpretations anchored in our analysis and presents a conclusion.

2 Print press discourse

Mirror of contemporary society, located in permanent socio-cultural revaluation, the texts of print press can disrupt or use a momentary political power. In contemporary society, the struggle's stake is no longer the social use of technology, but it is the huge production and dissemination of representations, informations and languages.

At present, the legitimacy of competence and credibility or reputation of political authority is increasingly in competition with mediatic credibility and the charisma already confirmed in public space. In political life we see how „heavy” actors are imposed, benefiting preferential treatment in their publicity and/or how insignificant actors, with reduced visibility, are ignored, even marginalized, notwithstanding their possibly higher reputation.

Free print press, in its various forms, assigns political significance to institutional activities and events in their succession; it should form the political life of a nation, from objective information to become the subject of public debate. In this case, the role of print press would be double:

1. secure information as a credible discourse to end a rumor; 2. enter politics in language forms, so they become consistently interpretable in a symbolic model of representations.

The press is designed to legitimize the actions of politicians, attending their visibility efforts, confirming or increasing their reputation. Print press includes essentially political discourses, containing both a specific orientation and a political commitment. The reader has the possibility to choose what and when to read, leaving time to reflection, too. Disproportionality is a risk to the reality described.

It is part of common sense understanding that political news are often biased, in particular when the owner of the media is a political actor himself. So the aim and target of political discourses in the press becomes persuading their audience.

No wonder why the people in power, if they intend to govern in peace, try to curb the enthusiasm of the media. Most of the times, through excellence in the elections, the print press is focused on topical issues, leading topics of public interest and events of internal and external social life. However, the perception of social reality depends on how it is presented. So the newspaper, like any commercial product, is dependent on aesthetic presentations that may distort any event-selection alternative to news items which are sensational and, often, negative (i.e. our comparative study).

3 The System GETARUNS

In this section we will present a detailed description of the system for Italian that we used in this experiment. The system is derived from GETARUNS, a multilingual

system for deep text understanding with limited domain dependent vocabulary and semantics, that works for English, German and Italian and has been developed in the past 20 years or so in several publications and conference presentations[3,5]. The deep version of the system, that works with a symbolic approach, has been scaled down in the last ten years to a version that can be used with unlimited text and vocabulary, again for English and Italian. The two versions can work in sequence in order to prevent failures of the deep version. Or they work separately to produce less constrained interpretations of the text at hand.

The "shallow" scaled version is a pipeline adapted for the opinion and sentiment analysis and results have already been published for English [6]. Now, the new current version which is used with Italian has been made possible by the creation of the needed semantic resources, in particular a version of SentiWordNet adapted to Italian and heavily corrected and modified. This version (see below) uses weights for the English WordNet and the mapping of sentiment weights has been done automatically starting from the linguistic content of WordNet glosses. However, this process has introduced a lot of noise in the final results, with many entries with a totally wrong opinion evaluation. In addition, there was a need to characterize uniquely only those entries that have a "generic" or "commonplace" positive, or negative meaning associated to them in the specific domain. This was deemed the only possible solution to the problem of semantic ambiguity, which could only be solved by introducing a phase of Word Sense Disambiguation which was not part of the system. However this was not possible for all entries. So, we decided to erase all entries that had multiple concepts associated to the same lemma, and had conflicting sentiment values. We also created and added an ad hoc lexicon for the majority of concepts (some 3000) contained in the texts we analysed, in order to increase the coverage of the lexicon. This was done again with the same approach, i.e. labelling only those concepts which were uniquely intended as one or the other sentiment, restricting reference to the domain of political discourse.

The system has been lately documented by our participation in the EVALITA (*Evaluation of NLP and Speech Tools for Italian*) challenge¹. It works in a usual NLP pipeline: the system tokenizes the raw text and then searches for Multiwords. The creation of multiwords is paramount to understanding specific domain-related meanings associated to sequences of words. This procedure is then extended to NER (*Named Entity Recognition*), which is performed on the basis of a big database of entities, lately released by JRC (*Joint Research Centre*) research centre.² Of course we also use our own list of entities and multiwords.

Words that are not recognized by simple matching procedures in the big wordform dictionary (500K entries), are then passed to the morphological analyser. In case also this may fail, the guesser is activated, which will at first strip the word of its affixes. It will start by stripping possible prefixes and then analysing the remaining portion; then it will continue by stripping possible suffixes. If none of these succeeds, the word will be labelled as foreign word if the final character is not a vowel; a noun otherwise. We then perform tagging and chunking. In order to proceed to the semantic level, each nominal expression is classified at first on the basis of the assigned tag: proper nouns

¹ <http://www.evalita.it/>

² <http://irmm.jrc.ec.europa.eu/>

are classified in the NER task. The remaining nominal expressions are classified using classes derived from ItalWordNet (*Italian WordNet*)³. In addition to that, we have compiled specialized terminology databases for a number of common domains including: medical, juridical, political, economic, and military. These lexica are used to add a specific class label to the general ones derived from ItalWordNet. And in case the word or multiword is not present there, to uniquely classify them. The output of this semantic classification phase is a vector of features associated to the word and lemma, together with sentence index and sentence position. These latter indices will then be used to understand semantic relations intervening in the sentence between the main governing verb and the word under analysis. Semantic mapping is then produced by using the output of shallow parsing and functional mapping algorithms which produce a simplified labelling of the chunks into constituent structure. These structures are produced in a bottom-up manner and subcategorization information – coming again from a fully specified subcategorization lexicon of Italian of some 17K entries - is only used to choose between the assignments of functional labels for argumenthood. In particular, choosing between argument labels like SUBJ, OBJ2, OBL which are used for core arguments, and ADJ which is used for all adjuncts requires some additional information related to the type of governing verb.

The first element for Functional Mapping is the Verbal Complex, which contains all the sequence of linguistic items that may contribute to its semantic interpretation, including all auxiliaries, modals, adverbials, negation, clitics. We then distinguish passive from active diathesis and we use the remaining information available in the feature vector to produce a full-fledged semantic classification at propositional level. The semantic mapping includes, beside diathesis:

- Change in the World; Subjectivity and Point of View; Speech Act; Factuality; Polarity.

At first we compute Mood and Tense from the Verbal Compound (hence VC) which, as said before, may contain auxiliaries, modals, clitics, negation and possibly adverbials in between. From Mood_Tense we derive a label that is the compound tense and this is then used together with Aspectual lexical properties of the main verb to compute Change_in_the_World. Basically this results into a subclassification of events into three subclasses: Static, Gradual, Culminating. From Change_in_the_World we compute (Point_of_)View, which can be either Internal (Extensional/Intensional) or External, where Internal is again produced from a semantic labeling of the subcategorized lexicon along the lines suggested in linguistic studies, where psych(ological) verbs are separated from movement verbs etc., . Internal View then allows a labeling of the VC as Subjective for Subjectivity and otherwise, Objective. Eventually, we look for negation which can be produced by presence of a negative particle or be directly in the verb meaning as lexicalised negation. Negation, View and Semantic Class, together with presence of absence of Adverbial factual markers are then used to produce a Factuality labeling.

One important secondary effect that carries over from this local labeling, is a higher level propositional level ability to determine inferential links intervening between propositions. Whenever we detect possible dependencies between adjacent VCs we check to see whether the preceding verb belongs to the class of implicatives. We are

³ http://www.ilc.cnr.it/iwndb/iwndb_php/

here referring to verbs such as “refuse, reject, hamper, prevent, hinder, etc.” on the one side, and “manage, oblige, cause, provoke, etc.” on the other (for a complete list see [14]). In the first case, the implication is that the action described in the complement clause is not factual, as for instance in “John refused to drive to Boston”, from which we know that “John did not drive to Boston”. In the second case, the opposite will apply, as in “John managed to drive to Boston”.

Eventually, we built a specialized coreference module which only look for mentions related to the two main political entities under scrutiny: Monti and Berlusconi. We searched for possible coreferential mentions to Berlusconi in English and American newspapers and we found the following possibly partial list: “Broadcaster; Businessman; Caretaker; Chairman; Chief; Creator; Entrepreneur; Executive; Film_Industrialist; Leader; Magnate; Millionaire; Minister; Mogul; Negotiator; Owner; Politician; Premier; President; Tycoon; Winner”. So we decided to include possible frequently recurring coreferential mentions in Italian newspaper in order to capture their almost real presence in texts and we came up with the following lists:

- Monti : Mario, Mario_Monti, Professore, nuovo_premier, presidente_del_PDL, capo_del_governo, premier, capo_del_Governo, presidente_del_Consiglio, presidente_del_consiglio, Presidente_del_Consiglio
- Silvio, Silvio_Berlusconi, Cavaliere, ex_premier, premier, presidente_del_PDL, capo_del_governo, capo_del_Governo, presidente_del_Consiglio, presidente_del_consiglio, Presidente_del_Consiglio

Notice the concept “premier” which can be applied to both entities but only in different time periods. In fact there are two additional temporally related concepts “ex_premier” applied only to Berlusconi, and “nuovo_premier”/new_premier applied to Monti. What we do in the case of “premier” is to use two separate list of concepts, one containing the concept in the period in which Berlusconi was still in power, and the other list, where the concept has been erased, when he resigned.

3 A comparative study

Whereas the aims of syntax and semantics in this system are relatively clear, the tasks of pragmatics are still hard to extract automatically. But, we have to recognize the huge relevance of pragmatics in analyzing political texts.

3.1 The corpus

For the elaboration of preliminary conclusions on the process of the change of the Italian government and president of government, we collected, stored and processed - partially manually, partially automatically -, relevant texts published by three national on-line newspapers having similar profiles⁴.

For analytical results to be comparable to those taken so far by second author [20,21], we needed a large corpus, especially considering five rigorous criteria that we list below:

⁴ www.corriere.it, www.liberoquotidiano.it, www.repubblica.it

1. Type of message: Selection of newspapers was made taking into account the type of opinions circulated by the Editorial: pro, against Berlusconi and impartial. The following newspapers were thus selected: Corriere della Sera (also called The People Newspaper); Libero (usually pro Berlusconi); and La Repubblica – (usually strongly against Berlusconi).

2. Period of time: The interval time chosen should be large enough to capture the lexical-semantic and syntactic richness found in the Italian press. It was divided into three time periods. We specify them here below with their abbreviations, used during analysis.

A month before the resignation of Berlusconi (12 November 2011), abbreviated to OMBB: October 12 to November 11, 2011;

The period between the presentation of Berlusconi's resignation and the appointment of Mario Monti as premier of the Italian Government, abbreviated with PTMB: 12 to 16 November 2011;

A month after the resignation of Berlusconi, abbreviated with OMAB: November 17 to December 12, 2011.

Two keywords were commonly used to select items from the Italian press, that is the name of the two protagonists: (Silvio) Berlusconi (and appellations found in newspaper articles: Silvio, Il Cavaliere, Il Caimano) and (Mario) Monti. We tried to select an archive rich enough for each of the three newspapers (meaning dozens of articles per day), the selected period of time as the one of interest, between average values. Text selection was made taking into account the subcriterion *Ordina per rilevanza* (order articles by relevance) that each web page of the corresponding newspapers made available. We then introduced a new subcriterion of selection: storing articles in the first three positions of each web page for every day of the research period. In particular we collected on average 250 articles per newspaper, that is 750 articles overall. Also number of tokens are on average 150K tokens per newspaper, i.e. 450K tokens overall. Computation time on a tower MacPro equipped with 6 Gb RAM and 1 Xeon quad-core was approximately 2 hours.

3.2 The syntactic and semantic analysis

In Fig. 1 below, we present comparative semantic polarity and subjectivity analyses of the texts extracted from the three Italian newspapers. On the graph we show differences in values for four linguistic variables: they are measured as percent value over the total number of semantic linguistic variables selected from the overall analysis and distributed over three time periods on X axis. Measures of polarity and subjectivity can only be measured in a relative and not in an absolute way. To display the data we use a simple difference formula, where Difference value is subtracted from the average of the values of the other two newspapers for that class. Differences may appear over or below the 0 line. In particular, values above the 0x axis mean they assume positive or higher than values below the 0x axis, which have a negative import. The classes chosen are respectively: 1. propositional level polarity with NEGATIVE value; 2. factivity or factuality computed at propositional level, which contains values for non factual descriptions; 3. subjectivity again computed at propositional level; 4. passive diathesis. We can now evaluate different attitudes and

styles of the three newspapers with respect to the three historical periods: in particular we can now appreciate whether the articles report facts objectively without the use of additional comments documenting the opinion of the journalist. Or if it is rather the case that the subjective opinion of the journalist is present only in certain time spans and not in others. Chronological difference is indicated by the three separate contiguous subsets into which the values are displayed below, OMBB comes Before OMAB, and the period in Between is placed at the for its lower intermediate significance.

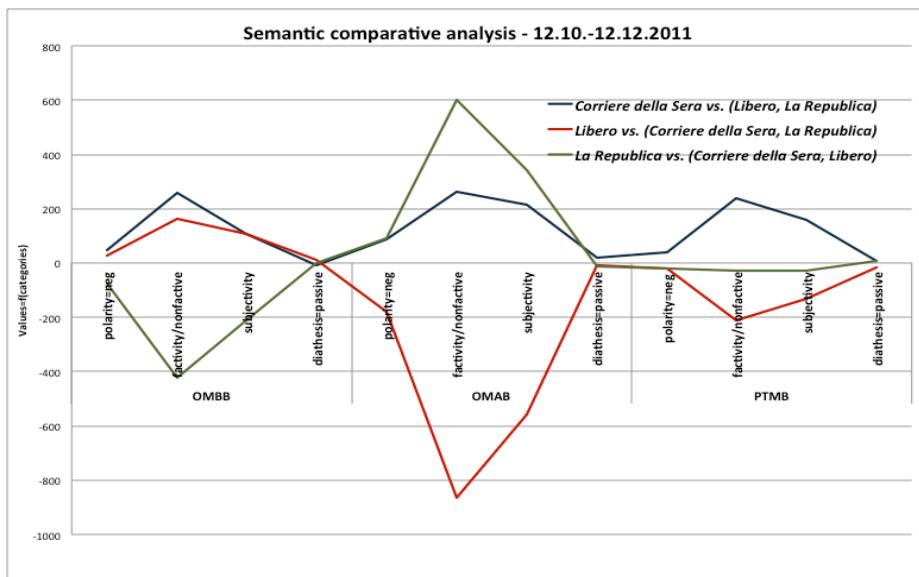


Fig. 1. Comparative semantic polarity analysis of three Italian newspapers.

So for instance, *Corriere*, the blue or darker line, has higher nonfactive values in two time spans, OMBB and PTMB; *Repubblica* values soar in OMAB. In the same period *Libero* has the lowest values; whereas in OMBB, *Libero* and *Corriere* have the highest values when compared with *Repubblica*. PTMB clearly shows up as a real intermediate period of turmoil which introduces a change: here *Repubblica* becomes more factual whereas *Libero* does the opposite. Subjectivity is distributed very much in the same way as factuality, in the three time periods even though with lesser intensity. *Libero* is the most factual newspaper, with the least number of subjective clauses. Similar conclusion can be drawn from the use of passive clauses, where we see again that *Libero* has the lowest number. The reasons for *Libero* having the lowest number of nonfactive clauses in OMAB, needs to be connected with the highest number of NEGATIVE polarity clauses, which is related to the nomination of Monti instead of Berlusconi, and is felt and is communicated to its readers as less reliable, trustable, trustworthy. Uncertainty is clearly shown in the intermediate period, PTMB, where *Corriere* has again the highest number of nonfactual clauses.

3.3 The pragmatic analysis

We show in this section the results outputted by GETARUNS when analysing the streams of textual data belonging to the three sections of the corpus (presented in section 4.1). In Fig. 2 we represent comparative differences between the three newspaper in the use of three linguistic variables for each time period. In particular, we plotted the following classes of pragmatic linguistic objects: 1. references to Berlusconi as entity (Silvio, Silvio_Berlusconi, Berlusconi, Cavaliere, Caimano); 2. references to Monti as entity (Monti, prof_Monti, professore, Mario_Monti, super_Mario); 3. negative words, that is overall negative content words. To capture coreference mentions to the same entity we built a specialized coreference algorithm.

With one month before Berlusconi's resignation (OMBB), we can highlight the opinions of the three dailies as follows: *Corriere della Sera* and *Libero* are concerned mostly with Berlusconi (see *Berlusconi occurrences*), with a remarkable difference however in terms of positive – *Libero* – vs negative – *Corriere* – comments. After Berlusconi resigned (OMAB) *Libero* is more concerned than the other two newspapers on Monti: negative appreciation is always higher with *Libero* and not with the other two. This can clearly be seen from the sudden dip of positive words. Finally in the intermediate period, both *Libero* and *Corriere* seem to be the most concerned with the new government, with the highest number of negative comments.

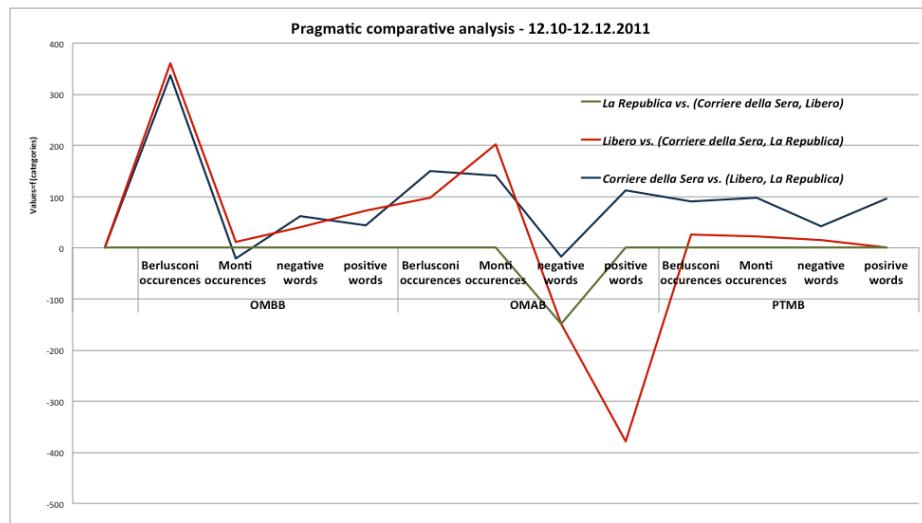


Fig. 2. Comparative pragmatic analysis of three Italian newspapers.

As shown in Fig.2, measuring the overall attitude with positive vs. negative affective content for each newspaper allows a clear cut subdivision in the three time periods. Table 1 below shows the same data in a more perspicuous manner.

Newspaper / time period	Corriere della Sera		Libero		La Repubblica	
	positive	negative	positive	negative	positive	negative
OMBB	35.81% 123.45	35.02% 39	32.28% 111.29	32.35% 36.03	31.91% 110	32.49% 36.32
PTMB	43.62% 134.6	45.71% 45.8	24.43% 75.4	22.75% 31.6	31.95% 98.6	31.54% 31.6
OMAB	37.78% 113.46	34.28% 37.92	27.56% 82.77	31.4% 34.73	34.69% 104.12	32.62% 37.96

Table 1. Sentiment analysis of three Italian newspapers

The percentages from Table 1 are organized as follows. Positive values are computed along time line distribution: for each newspaper, we compute the percentage referred to the each time slot. For instance, in OMBB positive values are distributed with the following subdivision in percent values: 35.81 for *Corriere*, 32.28 for *Libero*, and 31.91 for *Repubblica*. In other words, in OMBB, *Corriere* uses the most number of positive words. In fact, as can be easily noticed, *Corriere* is the newspaper that uses most positive keywords in all the three time periods. On the contrary, *Libero* is the newspaper that uses the least number of positive keywords. *Repubblica* lies in the middle. The second number included in the same cell is needed to account for differences in number of tokens, and this in turn is due to differences in number of days considered for each time period: 31 for OMBB, 5 for PTBM and 26 for OMAB. Average values for each time period for each newspaper in part confirm percent values but also give a deepest idea of the actual numbers at play.

Negative opinions are computed in the same way. These data can be interpreted as follow:

One month before Berlusconi's resignation (OMBB), both *Libero* and *Corriere della Sera* have more positive contents than *La Repubblica*, which can be interpreted as follows: Berlusconi's Government is considered a good one; in addition, *Libero*, has the lowest percentage of negative opinions about the current economic situation. In the intermediate period between Berlusconi's resignation and nomination of the new Prime Minister, Mario Monti (PTMB) we see that *Corriere* has by far the highest percentage of positive opinions, whereas *Libero* has the lowest. The other period, one month after the nomination of new prime minister, Mario Monti, (OMAB), we assist to a change of opinions. *Corriere della Sera* becomes more positive than other newspapers and also negative opinions are much higher: the new prime minister seems a good chance for the Italian situation; however, the economic situation is very bad. *Libero* – the newspaper owned by Berlusconi - becomes a lot less positive and less negative than the other two. This situation changes in the following time period, where *Libero* increases in positivity – but remains always the lowest value – and in negativity, but remains below the other two newspaper, on average. This can be regarded as a distinctive stylistic feature of *Libero* newspaper. As a whole, we can see that *Repubblica* is the one that undergoes less changes, if compared to *Libero* and *Corriere* which are the ones that undergo most changes in affective attitude.

We also saw above that *Libero* is the newspaper with the highest number of nonfactual and subjective clauses in the OMAB time period: if we now add this information to the one derived from the use of positive vs. negative words, we see that the dramatic change in the political situation is no longer shown by the presence of a strong affective vocabulary, but by the modality of presenting important concepts related to the current political and economic situation, which becomes vague and less factual after Berlusconi resigned.

Eventually, we were interested in identifying semantic linguistic common area (identification of common words), also called common lexical fields, and their affective import (positive or negative). From previous tables, it can be easily noticed that all three newspapers use words with strong negative import, but with different frequency. Of course, this may require some specification, seeing the political context analyzed. So we decided to focus on a certain number of specialized concepts and associated keywords that we extracted from the analysis to convey the overall attitude and feeling of the political situation. We collected in Table 2 below all words related to “Crisis Identification” (CIW for short) and noted down their absolute frequency of occurrence for each time interval.

<i>CIW OMBB</i>	<i>Corriere</i>	<i>Libero</i>	<i>Repub.</i>	<i>CIW OMAB</i>	<i>Corriere</i>	<i>Libero</i>	<i>Repub.</i>
1. crisis	124	71	94	1. crisis	50	21	110
sacrifice	4	14	4	sacrifice	9	23	16
rigour	5	4	4	rigour	23	18	10
austerity	0	6	6	austerity	6	2	0
2. battle	6	12	14	2. battle	14	4	8
dissent	2	8	8	dissent	0	4	0
dictator/ship	2	10	18	dictator/ship	2	6	2
3. fail/ure	8	13	9	3. fail/ure	21	8	15
collapse	10	6	12	collapse	8	2	4
drama/tic	12	14	18	drama/tic	4	0	8
dismiss/al	45	39	20	dismiss/al	3	2	15

Table 2. Crisis Identification words in two time periods

If we look at the list as being divided up into three main conceptualizations, we may regard the first one as denouncing the critical situation, the second one as trying to indicate some causes; and the last one as being related to the reaction to the crisis. It is now evident what the bias of each newspaper is, in relation to the incoming crisis:

- *Corriere della Sera* feels the “crisis” a lot deeper before Berlusconi’s resignation, than afterwards when Monti arrives; the same applies to *Libero*. *La Repubblica* feels the opposite way. However, whereas “austerity” is never used by *La Repubblica* after B.’s resignation and it was used before it, this is the opposite of what *Corriere della Sera* does, the word appears only after B.’s resignation, never before. As to the companion word “sacrifice”, *Libero* is the one that uses it the most, and as expected its appearance increases a lot after B.’s resignation, together with the companion word

“rigour” that has the same behaviour. This word confirms *Corriere*’s attitude towards Monti’s nomination: it will bring “austerity, rigour and sacrifice”.

- in the second half, the other interesting couple of concepts is linked to “battle, dissent, dictator”. In particular, “battle” is used in the opposite way by *Corriere della Sera* when compared to the other two newspapers: the word appears more than the double in the second period, giving the impression that the new government will have to fight a lot more than the previous one. As to “dissent”, all three newspapers use it in the same manner: it disappears in both *Corriere della Sera* and *La Repubblica*, and it is halved in *Libero*. Eventually the “dictator/ship” usually related to B. or to B.’s government: it is a critical concept for *La Repubblica* in the first period, and it almost disappears in the second one.

- as to the third part of the list, whereas *Libero* felt the situation “dramatic” before B.’s resignation, the dramaticity disappears afterwards. The same applies in smaller percentage to the other two newspapers. Another companion word, “collapse” has the same behaviour: Monti’s arrival is felt positively. However, the fear and the rumours of “failure” is highly felt by *Corriere della Sera* and *La Repubblica*, less so by *Libero*. This is confirmed by the abrupt disappearance of the concept of “dismiss/al” which dips to the lowest with *Libero*.

Eventually, in order to better compare specialized keywords we carefully chose and reclassified a small subset of all lemmata – 100 concepts – using a subset of labels that were suggested by Linguistic Inquiry and Word Count (LIWC)[9], a text analysis software program designed that we used in a previous experiment on Romanian [7].

The result of this new classification are highlighted here below, where we list for each newspaper the best performance in term of number of occurrences, for the first 16 classes in a given time interval: the same conclusion can be now reached by noting that *Libero* has opposite attitudes to *Repubblica*, and the latter has opposite attitudes to *Corriere*.

Newsp/t.perds	<i>OMBB</i>	<i>PTMB</i>	<i>OMAB</i>
LIBERO	1. sadness 2. results	1. sadness 2. results	3. rational 4. intuition
CORRIERE	4. intuition 5. negative 10. uncertain 11. failure 12. work 13. social 14. politics 15. positive 16. emotive	3. rational 4. intuition 5. anxiety 8. financial 10. uncertain 12. work 13. social 14. politics	6. anger 11. failure 13. social 14. politics 15. positive 16. emotive
REPUBBLICA	5. anxiety 6. anger 7. inhibition 8. financial 9. negative	6. anger 7. inhibition 11. failure 15. positive 16. emotive	1. sadness 2. results 5. anxiety 7. inhibition 8. financial 9. negative 10. uncertain 12. work

Table 3. Selected pragmatic features for three newspapers in 3 time periods

4 Conclusion

The analysis of the case study we proposed in this paper aims at testing if a linguistic perspective anchored in natural language processing techniques (in this case, the scaled version of GETARUNS system) could be of some use in evaluating political discourse in print press. If this proves to be feasible, then a linguistic approach would become a very relevant to an applicative perspective, with important effects in the optimization of the automatic analysis of political discourse.

However, we are aware that this study only sketches a way to go, and a lot more should be studied until a reliable discourse interpreting technology will become a tool in researcher's hands. We should also be aware of the dangers of false interpretation. For instance, if we take as example the three newspapers we used in our experiments, differences at the level of lexicon and syntax, which we have highlighted as differentiating them, should be attributed only partially to their idiosyncratic rhetorical styles, because these differences could also have editorial roots. According to common opinion, at least, *Corriere della Sera*, should embody an impartial opinion, *Libero*, pro Berlusconi and *La Repubblica*, against him. But differences are more subtle, and in fact, in some cases, we could likewise classify *Libero* as being impartial, *Corriere della Sera* as being pro current government and *La Repubblica* as the only one being more critical on the current government disregarding its political stance. It remains yet to be decided the impact that the use of certain syntactic structures could have over a wider audience of political discourse. In other words, this study may show that automatic linguistic processing is able to detect tendencies in the manipulation of the interlocutor with the hidden role of detouring the attention of the audience from the actual communicated content in favor of the speaker's intentions.

Different intensities of emotional levels have been clearly highlighted, but we intend to organize a much more fine-grained scale of emotional expressions. It is a well-known fact that the audience can be easily manipulated (e.g., the social and economic class) by a social actor (journalist, political actor) when their themes are treated with excessive emotional tonalities (in our study, common negative words). In the future, we intend to extend the specialized lexicon for political discourse in order to individuate more specific uses of words in context, of those words which are ambiguous between different semantic classes, or between classes in the lexicon and outside the lexicon (in which case they would not have to be counted). We believe that GETARUNS has a range of features that make it attractive as a tool to assist any kind of communication campaign. We wish it to be rapidly adapted to new domains and to new languages (i.e. Romanian), and be endowed with a user-friendly web interface that offers a wide range of functionalities. The system helps to outline distinctive features which bring a new and, sometimes, unexpected vision upon the discursive feature of journalists' writing.

Acknowledgments: In performing this research, the second author was supported by the POSDRU/89/1.5/S/63663 grant.

References

1. Bos, Johan & Delmonte, Rodolfo (eds.): "Semantics in Text Processing (STEP), Research in Computational Semantics", Vol.1, College Publications, London (2008).
2. Esuli, A. and F. Sebastiani. Sentiwordnet: a publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation LREC, 6, 2006.
3. Delmonte, R. (2007). Computational Linguistic Text Processing – Logical Form, Logical Form, Semantic Interpretation, Discourse Relations and Question Answering, Nova Science Publishers, New York.
4. Delmonte, R., Tonelli, S., Tripodi, R.: Semantic Processing for Text Entailment with VENSES, published at <http://www.nist.gov/tac/publications/2009/papers.html> in TAC 2009 Proceedings Papers (2010).
5. Delmonte, R. (2009). Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution, Nova Science Publishers, New York.
6. Delmonte R. and Vincenzo Pallotta, 2011. Opinion Mining and Sentiment Analysis Need Text Understanding, in "Advances in Distributed Agent-based Retrieval Tools", "Advances in Intelligent and Soft Computing", Springer, 81-96.
7. Gîfu, D. and Cristea, D.: Multi-dimensional analysis of political language, in J. J. (Jong Hyuk) Park, V. Leung, T. Shon, Cho-Li Wang (eds.) In Proc. of 7th FTRA International Conference on Future Information Technology, Application, and Service – FutureTech-2012, Vancouver, vol. 1, Springer (2012).
8. Hobbs, J. R., Stickel, M., Appelt, D., and Martin, P.: "Interpretation as Abduction", SRI International Artificial Intelligence Centre Technical Note 499 (1990).
9. Pennebaker, James W., Booth, Roger J., Francis, Martha E.: "Linguistic Inquiry and Word Count" (LIWC), at <http://www.liwc.net/>.
10. Kim, S.-M. and E. Hovy. Determining the sentiment of opinions. In Proceedings of the 20th international conference on computational linguistics (COLING 2004), page 1367–1373, August 2004.

11. Pang, B. and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL), page 271–278, 2004.
12. Polanyi, Livia and Zaenen, Annie: “Contextual valence shifters”. In Janyce Wiebe, editor, Computing Attitude and Affect in Text: Theory and Applications. Springer, Dordrecht, 1–10 (2006).
13. Pollack, M., Pereira, F.: “Incremental interpretation”. In Artificial Intelligence 50, 37-82 (1991).
14. Sauri R., Pustejovsky, J.: “Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text”, Computational Linguistics, 38, 2, 261-299 (2012).
15. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M.: “Lexicon-based methods for sentiment analysis”. In Computational Linguistics 37(2): 267-307 (2011).
16. Turney, P.D. and M.L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), pages 15–346, 2003.
17. Wiebe, Janyce, Wilson, Theresa, Cardie, Claire: “Annotating expressions of opinions and emotions in language”. In Language Resources and Evaluation, 39(2):165–210 (2005).
18. Reyes A., Rosso P., Buscaldi D. (2012). From Humor Recognition to Irony Detection: The Figurative Language of Social Media. In: Data & Knowledge Engineering, vol. 74, pp.1-12.
19. Reyes A., Rosso P. (2013). On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. In: Knowledge and Information Systems. DOI: <http://dx.doi.org/10.1007/s10115-013-0652-8>

Use of Language and Author Profiling: Identification of Gender and Age

Francisco Rangel^{1,2}, Paolo Rosso²

¹ Autoritas Consulting S.A., C/ Lorenzo Solano Tendero 7
28043 Madrid, Spain
francisco.rangel@autoritas.es
<http://www.kicorangel.com>

² Natural Language Engineering Lab, ELiRF,
Universitat Politècnica de València, Camino de Vera S/N
46022 Valencia, Spain
prosso@dsic.upv.es
<http://users.dsic.upv.es/~prosso>

Abstract. “In the beginning was the Word, and the Word was with God, and the Word was God”. Thus, John 1:1¹ begins his contribution to the Holy Bible (one of the most-distributed book in the world with hundreds of millions of copies²), the importance of the word lies in the essence of human beings. The discursive style reflects the profile of the author, who decides, often unconsciously, about how to choose and combine words. This provides valuable information about the personality of the author. In this paper we present our approach to identify age and gender of authors based on their use of language. We propose a representation based on stylistic features and obtain encouraging results with a SVM-based approach on the PAN-AP-13³ dataset.

1 Introduction

Knowing the profile of an author could be of key importance. For instance, from a forensic linguistics perspective being able to know what is the linguistic profile of a suspected text message (language used by a certain type of people) and identify characteristics (language as evidence) just by analyzing the text would certainly help considering suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products.

In previous work we carried out a statistical study of how the language is used in Spanish in different channels of Internet, concretely what grammatical categories

¹ <http://www.biblegateway.com/passage/?search=John+1&version=KJV>

² http://en.wikipedia.org/wiki/List_of_best-selling_books

³ Dataset for the Author Profiling task of the PAN 2013

<http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html>

people use in channels such as Wikipedia⁴, newsletters, blogs, forums, Twitter⁵ and Facebook⁶. In a recent work we investigated how the use of language could provide us enough evidences to identify the six basic emotions of Ekman⁷. We proposed a set of stylistic features and obtained competitive results in the identification of such emotions. We also carried out an exhaustive analysis of how the language varies by gender, topic and emotion, and all the possible combinations.

Based on the results of our previous works, in this paper we focus on the cognitive traits that make us different by gender and age. For that, we propose a set of features to represent texts written by anonymous authors, on the basis of stylistic features. With this set of features we aim to model the differences by age and gender in order to use them in a machine learning approach. We used a SVM method that we trained and tested with the PAN-AP-13 dataset. The obtained results are encouraging, although more in-depth features need to be investigated.

In Section 2 we present the state of the art, describing related work on author profiling. Furthermore we present the theoretical framework based on research on neurology. In Section 3 we describe our approach in detail together with the proposed features. In Section 4 we present the dataset used, the machine learning method and the evaluation measures. In Section 5 we discuss the experimental results. In Section 6 we draw some conclusions and discuss the future work.

2 Related work

Several are the interesting works on author profiling from the perspective of the common theoretical framework which involves several disciplines such as psychology, (computational) linguistics or even neurology.

2.1 Computational linguistics approaches

Several areas such as psychology, linguistics and, more recently, natural language processing are interested on studying how the use of the language varies according to the profile of the author. Pennebaker et al. (2003) connected the use of the language with traits such as gender, age, native language and so on. Argamon et al. (2003) used function words and the part-of-speech to predict gender of the authors of written texts from the British National Corpus, and Holmes and Meyerhoff (2003); Burger and Henderson (2011) have also investigated in obtaining age and gender from formal texts. Authors like Koppel et al. (2003); Schler et al. (2006); Goswami et al. (2009) used combinations of simple lexical and syntactic features to determine the gender and age of authors of anonymous blog posts. Peersman et al. (2011) retrieved a dataset from Netlog⁸, with self-annotated age and gender of their authors and

⁴ <http://dumps.wikimedia.org/eswiki/20121227/eswiki-20121227-pages-meta-current.xml.bz2>

⁵ <https://twitter.com/>

⁶ <https://www.facebook.com/>

⁷ Joy, surprise, sadness, disgust, anger, fear

⁸ <http://www.netlog.com/>

Goswami et al. (2009) demonstrated that the use of language in blogs correlates with age (for example, with the increase of the use of prepositions and determiners), but could not determine similar correlation with gender. Zhang and Zhang (2010) experimented with short segments of blog posts and Nguyen et al. (2013) studied the use of language and age in Twitter, the most well-known platform of short texts (140 characters long). All of them based their studies on gathering stylistic features like non-dictionary words as slang words, part-of-speech, function words, hyperlinks, the average length of the sentences, and sometimes combined with content features as single words with the highest information gain.

2.2 Author profiling tasks

The task of obtaining author profiles has an emerging interest in the scientific community, as can be seen in the number of related tasks around the topic. The task on *Author Profiling at PAN 2013*⁹ encouraged researchers to identify age and gender of the authors of a large amount of anonymous texts (Rangel et al., 2013). Participants had to infer from blog posts what age and gender the authors are, in a real scenario with a large-size corpus and high amount of spam data, for example, automatically generated by robots.

Similarly, the shared task on Native Language Identification at *BEA-8 Workshop*¹⁰ promotes researchers to identify native language of an author based on a sample of their writing. Finally, the task on *Personality Recognition at ICWSM 2013*¹¹ intends to be a common research framework where to investigate the Big Five traits¹².

In a similar vein, the interest in this type of research is evident in the Kaggle¹³ platform, where companies and research departments can share their needs and independent researchers can join the challenge of solving them. We can find challenges as *Psychopathy Prediction Based on Twitter Usage*¹⁴, *Personality Prediction Based on Twitter Stream*¹⁵ or *Gender Prediction from HandWriting*¹⁶. This shows the rise of interest on this kind of problems.

2.3 Neurology: A theoretical framework

Neurology is the science which focuses its interest on treating disorders on neural system, such as aphasia. Aphasia is the loss of ability to produce or understand language due to lesions in brain areas related to these functions. At the end of XIX century and as result of studies about that disease, the German neurologist and

⁹ <http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html>

¹⁰ <https://sites.google.com/site/nlisharedtask2013/>

¹¹ <http://mypersonality.org/wiki/doku.php?id=wcpr13>

¹² Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

¹³ <http://www.kaggle.com/>

¹⁴ <http://www.kaggle.com/c/twitter-psychopathy-prediction>

¹⁵ <http://www.kaggle.com/c/twitter-personality-prediction>

¹⁶ <http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting>

psychiatrist Karl Wernicke and the French physician, anatomist and anthropologist Paul Pierre Broca defined two brain areas involved in the comprehension and production of the language, respectively Wernicke's Area and Broca's Area. (Falk, D., 2004)

Wernicke's area is in the cerebral cortex in the posterior half of the superior temporal gyrus and in the adjacent part of the middle temporal convolution, and its main role is the auditory decoding in the linguistic function, related with the language comprehension and with the control of the content of the message.

Broca's area is in the third inferior frontal gyrus, in the frontal lobe of the left hemisphere of the brain, for the vast majority of people, the hemisphere which rules the language. It controls many of the social skills of people, processes the grammar, is involved in the production of the speech, in the processing of the language and in its comprehension. It controls the ability to express and conciliate emotions, the skill for reading facial emotion in other people, the emotions and the skill for establishing social relationships. This area seems to be responsible for processing style words.

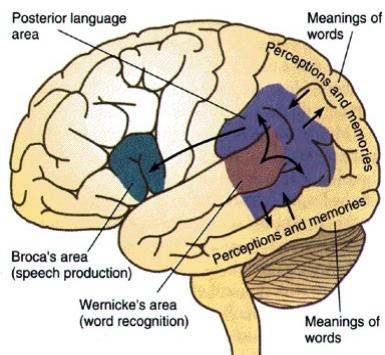


Fig. 1. Broca and Wernicke's areas of the brain

There are two basic questions to answer when we write a speech, WHAT to say and HOW to say it. The first question relates to the object that we want to communicate and defines the content of the speech. It is not a personal issue but a matter referred to what is being communicated. The second question responds to the way the author is going to communicate the content, for example, the style of the discourse itself. Therefore, it is a matter inherent to the communicator, to the way the communicator builds his discourse, and it is determined essentially by his profile. From the viewpoint of the receiver, the questions turn on WHAT is said and WHO is saying what, and both questions are related to two kind of words, those oriented to communicate contents, to answer the question WHAT, and those oriented to connect the discourse, to give its style and form, to answer the question HOW/WHO.

3 Automatic identification of gender and age based on stylistic features

We focused our interest on the cognitive approach based on the neurology studies of Broca and Wernicke and tried to represent the way the users express themselves, the way they use the language, that is, the style authors write. Based on the study of Pennebaker (2011) for English, where stylistic features identify some profile traits, we carried out some statistical research in Spanish analyzing a large number of documents¹⁷ from Wikipedia, newsletters, forums, blogs, Twitter and Facebook and obtaining the frequency of use of the different grammatical categories.

Table 1. Distribution of grammatical categories per channel

POS	WIKI	NEWS	BLOGS	FORUMS	TW	FB
ADJ	13.57	12.50	13.67	9.27	6.62	12.06
ADV	2.78	3.46	3.87	4.74	6.30	3.49
CONJ	1.52	2.10	1.80	4.18	7.00	2.64
Q	3.34	4.47	4.15	5.34	5.53	4.29
DET	2.88	3.48	2.78	4.18	6.40	4.02
INTJ	0.35	0.04	0.06	0.42	0.38	0.07
MD	0.01	0.03	0.02	0.00	0.00	0.00
PREP	4.00	5.49	5.07	8.94	13.81	6.15
PRON	0.65	0.92	1.12	2.22	3.32	1.39
NOM	50.33	47.05	46.59	42.63	34.08	47.07
VERB	20.55	20.47	20.88	18.08	16.56	18.83

Table 2. Frequency of person and number in pronouns and verbs

POS	PER	NUM	WIKI	NEWS	BLOG	FOR	TW	FB
PRON	1	SIN	13.61	14.58	18.85	54.47	65.81	22.3
		PLU	0.00	0.00	0.00	0.00	0.00	0.00
	2	SIN	4.58	1.18	2.23	1.54	3.53	3.95
		PLU	1.92	1.75	5.31	4.61	5.62	3.49
	3	SIN	55.06	50.75	39.26	24.08	12.70	34.68
		PLU	13.42	18.22	16.93	8.91	3.35	17.14
OTHER			11.41	13.52	17.42	6.39	8.99	18.44
VERB	1	SIN	19.95	17.41	17.50	28.94	24.00	16.61
		PLU	2.10	2.42	4.19	2.68	4.68	4.89
	2	SIN	6.02	1.55	3.58	3.55	6.77	2.95
		PLU	0.46	0.42	0.69	0.98	1.65	0.76
	3	SIN	31.40	34.00	29.92	28.80	31.21	31.21
		PLU	40.07	44.20	45.11	35.05	31.69	43.59

¹⁷ Number of documents per channel: Wikipedia: 3,987,179 Newsletters: 5,191,694 Blogs: 1,083,709 Forums: 673,664 Twitter: 23,873,371 Facebook: 576,723

Table 1 shows the similitude between Wikipedia, newsletters and blogs in the use of adjectives, nouns and verbs to describe objects, people, places and situations. Forums is highlighted for the high use of prepositions, adverbs and pronouns due to the need of authors for directly describing their problems and searching for a solution. In Twitter, people use pronouns and verbs in first person (Table 2) with the highest frequency. This confirms such channel as ego-centered, where authors try to communicate personal thoughts, together with a low use of verbs and high use of adverbs and prepositions, following Twitter's main motto: “what are you thinking about?”, “what are you doing?” or “what is happening?”.

Following, we employed these findings on the use of grammatical categories in order to identify emotional profile in texts. Texts were classified into the six basic emotions (joy, surprise, anger, disgust, fear, sadness). We analyzed the distribution of the use of grammatical categories by gender in a dataset of 1,200 Facebook comments. Results are shown in Table 3.

Table 3. Distribution of grammatical categories by gender

POS	ALL	MALE	FEMALE
ADJ	6.49	6.53	6.45
ADV	3.93	3.94	3.91
CONJ	9.51	9.55	9.46
Q	5.46	5.76	5.12
DET	7.25	6.81	7.74
INTJ	0.23	0.18	0.30
MD	0.00	0.00	0.00
PREP	6.06	6.25	5.85
PRON	2.45	2.24	2.67
NOM	31.89	32.21	31.53
VERB	15.38	15.44	15.32

We can appreciate some important variations in the use of the grammatical categories by gender, for instance, as found for English (Pennebaker, 2011), we also verified for Spanish that men use more prepositions than women (+6.84%), perhaps because they try to hierarchically categorize things into their environment, and women use more pronouns (+19.20%), determinants (+13.66%) and interjections (+66.67%) than men perhaps because they are more interested in social relationships. Such conclusions appear to be parallel with content and style, with Wernicke and Broca's areas and we thought that using such stylistic features could help us to determine gender with some accuracy. We also have the intuition that such features could help us to identify age. Thus, we proposed the following features:

- Frequencies: Ratio between number of unique words and total number of words, words starting with capital letter, words completely in capital letters, length of the words, number of capital letters and number of words with flooded characters (e.g. Heeeeelloooo);
- Punctuation marks: Frequency of use of dots, commas, colon, semicolon, exclamations, question marks and quotes;

- Part-of-speech: Frequency of use of each grammatical category, number and person of verbs and pronouns, mode of verb, proper nouns (NER) and non-dictionary words (words not found in dictionary);
- Emoticons¹⁸: Ratio between the number of emoticons and the total number of words, number of the different types of emoticons representing emotions: joy, sadness, disgust, angry, surprised, derision and dumb;
- Spanish Emotion Lexicon (SEL) (Sidorov et. al, 2012) : We obtained the lemma for each word and then its *Probability Factor of Affective Use* value from the SEL dictionary. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion, building one feature per emotion.

We do not use any content/context dependent features in order to obtain more independence from the topics.

4 Methodology

4.1 Training and test datasets

The PAN-AP-13 dataset consists of a large number of anonymous authors labeled with gender and age. For the age group, on the basis of previous work (Koppel et al., 2003) the following classes are considered: 10s (13-17), 20s (23-27) and 30s (33-47). The data is balanced by gender but not by age. Each author can contain from one to tens of posts. The distribution of the number of authors per dataset is shown in Table 4.

Table 4. Distribution of number of authors by age

AGE	NUM. OF AUTHORS	
	TRAIN	TEST
10s	2,500	240
20s	42,600	3,840
30s	30,800	2,720

4.2 Machine learning approach and performance measures

We used the Support Vector Machine method implemented in Weka¹⁹. We experimented with different parameters and finally we used a Gaussian kernel with $g=0.01$ and $c=2,000$.

In order to be able to compare our results with the ones obtained by the teams participating in the PAN 2013 task on Author Profiling, we used Accuracy as

¹⁸ http://es.wikipedia.org/wiki/Anexo:Lista_de_Emoticonos

¹⁹ <http://www.cs.waikato.ac.nz/ml/weka/>

“closeness of agreement between a measured quantity value and a true quantity value of a measurand”. Concretely, we perform the ratio between the number of authors correctly predicted by the total number of authors.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

5 Discussion of the experimental results

In Table 5 our proposal is ranked together with the final results of PAN-AP task²⁰, separately for age and gender.

Table 5. PAN ranking for Author Profiling by Gender and by Age (Spanish)

POS	TEAM	GENDER		POS	TEAM	AGE
1	Santosh	0.6473		1	Pastor	0.6558
2	Pastor	0.6299		2	Santosh	0.6430
3	Haro	0.6165		3	(Rangel)	0.6350
4	Ladra	0.6138		4	Haro	0.6219
5	Flekova	0.6103		5	Flekova	0.5966
6	Jankowska	0.5846		6	Ladra	0.5727
7	(Rangel)	0.5713		7	Yong	0.5705
8	Kern	0.5706		8	Ramirez	0.5651
9	Jimenez	0.5627		9	Aditya	0.5643
10	Ayala	0.5526		10	Jimenez	0.5429
11	Cagnina	0.5516		11	Gillam	0.5377
12	Yong	0.5468		12	Kern	0.5375
13	Mechti	0.5455		13	Moreau	0.5049
14	Weren	0.5362		14	Meina	0.4930
15	Meina	0.5287		15	Weren	0.4615
16	Ramirez	0.5116		16	Jankowska	0.4276
17	Baseline	0.5000		17	Cagnina	0.4148
18	Aditya	0.5000		18	Hidalgo	0.4000
19	Hidalgo	0.5000		19	Farias	0.3554
20	Farias	0.4982		20	Baseline	0.3333
21	Moreau	0.4967		21	Ayala	0.2915
22	Gillam	0.4784		22	Mechti	0.0512

We achieved the 7th position in gender prediction and 3rd position in age prediction. Based on such results, we conclude that the proposed stylistic features perform better for age than for gender identification. Perhaps this is due to the fact that the writing style depends more on the age of the author than on the gender, confirming what

²⁰ <http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/pan13-ap-final-results.pdf>

stated in (Goswami et al., 2009) about the correlation between age and use of language. But in any case, the task of identifying age seems to be easier than identifying gender, task which seems to be very difficult because the values obtained are not so high compared to the baseline (50%).

6 Conclusions and future work

We focused our interest on the cognitive approach based on the neurology studies of Broca and Wernicke and tried to represent the way the users express themselves, the way they use the language, that is, the style in which authors write. We carried out some experiments and some important variations in the use of the grammatical categories by gender were appreciated. For example, men use more prepositions than women because they try to hierarchically categorize things into their environment, and women use more pronouns, determinants and interjections than men because they are more interested in social relationships.

We conclude that stylistic features help to identify age and gender of anonymous authors although the task seems to be very difficult mainly for gender detection. We obtained competitive results in comparison with the ones obtained by the PAN-AP task participants. This encourages us to follow the research in this direction in order to understand better how people use language to express themselves and how this could help us to identify the profile of an author.

We must bear in mind with the differences between languages, for example between English and Spanish. For instance, in Spanish the use of pronouns is generally elliptical and it is a choice of the author to use them perhaps to emphasize something, as well as the use of prepositions or determinants in English is more regulated than in Spanish. Due to such specificities, we plan to investigate our proposal to different languages as English.

The features we use for modeling the discursive style are preliminary and simples. As future work we are interested in analyzing the discourse in order to investigate further how people use different words of the different grammatical categories, how they place them in the sentence, and how such stylistic decisions provide us information about the author profile. We also plan to research on the relationship between the demographics such as the gender and age with the emotional profile of the authors and their personality traits, trying to link such tasks in order to build a common framework to allow us to better understand how people use language from a cognitive linguistics viewpoint.

Acknowledgements

The work of the first author was partially funded by Autoritas Consulting SA and by Ministerio de Economía de España under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the second author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS – Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

- Argamon, S., Koppel, M., Fine, J., Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3): 321–346.
- Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. (2011). Discriminating gender on Twitter. In EMNLP'11: 1301-1309.
- Falk, D. (2004). Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and Brain Sciences* 27, 491-541.
- Goswami, S.; Sarkar, S.; and Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In ICWSM'09.
- Holmes, J., and Meyerhoff, M. (2003). *The handbook of language and gender*. Oxford: Blackwell.
- Koppel, M., Argamon, S., Shimoni, A. R. (2003). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17: 401-412.
- Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T. (2013). “How Old Do You Think I Am?”: A Study of Language and Age in Twitter. The 7th International AAAI Conference on Weblogs and Social Media. ICWSM'13.
- Peersman, C., Daelemans, W., Van Vaerenbergh, L. (2011). Predicting Age and Gender in Online Social Networks. SMUC'11.
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*. (54): 547–577.
- Pennebaker, J.W. (2011) *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press.
- Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. (2013) Overview of the Author Profiling Task at PAN 2013. In: Forner P., Navigli R., Tufis D.(Eds.), *Notebook Papers of CLEF 2013 LABs and Workshops*, CLEF-2013, Valencia, Spain, September 23-26.
- Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2006) Effects of Age and Gender on Blogging. *American Association for Artificial Intelligence*.
- Sidorov, G., Miranda Jiménez, S., Viveros Jiménez, F., Gelbukh, A., Castro Sánchez, N., Velásquez, F., Díaz Rangel, I., Suárez Guerra, S., Treviño, A., Gordon, J. (2012) Empirical Study of Opinion Mining in Spanish Tweets. *LNAI* 7629-7630.
- Zhang, C., and Zhang, P. (2010). Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA.