QITL-5

Proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics

University of Leuven,
12-14 September 2013

KU LEUVEN

In co-operation with
Ghent University

Proceedings of Quantitative Investigations in Theoretical Linguistics 5

12–14 September 2013
University of Leuven

Edited by Thomas Wielfaert, Kris Heylen and Dirk Speelman
Cover design by Costanza Asnaghi

Proceedings of Quantitative Investigations in Theoretical Linguistics 5
12–14 September 2013
University of Leuven
http://wwwling.arts.kuleuven.be/QITL5/

Invited Speakers
Jennifer Hay, University of Canterbury, New Zealand
Laura Janda, University of Tromsø, Norway
Søren Wichmann, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Programme Committee
Antti Arppe, University of Alberta
Marco Baroni, University of Trento/CIMeC
Gerlof Bouma, University of Gothenburg
Douglas Biber, Northern Arizona University
Joan Bresnan, Stanford University
Michael Cysouw, Philipps University Marburg
Walter Daelemans, University of Antwerp
Peter de Swart, University of Nijmegen
Dagmar Divjak, University of Sheffield
Cédrick Fairon, Université catholique de Louvain (UCL)
Gaëtanelle Gilquin, Université catholique de Louvain (UCL)
Dylan Glynn, Lund University
Stefan Th. Gries, University of California, Santa Barbara
Stefan Grondelaers, University of Nijmegen
Martin Hilpert, Université de Neuchâtel
Daniel Ezra Johnson, Lancaster University
Merja Kytö, University of Uppsala
Anke Lüdeling, Humboldt University in Berlin
Thomas Mayer, Philipps University Marburg
Tom Ruette, Humboldt University in Berlin
Clara Vanderschueren, Ghent University
Ruprecht von Waldenfels, University of Bern
Tom Wasow, Stanford University
Amir Zeldes, Humboldt University of Berlin

Organizing Committee
Dirk Speelman, University of Leuven
Dirk Geeraerts, University of Leuven
Kris Heylen, University of Leuven
Gert De Sutter, Ghent University
Timothy Colleman, Ghent University

Sponsors
Research Foundation Flanders (FWO)
Faculty of Arts (Faculteit Letteren), University of Leuven

Preface

Ever since the first Quantitative Investigations in Theoretical Linguistics conference in 2002 in Osnabrück, QITL conferences have taken up a special position in the landscape of linguistics conferences. Not just a special position, we believe, but a position that is to be cherished. Not only is it one of the rare forums where researchers from all subdisciplines of linguistics interested in quantitative linguistic methodology can meet and share insights, it also is a place where proponents of quantitative linguistics and empirical methodology from various theoretical linguistic backgrounds compare and discuss their findings. Most of all, QITL rightfully advocates an approach to quantitative research in which the relation between quantitative methods and theoretical insights is made very explicit.

In other words, QITL is not just about quantitative approaches, but is also and most importantly about the use of quantitative methods in support of theory building and in support of the falsification of theory-driven hypotheses. We believe this focus on linguistic methodology and on how it can inform linguistic theory is becoming ever more important now that we witness a turn towards empirical methodology throughout linguistics.

However, the relevance of its topic is not the only asset of QITL, there is also the format. With its single track sessions, its ample time for discussion and its relatively moderate size, QITL has often created ideal circumstances for fruitful (and enjoyable) discussion.

For all these reasons we have tried to adhere to the QITL tradition as much as possible, and we feel honoured and excited to have the opportunity to contribute to this forum and its continuation. We want to thank all authors and participants for their interesting contributions, we want to thank all members of the programme committee for their time and energy and for doing a terrific job and finally we also want to thank all members of the QLVL team who helped out with the practical organisation of the conference.

Leuven, September 2013,

Dirk Speelman, Dirk Geeraerts, Kris Heylen, Gert De Sutter and Timothy Colleman

Contents

Skewed word distributions affect speech production and perception

Jennifer Hay
University of Canterbury, New Zealand
jen.hay@canterbury.ac.nz

Experienced-based models of speech production and perception often privilege the word as a unit of representation. Linguistic evidence presented in favour of detailed word-level representations usually involves word-frequency effects. Such effects are compatible with the idea of phonetically detailed word-level representations. But many frequency effects are also compatible with more abstract models, in which production or perception contain an overall bias in a certain direction, applying across the board in more frequent or more predictable contexts.

This paper attempts to explore the predictions of experience-based representations at the word level in more detail. In particular, it explores the characteristics of words which are not evenly distributed across particular linguistic or social habitats. Some words occur more sentence medially, for example. And some occur more sentence finally. Some words are used more by older speakers, and some by younger speakers. I argue that the uneven linguistic and social distributions of words affects their representations. The consequences of these uneven statistical distributions can be seen in production, perception and in the trajectory of sound change.

The Big Questions
Need Multipurpose Portable Solutions

Laura A. Janda
University of Tromsø
laura.janda@uit.no

I offer "linguistic profiles" as a suite of methodological ideas bridging the gap between key theoretical issues in linguistics and quantitative models. Collectively linguistic profiles make it possible to operationalize theoretical questions about the structure of languages so that data can be collected and analyzed. As linguists we should strive to create investigative resources that are portable across languages and have multipurpose applications for language pedagogy and support of endangered languages in addition to linguistic research.

The Big Questions I focus on are:

1) What is the relationship between form and meaning?

2) What is the relationship between lexicon and grammar?

3) What is the structure of linguistic categories?

4) What is the structure of linguistic constructions?

All of these issues are controversial in linguistic theory. While some linguists separate form from meaning, others insist that there is no form without meaning, which means that there are no semantically empty forms, and that difference in form necessarily reflects difference in meaning, with the entailment that there are no true synonyms. A distinction between lexicon and grammar is assumed in theories that assign various phenomena to one or the other, however other theories view lexicon and grammar as parts of a single continuum lacking a clear boundary. Crucially, it is asserted that meaning is not the exclusive privilege of the lexicon, but that grammatical categories such as case, aspect, person, etc. have meaning as well. Although it has been presumed since Aristotle that linguistic categories are discretely bounded, many linguists now believe that categories may be fuzzy and overlapping, structured around prototypes. Grammatical constructions can be modeled as hierarchical structures, often diagrammed as trees, but there is growing evidence that grammatical structure is flat, relying on locally-available sequential cues (Frank et al. 2012).

These Big Questions are not in themselves quantifiable. Linguistic profiles make it possible to approach these questions empirically and from a variety of angles. These include:

1) Grammatical profiling -- examining the relationship between the frequency distribution of grammatical forms and grammatical and lexical categories (Janda and Lyashevskaya 2011, Eckhoff and Janda forthcoming);

2) Constructional profiling -- examining the relationship between the frequency distribution of grammatical constructions and meaning (Sokolova, Janda and Lyashevskaya 2012);

3) Collostructional profiling -- examining the relationship between a construction and the words that most frequently fill its slots (Kuznetsova 2013);

4) Semantic profiling -- examining the relationship between meanings (measured by independently assigned semantic tags) and forms (morphemes, words; Janda and Lyashevskaya forthcoming);

5) Radial category profiling -- examining differences in the frequency distribution of uses across two or more near-synonyms (Nesset et al. 2011, Endresen et al. 2012).

Linguistic profiles aim at the Big Questions, but are themselves agnostic about both the theory involved and the statistical methods used. Profiles are a way of organizing measures that can be evaluated in many ways, including: chi-square, Fisher test, hierarchical clustering, componential analysis, regression, conditional inference trees and random forests, and naive discriminative learning.

All linguistic profiling methods take the form-meaning relationship as their point of departure. We should create open-source resources for languages that will make it possible to extract the data needed for linguistic profiling. These resources will include disambiguators and parsers and can be modeled after the Giellatekno language technology resources at the University of Tromsø (URL). In addition to facilitating linguistic research, these resources can serve multiple purposes in the building of tools for language pedagogy, (real, not statistical) machine translation, and documentation and revitalization for minority indigenous languages.

References

Eckhoff, H. M. & Janda, L. A. Forthcoming. "Grammatical Profiles and Aspect in Old Church Slavonic". *Transactions of the Philological Society*.

Endresen, A., Janda, L. A., Kuznetsova, J., Lyashevskaya, O., Makarova, A., Nesset, T. & Sokolova, S. 2012. "Russian 'purely aspectual' prefixes: Not so 'empty' after all?", co-authored with *Scando-Slavica* 58:2, 231-291.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B, 279,* 4522–4531. doi: 10.1098/rspb.2012.1741

Janda, L. A. & Lyashevskaya, O. 2011. "Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian". *Cognitive Linguistics* 22:4 (2011), 719-763.

Janda, L. A. & Lyashevskaya, O. Forthcoming. "Semantic Profiles of Five Russian Prefixes: *po-*, *s-*, *za-*, *na-*, *pro-*", co-authored with Olga Lyashevskaya. *Journal of Slavic Linguistics*.

Kuznetsova, J. 2013. *Linguistic Profiles. Correlations between Form and Meaning*. PhD dissertation, University of Tromsø.

Nesset, T., Endresen, A. & Janda, L. A. 2011. "Two ways to get out: Radial Category Profiling and the Russian Prefixes *vy-* and *iz-*". *Zeitschrift für Slawistik* 56:4, 377-402.

Sokolova, S., Lyashevskaya, O. & Janda, L. A. 2012. "The Locative Alternation and the Russian 'empty' prefixes: A case study of the verb *gruzit'* 'load'", co-authored with Svetlana Sokolova[1] and Olga Lyashevskaya[2]. In: D. Divjak & St. Th Gries (eds.). *Frequency effects in language representation* (Trends in Linguistics. Studies and Monographs. 244.2), 51-86. Berlin: Mouton de Gruyter.

The automated classification of the world's languages: can it go deeper?

Søren Wichmann
Max Planck Institute for Evolutionary Anthropology
wichmann@eva.mpg.be

Within the Automated Similarity Jugdment Program or ASJP project, 40-item word lists from more than 6,000 languages and dialects have been compiled for the purpose, inter alia, of arriving at a consistent and objective classification of the world's languages. This has mostly been carried out using a version of the Levenshtein distance. Inspection of a single tree of the world's languages, as well as several studies of individual families, have rendered the strengths and weaknesses of the method evident. Notably, the method is very reliable in distinguishing and correctly clustering families down to a time depth of around 4,200 BP, but beyond that it gets increasingly less reliable. The high reliability at the level of relatively shallow families justifies using a pre-established conservative classification of the world's language families-produced by Harald Hammarström-for clustering the languages prior to a classification through ASJP using average similarities across families. This reduces the noise from individual languages for which wordlists are incomplete, as well as the effects of accidental similarities between single pairs of languages. At the same time, pairwise comparisons of all languages families in the world, involving a total of 57,630 pairs, offers the possibility of empirically estimating whether or not the similarity found for a given pair is or isn't due to chance. Other probabilistic approaches to historical linguistic language comparison have had to rely on less reliable theoretical probabilities of matches between sound segments.

One of the results of this approach is a remarkable grouping of 25 out of 27 Australian families in the database into a single cluster, supporting a common view, but one that has never been supported by strong lexical evidence, that there is an Australian super-family comprising all or nearly all of the languages on this continent. Holman et al. (2011) estimated Australian to be 5,296 years old, so it appears that ASJP can still provide useful results even at a time depth around a millennium greater than the time depth at which it routinely performs reliably. It remains, however, to investigate whether the results for Australian are more likely to be due to lexical diffusion rather than inheritance (genealogical relatedness).

This paper will give an overview of the nuts and bolts of ASJP and different studies of its performance, and will introduce the new results for pairwise comparisons of language families. Statistical considerations of how to validate results for deep genealogical relations will be offered. Finally, the case of Australian is presented in some detail.

A quantitative investigation of semantic properties of determiners using factorization techniques

Marta Abrusan Tim Van de Cruys
IRIT & CNRS, Toulouse, France
{marta.abrusan,tim.vandecruys}@irit.fr

In this research, we explore the use of a number of well-established statistical methods for the investigation of semantic properties of quantificational determiners, with a particular emphasis on its interaction with the mass-count distinction. We combine these results with a statistical analysis of aspectual distinctions and *actionsart*.

Specifically, we explore the use of a distributional word space model – based on a massive text corpus of about four billion words – in order to capture semantic generalizations about natural language determiners. From our corpus, we extract co-occurrences of the most common quantificational determiners with the most frequent nouns, and we analyze the resulting co-occurrence frequencies using a number of well-known factorization techniques.

First, we look at singular value decomposition, a well-known method from linear algebra. This method has been used in order to automatically capture latent semantic properties of words from their simple co-occurrence frequencies, in techniques such as latent semantic analysis (Landauer and Dumais, 1997). Our analysis shows a number of interesting generalizations – known from formal semantic analysis – such as the mass-count distinction.

Next, we explore the use of tensor factorization methods (Kolda and Bader, 2009). Previous research has shown that tensor algebra is a suitable tool for the modeling of language phenomena, and our research investigates its usefulness for the modeling of determiners. Up till now, most research in distributional semantics uses a simple matrix model as its basic mathematical object, which is well suited for modeling the semantics of individual words (representing these words by their different contexts). However, if one wants to model the interaction between multiple words (in this case the interactions between determiners and nouns), we must take into account multiple co-occurrences. Multiple co-occurrences can be modeled in the form of tensors, which are the generalization of a matrix to more than two dimensions. Using tensor algebra, we explore the correlations of determiners and nouns, together with the words in the surrounding context of the determiner-noun phrase. In doing so, our model aims to capture more advanced latent semantic characteristics, akin to generalizations known from formal semantic theories, in a fully automatic way.

With the above methods, we not only capture interesting generalizations about determiners, but we also construct, as a side effect, an ordering of the most frequent nouns based on their score on e.g. the mass-count scale. This information, in turn, can be used in the statistical analysis of various other semantic phenomena, such as the aspectual properties of verbal complexes. This is because, as is well known, the aspectual and *actionsart* properties of verbal complexes depend to a large degree on the nature of the object: in particular whether or not the object is count or mass, or more generally, quantized or cumulative. We combine our results with previous studies in distributional semantics on aspect.

So far distributional approaches and related factorization methods have focused on the induction of the semantics of lexical (content) words. Our research shows that the aforementioned factorization techniques are a fruitful approach for the investigation of fine-grained semantic properties of function words as well as related formal properties of lexical elements.

References

Tamara G. Kolda and Brett W. Bader. 2009. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychology Review*, 104:211–240.

Global Autocorrelation and Dialect Studies: The Role of Significance

Costanza Asnaghi
Università Cattolica del Sacro Cuore; KU Leuven
costanza.asnaghi@unicatt.it

This study debates whether it is sensible to include or better to exclude variables that do not exhibit statistically significant patterns in dialect studies. The argumentation is based on the results of a previous regional lexical variation survey of California English, which examined 45 continuous lexical alternation variables in 334 online newspapers across 273 California locations (Asnaghi, 2013). The frequencies of the 45 lexical alternation variables were gathered through site-restricted web searches (Grieve et al, submitted).

The investigation begins with the explanation of the choice for the spatial weighting function based on the assumption that language in neighboring locations is likely to be more similar than language in locations which are far apart: with a reciprocal weighting function, the results of this research are highlighted based on the distances between cities. Next, an overview of global and local methods for spatial autocorrelation, comparing pairs of values in the spatial distribution of each variable, is provided: in particular, Moran's *I* (Moran, 1948) is used to analyze global spatial autocorrelation, and Getis-Ord *Gi* (Ord and Getis, 1995; Grieve, 2011) is used to analyze local spatial autocorrelation.

The list of 45 variables is ranked according to their significance score (*p*-value). Thirty out of 45 variables are found to display significant patterns of autocorrelation on the global level.

Factor analysis is then calculated for the autocorrelated values of the variables from the California survey in two different ways: the first analysis is conducted on the comprehensive set of 45 variables, and the second analysis is conducted on the 30 significant variables only.

The comparison of the representations resulting from the two factor analyses shows that the maps from the two sets of variables reasonably align, with a preference for those incorporating the complete set of variables. It could be claimed that this preference is based on arbitrary judgement. In fact, it is not necessarily true that a more cohesive pattern, as detected in factor 1 and factor 3 for the complete set of variables (Figure 1 and 5) rather than in factor 3 and factor 1 for the significant variables only (Figure 2 and 6), is sound proof of a more accurate representation of language usage. Nonetheless, the choice for the all-variables representation for factor 2 (Figure 3) is based on socio-economic grounds. Introducing an external element, namely the socio-economic one, in the criteria for the selection of a model can base the choice on a more realistic rationale.

Apart from the comparison between pattern representations, there are other reasons for retaining all variables in a dialect research: even if a variable does not return a significant Moran's *I*, it can still show clear patterns of local spatial autocorrelation. As a general rule, the stronger the Moran's *I* score is, the stronger the local autocorrelation clusters are; nonetheless, the elimination of lower Moran's *I* variables leads to an exclusion of important patterns that the local autocorrelation analysis identified. In fact, not significant Moran's *I* variable patterns side with other significant variables, strengthening important information.

References

Asnaghi, Costanza. 2013. *An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches*. Ph.D. thesis, Universtà Cattolica del Sacro Cuore

and KU Leuven.

Grieve, Jack. 2011. A Regional Analysis of Contraction Rate in Written Standard American English. *International Journal of Corpus Linguistics* 16 (4): 514–546.

Grieve, Jack, Costanza Asnaghi, and Tom Ruette. Submitted. Site-Restricted Web Searches for Data Collection in Regional Dialectology. *American Speech*.

Moran, Patrick. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2): 243–251.

Ord, J. Keith and Arthur Getis (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27 (4): 286–306.

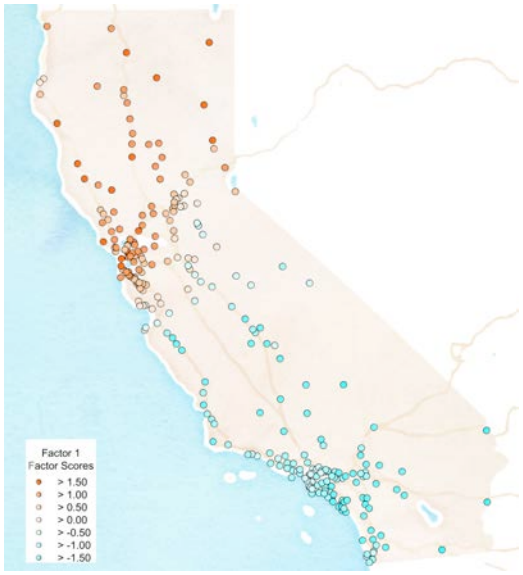

Figure 1: Map of Factor 1 from Complete Set of Variables (20% Variance Explained).

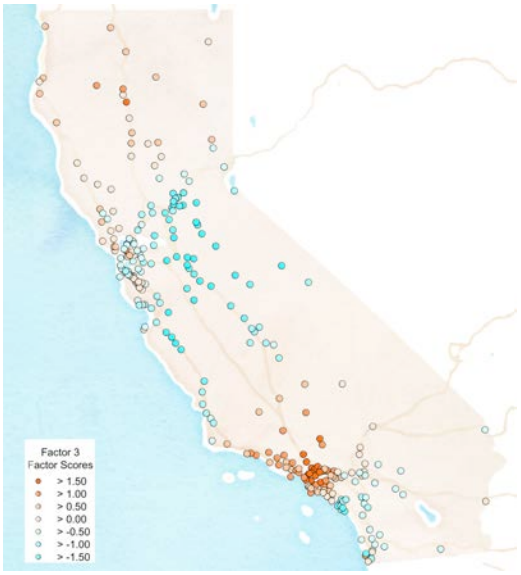

Figure 2: Map of Factor 3 from Significant Variables Only (17% Variance Explained).

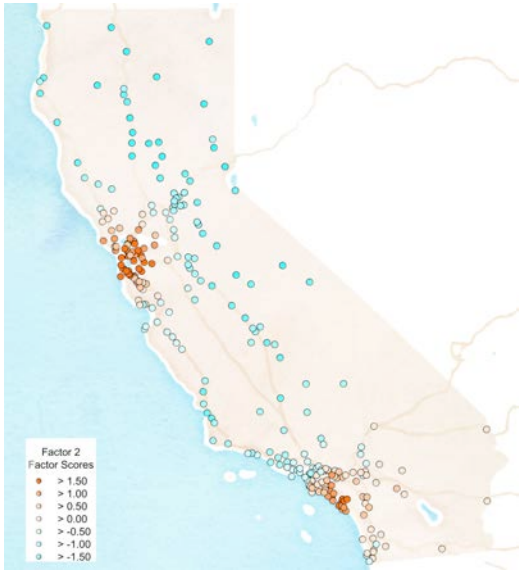

Figure 3: Map of Factor 2 from Complete Set of Variables (18% Variance Explained).

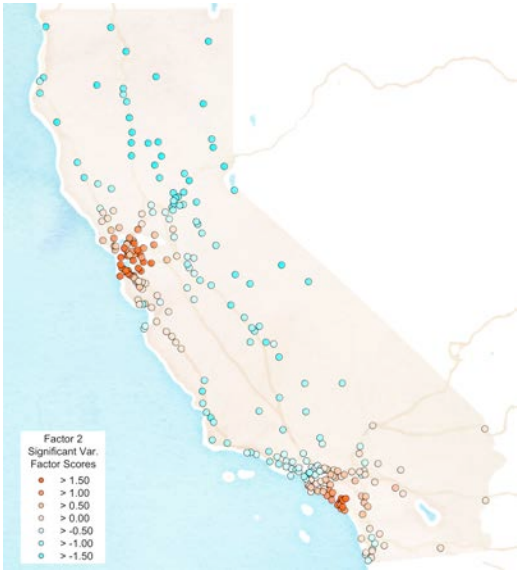

Figure 4: Map of Factor 2 for Significant Variables Only (20% Variance Explained).

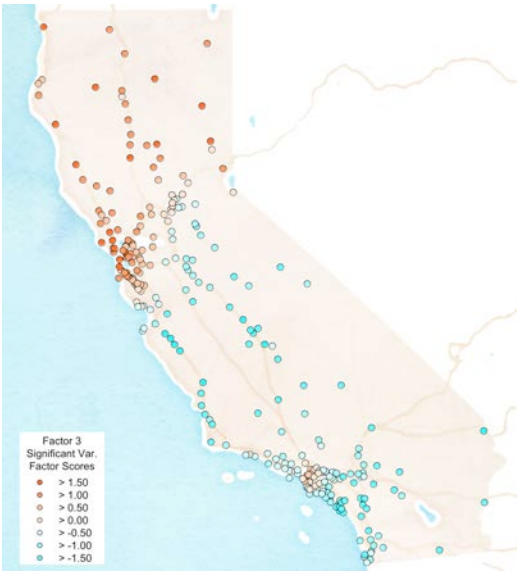

Figure 5: Map of Factor 3 from Complete Set of Variables (12% Variance Explained).

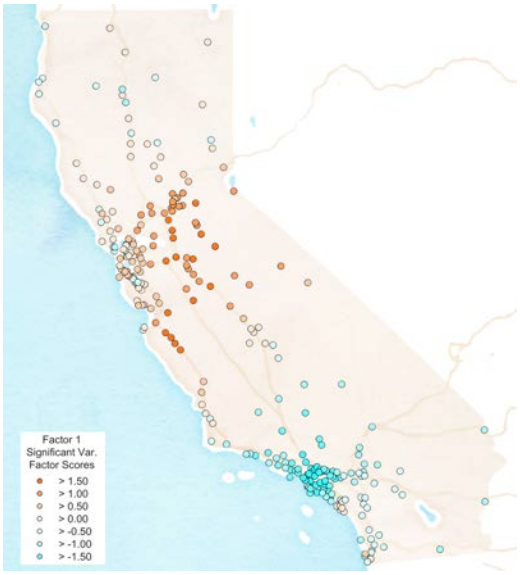

Figure 6: Map of Factor 1 for Significant Variables Only (22% Variance Explained).

Dissociating grammaticality and word-order choice: A case study on object pronouns in German

Markus Bader Sascha Dümig
Goethe University Frankfurt
bader@em.uni-frankfurt.de duemig@lingua.uni-frankfurt.de

1 Introduction

Determining the factors governing word order variation has become a central task of linguistic research. A long-standing question in this regard is how different factors interact, in particular performance factors like constituent weight and grammatical factors like animacy and definiteness. Can certain factors be reduced to a single basic factor (e.g. weight as claimed in Hawkins, 1994)? More recently, a further question concerning word order variation has gained some prominence: What is the relationship between the frequency of a given word order option and its perceived grammaticality/acceptability?

In order to address these questions, we present a corpus study and two experiments. The particular phenomenon under investigation is the placement of personal pronouns functioning as objects in German embedded clauses. Like other types of objects, object pronouns can precede or follow the subject, as in (1). According to descriptive grammars, subject-object (SO) and object-subject (OS) order are equally acceptable.

(1) Peter sagt, dass der Opa ihn besucht / dass ihn der Opa besucht
Peter says that the grandpa him visits that him the grandpa visits
'Peter says that grandpa will visit him.'

2 Corpus Data

In an earlier corpus study on object pronouns, Heylen (2005) showed that a range of different factors jointly determine the order of a non-pronominal subject and a pronominal object. Because Heylen's study was based primarily on reflexive pronouns (n=816) and contained only a small number of personal pronouns (n=179), we conducted a new corpus study. We extracted from the deWac corpus (Baroni et al., 2009) about 3600 complementizer-introduced subordinate clauses containing a personal pronoun object immediately preceded or followed by the subject. Overall, OS order occurred in about 62% of all cases. The sentences were coded for several properties, including length, animacy and definiteness of the subject and base-order of the verb (SO verbs including action verbs and OS verbs including object-experiencer verbs).

(i) Figure 1 (left) shows the effect of length for definite NPs in three types of sentences: animate subject/SO verb, inanimate subject/SO verb, and inanimate subject/OS verb. As shown by this figure, the weight of the subject has an effect on order even if definiteness, animacy, and base order are held constant. (ii) Figure 1 (right) shows the effect of definiteness and animacy for NPs consisting of two words. As shown by this figure, definiteness and animacy have a strong effect on order even if length is held constant. A logistic regression analysis including all factors simultaneously shows that weight, grammatical properties of the subject, and grammatical properties of the verb, all contribute to the probability of positioning an object pronoun before or after the subject.

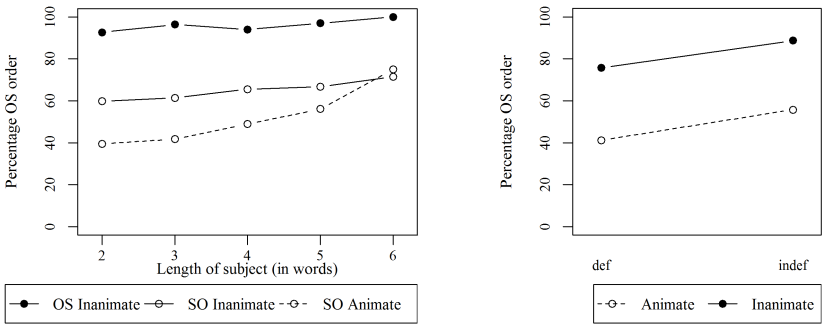

Figure 1: Percentage of OS order depending on subject length (left) and subject animacy and definiteness (right).

Table 1: Mean Percentages of sentences with OS order (standard error) in Experiment 1.

	SO main clause		OS main clause	
	Animate subject	Inanimate subject	Animate subject	Inanimate subject
Short	47 (7.7)	79 (5.6)	55 (6.2)	84 (4.6)
Long	60 (7.1)	75 (5.2)	74 (6.0)	90 (3.7)

3 Experimental Data

Bresnan and colleagues (e.g., Bresnan, 2007; Bresnan & Ford, 2010) have found close correspondences between corpus frequencies and rating judgments. However, their rating procedure conflates syntactic choice and acceptability. Participants had to rate the naturalness of the two alternatives of a syntactic alternation by distributing exactly 100 points across the two alternatives. For example, a pair consisting of a very natural and a rather unnatural alternative could get something like 90:10 points. Two equally natural alternatives could accordingly get only about 50 points each even if both were fully acceptable.

In order to dissociate syntactic choice and acceptability we ran two independent experiments. In a production experiment, participants first read a sentence (e.g., “Der Opa ärgert den Lehrer”- ’Grandpa bothers the teacher’). This sentence always contained an object-experiencer psych-verb. After a visual prompt with a main clause (“Der Lehrer sagt” - ’The teacher says’), the initial main clause had to be orally repeated from memory in the form of an embedded clause (“dass ihn der Opa/der Opa ihn ärgert” - ’that grandpa bothers him’). 40 sentences were constructed, each appearing in 8 versions according to the following three factors: (i) The order of subject and object within the main clause was either SO or OS. (ii) The subject was either animate or inanimate. (iii) The subject was either short (two words) or long (four words). The results are shown in Table 1. All main effects and the interaction between animacy and length were significant. The results from the production experiment are thus in close correspondence with the data obtained in the corpus study.

The second experiment was a standard magnitude estimation experiment (Bard et al., 1996). The material for this experiment was derived from the material of the production experiment. Each experimental sentence started with a main clause (“Der Lehrer sagt” - ’The teacher says’) which was followed by an embedded clause with either SO or OS order (“dass der Opa ihn ärgert” or “dass ihn der Opa ärgert”- ’that grandpa bothers him’). Two additional factors were the animacy and the length of the subject, as in the production experiment. The results are shown in Table 2.

The only significant effect in the magnitude estimation experiment was due to the factor length: Short sentences were rated as somewhat more acceptable than long sentences. Crucially, the factor order was not significant, nor any interaction involving this factor. Thus, even in those conditions where the production experiment and the corpus study show a clear preference for either SO or OS, the two orders do not differ in terms of acceptability.

Table 2: Mean acceptability scores (standard error) obtained in Experiment 2.

	Animate subject		Inanimate subject	
	SO	OS	SO	OS
Short	.34 (.036)	.33 (.031)	.34 (.034)	.35 (.035)
Long	.31 (.033)	.31 (.032)	.29 (.032)	.32 (.032)

4 Conclusion

The corpus study and the production experiment show that the position of an object pronoun relative to a non-pronominal subject in German is jointly determined by a complex interplay of grammtical properties and weight. Neither can be reduced to the other. In combination with the magnitude estimation experiment, the present results present a case where syntactic choice varies independently of acceptability (see also Featherston, 2005; Kempen & Harbusch, 2008). This suggests that probabilistic constraints are primarily a matter of the performance mechanisms, not the grammar.

References

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin: Mounton de Gruyter.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Featherston, Sam. 2005. The decathlon model of empirical syntax. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence*, 187–208. Berlin: de Gruyter.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence*, 241–263. Berlin: de Gruyter.

Kempen, Gerard & Karin Harbusch. 2008. Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In Anita Steube (ed.), *The discourse potential of underspecified structures*, 179–192. Berlin: de Gruyter.

Informativity is a predictor of semantic transparency in English compound nouns

Melanie J. Bell
Anglia Ruskin University
melanie.bell@anglia.ac.uk

Semantic transparency is known to play an important role in the storage and processing of complex words (e.g. Marslen-Wilson et al., 1994), and human raters of transparency achieve high levels of agreement (Sprouse, 2011), yet the phenomenon itself is poorly understood. For example, despite the fact that transparency is generally believed to be a gradient phenomenon (e.g. Wurm, 1997), most studies treat it as if it were categorical. In the case of bimorphemic compounds, a four-way distinction is often used, based on the perceived transparency of the constituents: transparent-transparent (e.g. *car-wash*), transparent-opaque (e.g. *jailbird*), opaque-transparent (e.g. *strawberry*) and opaque-opaque (e.g. *hogwash*) (Libben et al., 2003). In contrast, this paper presents a model of compound transparency as a continuous rather than discrete variable, which shows that the transparency of the constituents, and hence of the compound, is related to their frequency and other measures of informativity. It is the first study to show that information content, measured in terms of distribution, is predictive of perceived transparency.

The model uses the publicly available dataset collected for and described in Reddy et al. (2011). These authors selected a set of 90 English compound nouns from the ukWaC corpus. For each of the 90 compounds, Reddy et al. (ibid.) obtained literality ratings from human raters, who were asked to rate either (a) how literal they perceived the compound to be, or (b) how literally the first constituent was used in the compound or (c) how literally the second constituent was used in the compound. Each of these tasks was completed by thirty raters for each compound. To this dataset, I added various measures of informativity, namely the frequency and 'family size ratio' of each compound constituent as extracted from the BNC. Family size ratio is the number of compound types in which the constituent occurs in the same position (the positional morphological family size) divided by the number of types in which it occurs in the other position; in other words, it is a measure of the tendency of a constituent to occur in the left or right-hand position (cf. Baayen, 2010). The assumptions are firstly that more frequent constituents are more expected and hence less informative when they occur, and secondly that a constituent which occurs mainly in the left-hand (modifier) position of compounds will be less expected to occur in the right-hand (head) position, and will therefore be more informative when it does so (and vice versa). In addition, the data was coded for the semantic relation between the constituents (using the classification of Levi, 1978), for metaphorical shift in the meaning of either constituent or the compound as a whole, and for the extent of lexicalisation as measured by 'spelling ratio' (the proportion of tokens written unspaced, cf. Bell and Plag, 2012). These informativity and semantic variables were used as predictors in ordinary least squares regression analyses with literality of the compound or its constituents as the dependent variables.

The final model for overall literality of the compound, as given by the human raters, is shown in Table 1, where positive coefficients indicate a tendency towards higher literality and negative coefficients indicate a tendency towards lower literality. It can be seen that both types of predictor, semantic and frequency-based, are statistically significant, with significant interactions between the informativity measures. As might be expected, literality rating is lower when either constituent (N1 or N2), or the whole compound (NN), is metaphorical. Literality also falls as the proportion of unspaced tokens increases (i.e. as lexicalisation increases). On the other hand, certain semantic relations ('N2 is for N1' and

'N2 is in N1') are associated with greater literality. Most significant for this paper, however, are the two interaction effects, which are shown in Figure 1.

	Coef	S.E.	t	Pr(>|t|)
Intercept	-1.7358	0.3542	-4.90	<0.0001
N1 is metaphorical	-1.1312	0.0975	-11.61	<0.0001
N2 is metaphorical	-1.4783	0.0906	-16.32	<0.0001
NN is metaphorical	-2.0248	0.0867	-23.37	<0.0001
spelling ratio	-0.1078	0.0224	-4.82	<0.0001
semantic relation = In	0.2908	0.1215	2.39	0.0169
semantic relation = For	0.1921	0.0781	2.46	0.0140
logFreqN1	0.3715	0.0215	17.29	<0.0001
logFamSizeRatioN1	-1.8283	0.2436	-7.51	<0.0001
logFreqN2	0.2391	0.0344	6.95	<0.0001
logFamSizeRatioN2	3.9074	0.4906	7.96	<0.0001
logFreqN1*logFamSizeRatioN1	0.2262	0.0301	7.51	<0.0001
logFreqN2*logFamSizeRatioN2	-0.4134	0.0555	-7.44	<0.0001

Table 1: Final model for compound literality using semantic and frequency-based predictors, R^2 adj = 0.572

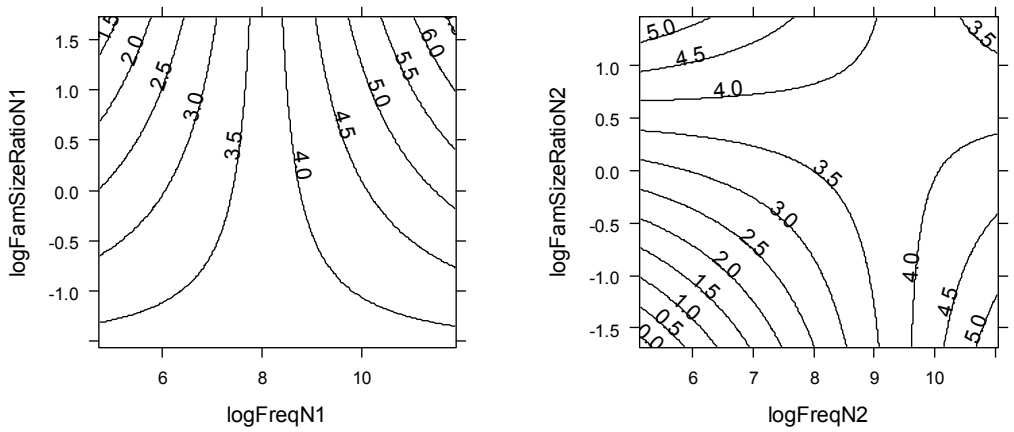

Figure 1: Partial effects in the final model for compound literality, informativeness predictors

The numbers on the contour lines in this figure give the ratings for literality: the human raters used a scale of 0-5, but the graphs extend beyond this range. Looking first at the left-hand plot we see that, in general, as the frequency of the left-hand noun (N1) increases, literality rating also increases. However, this effect is strongest when N1 has a high family size ratio, in other words when it typically occurs as a modifier. Overall, compounds are rated as most literal when N1 is a frequent word that typically occurs in the modifier position, and is therefore relatively expected and uninformative in that position. Looking at the right-hand plot, we see that literality is lowest when the frequency of the right-hand noun (N2) is low and its family size ratio is also low: in other words, when it is a low frequency word occurring in its non-preferred position, and is therefore highly unexpected and informative. Models for the literality of the individual constituents follow very similar patterns. On the assumption that literality is a measure of semantic transparency, this is the first evidence that transparency can be at least partially understood as the inverse of informativity.

References

Baayen, R. Harald. 2010. The directed compound graph of English: An exploration of lexical connectivity and its processing consequences. In *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, 383-402. Hamburg: Buske.

Bell, Melanie J. and Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48(3): 485–520.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Libben, Gary, Martha Gibson, Yeo Bom Yoon and Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84 (1): 50–64.

Marslen-Wilson, William, Lorraine Komisarjevsky Tyler, Rachelle Waksler and Lianne Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101(1): 3–33.

Reddy, Siva, Diana McCarthy and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Conference on Natural Language Processing, Chiang Mai, Thailand*, 210–218. AFNLP. All data available at: http://sivareddy.in/papers/files/ijcnlp_compositionality_data.tgz.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1): 155-167.

Wurm, Lee H. 1997. Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language* 37 (3): 438–461.

"Some rise by sin, and some by virtue fall"* Lexical convergence between Belgian Dutch and Netherlandic Dutch

Jocelyne Daems Kris Heylen Dirk Geeraerts
KU Leuven
{jocelyne.daems; kris.heylen; dirk.geeraerts}@arts.kuleuven.be

1 Background

Having more than one national variety, Dutch is considered a pluricentric language (Clyne, 1992). The main national varieties are Netherlandic Dutch (spoken in the Netherlands) and Belgian Dutch (spoken in Flanders, the northern part of Belgium). Interestingly, the process of linguistic standardization evolved differently in both regions. While the Netherlands independently developed a standard variant of Dutch, the standardization process of Belgian Dutch was delayed due to the influence of French. When the standardization of Dutch in Flanders resumed its process, an explicit exonormative orientation was adopted. Instead of developing a Belgian Dutch standard, convergence with the (long established) Netherlandic Dutch norm was promoted, aiming for a uniform Standard Dutch (Geeraerts, 2003).

To measure the convergence between the two national varieties, we will compare the word choice in the lexical field of sins and virtues. As such, this study represents a follow-up of Geeraerts et al. (1999), which looked at uniformity levels for clothing and football concepts in 1950, 1970 and 1990. The study confirmed the tendency of convergence between the two national varieties over the investigated time span and its attribution to the exonormative orientation of Belgian Dutch. In addition, from a synchronic point of view the distance between the standard and substandard language was distinguished as larger in Belgium than in the Netherlands. Although the results are readily interpretable and largely parallel for both lexical fields, their extrapolation to other lexical fields or other parts of speech requires further research. For instance, building on this tradition, Impe and Speelman (2007) investigate the role of attitudes vis-à-vis different varieties of Belgian Dutch and Plevoets (2008) zooms in on the morphological characteristics of the substandard Belgian Dutch variety, also called Colloquial Belgian Dutch (CBD). CBD is also elaborated on more generally in Geeraerts (2011) and from a lexical point of view in Zenner et al. (2009). Closer to the original study of Geeraerts et al. (1999) are for instance Grondelaers et al. (2001b), through their inclusion of content words and prepositions, and the exploration of the methodological possibilities in lexical lectometry by Ruette (2012). Then, with the extension to sins and virtues, we acquire not only more data (i.e. a new lexical field), but we can also examine the role of part of speech and the impact of the lexical field itself.

2 Method

This paper sets out to empirically test to what extent there currently is convergence between Belgian Dutch and Netherlandic Dutch, both in the standard and in the substandard language variety. Focusing on lexical uniformity, we rely on the onomasiological measure of lexical variation designed by Geeraerts et al. (1999), which calculates the differences in lexicalization preferences for a given concept in the two regions. For example, Table 1

*William Shakespeare, *Measure for Measure*, Act II, Scene 1

shows the concept NIJD 'envy', which can be lexicalized by the six near-synonyms *afgunst*, *ijverzucht*, *jaloersheid*, *jaloezie*, *na-ijver* and *nijd*. We call the profile for NIJD the whole of the alternative lexicalizations within a source (e.g. Belgian Dutch quality newspapers) together with its specific frequency distribution. The degree of uniformity between Belgian Dutch and Netherlandic Dutch can then be measured in terms of overlapping lexicalization preferences. That is, when summing the smallest relative value for each term from the two profiles, we get the proportion of the two profiles' shared lexicalisation preferences, or in other words, the degree of uniformity: (23.87 + 0.03 + 0.67 + 57.57 + 3.73 + 4.15) = 90.03%).

NIJD	Neth.Dutch	%	Belg.Dutch	%
afgunst	399	24.42	746	23.87
ijverzucht	1	0.06	1	0.03
jaloersheid	11	0.67	126	4.03
jaloezie	1094	66.96	1799	57.57
na-ijver	61	3.73	246	7.87
nijd	68	4.15	207	6.61

Table 1: Lexicalization preferences for NIJD in quality newspapers

We also incorporate a stratificational dimension by looking at the uniformity among standard and substandard language, which expectedly is lower in Belgian Dutch than in Netherlandic Dutch due to the delayed (and supposedly incomplete) standardization of Belgian Dutch. In this respect, Grondelaers et al. (2001a) demonstrate the value of Usenet, an online newsgroup system, as a source for CBD material. Finally, to get a better understanding of the role of exogenous and endogenous terms, of words of foreign origin, and of terms either propagated or rejected in the purist literature, we measure their proportion for each concept by taking into account the weighted relative frequency of these terms.

3 Data and results

On the basis of a data set of more than 550 million words of Belgian Dutch and Netherlandic Dutch, we apply the methodology of profile-based uniformity to concepts of sins (e.g. NIJD 'envy') and virtues (e.g. IJVER 'diligence'). Focusing on uniformity levels for both nouns and adjectives, we are able to look at the influence of part of speech. The impact of register on uniformity is measured by comparing uniformity tendencies in Usenet material and quality newspapers. Preliminary results confirm the high level of convergence between standard Belgian Dutch and standard Netherlandic Dutch, while the levels are significantly lower for the substandard variants. In addition, uniformity levels for virtues rather than sins show large discrepancies, with Belgian Dutch scoring rather low and Netherlandic Dutch much higher.

The study of the lexical field of sins and virtues fits in with a larger project which analyses 40 emotive concepts, 20 IT concepts and 20 traffic concepts. A similar study is found in Zenner et al. (2012) on 149 person reference nouns (such as RUGZAKTOERIST 'backpacker'). The various natures of these lexical fields, in particular with regard to the contact between the two national varieties and the proportion of foreign terms, allow for various comparisons. Lastly, our results will be set against the uniformity levels obtained by Geeraerts et al. (1999).

References

Clyne, Michael (ed.). 1992. *Pluricentric languages: Differing norms in Different nations.* BerlinNew York: Mouton de Gruyter.

Geeraerts, Dirk. 2003. Cultural models of linguistic standardization. In: Dirven, Roslyn

and Pütz (eds.), *Cognitive Models in Language and Thought. Ideology, Metaphors and Meanings*, 25–68. Berlin: Mouton de Gruyter.

Geeraerts, Dirk. 2011. Colloquial Belgian Dutch. In: Soares da Silva, Torres and Gonçalves (eds.), *Línguas Pluricêntricas. Variação Linguística e Dimensóes Sociocognitivas*, 61–74. Braga: Publicações da Faculdade de Filosofia, Universidade Católica Portuguesa.

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen.* Amsterdam: P.J. Meertens-Instituut.

Grondelaers, Stefan, Dirk Geeraerts, Dirk Speelman and José Tummers. 2001. Lexical standardisation in internet conversations. Comparing Belgium and The Netherlands. In: Fontana, McNally, Turell and Enric Vallduví (eds.), *Proceedings of the First International Conference on Language Variation in Europe*, 90–100. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, Unitat de Investigació de Variació Lingüística.

Grondelaers, Stefan, Hilde Van Aken, Dirk Speelman and Dirk Geeraerts. 2001. Inhoudswoorden en preposities als standaardiseringsindicatoren. De diachrone en synchrone status van het Belgische Nederlands. *Nederlandse Taalkunde* 6: 179–202.

Impe, Leen and Dirk Speelman. 2007. Vlamingen en hun (tussen)taal – een attitudineel mixed guiseonderzoek. *Handelingen van de Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde* 61: 109–128.

Plevoets, Koen. 2008. Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands. Unpublished PhD thesis, *KU Leuven*.

Ruette, Tom. 2012. Aggregating Lexical Variation. Towards large-scale lexical lectometry. Unpublished PhD thesis, *KU Leuven*.

Ruette, Tom, Dirk Speelman and Dirk Geeraerts. 2011. Measuring the lexical distance between registers in national varieties of Dutch. In: Soares da Silva, Torres and Gonçalves (eds.), *Línguas Pluricêntricas. Variação Linguística e Dimensóes Sociocognitivas*, 541–554. Braga: Publicações da Faculdade de Filosofia, Universidade Católica Portuguesa.

Zenner, Eline, Dirk Geeraerts and Dirk Speelman. 2009. Expeditie Tussentaal: Leeftijd, identiteit en context in "Expeditie Robinson". *Nederlandse Taalkunde* 14: 26–44.

Zenner, Eline, Dirk Speelman and Dirk Geeraerts. 2012. Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics* 23: 749–792.

Temporal information affects implicit causality biases in pronoun resolution

Jeruen E. Dery Dagmar Bittner
Zentrum für Allgemeine Sprachwissenschaft
{dery, bittner}@zas.gwz-berlin.de

1 Overview

Implicit causality (IC) refers to the observation that certain verb classes tend to prefer statistically reliable causal antecedents (Garvey and Caramazza, 1974). These causal biases can affect processes of pronoun resolution. For example, in *John confessed to/punished Bill because he…*, it was observed that the pronoun more likely refers back to *John* given verbs that are biased toward the first-mentioned noun phrase (NP1) such as *confess*, but more likely refers back to *Bill* given verbs that are biased toward the second-mentioned NP (NP2) such as *punish*. Hence, one long-standing debate in IC concerns whether the IC bias is part of the semantics of the verbs (Brown and Fish, 1983; Crinean and Garnham, 2006) or are just probabilistic notions derived by abstracting over likely causes of events (Pickering and Majid, 2007; Bott and Solstad, submitted).

In our study, we use the discourse's temporal dimension to contribute to this debate and see whether a verb's IC bias can be affected by the discourse's temporal properties. Recent findings (Solstad, 2010; Bott and Solstad, submitted) suggest that there are strong associations between different explanation types and the referent that this explanation is attributed to. In other words, a verb's IC bias may be due to differences in the types of causes typically associated with the events these verbs describe. Events can occur as a result of a) a simple cause (SC) that directly causes an eventuality; b) externally-anchored reasons (ER), in which the source of an agent's attitude is external to the agent; and c) internally-anchored reasons (IR), in which the source of the agent's attitude is internal to the agent. Bott & Solstad (submitted) suggest that a verb's IC bias is due to this difference: internal reasons tend to co-occur with an NP1 bias, while external reasons co-occur with an NP2 bias. Example (1) below illustrates these three different explanation types.

(1) a. SC: John disturbed Mary because he was making lots of noise.
b. ER: John disturbed Mary because she had damaged his bike.
c. IR: John disturbed Mary because he was angry at her.

2 Experiments

We hypothesize that manipulating the discourse's temporal dimension will have an effect on whether an event is caused by external or internal reasons; and on its IC bias. Experiment 1 used 124 German verbs in an offline comprehension-to-production sentence-continuation task. Participants provided plausible continuations to simple active present-tense sentences and the connective *because*, as in (2).

(2) Karlo hilft Jenny, weil _____.

We annotated participants' responses for choice of referent and temporal location of the explanations they provided. We found a significant relationship between the explanations' temporal location and referent choice: explanations temporally located in the past were more likely to have NP2 reference, while those located in the present were more likely to have NP1 reference.

Experiment 2 was another sentence-continuation task with a 3x2 design crossing IC type (i.e., NP1, NP2, non-IC); and temporal location (i.e., today, yesterday) of the explanation clause, as in (3).

(3) Karlo hilft Jenny, weil _____ heute/gestern _____.

Unlike Experiment 1, we measured participants' conceptualization time (the time it took them to evoke a plausible continuation). Participants viewed the prompts on a computer, and were instructed to think of a plausible continuation. They were to press a button on the keyboard only when they have evoked a continuation in their head. A message box then appears on the computer where participants typed their responses. Conceptualization times were recorded, measured from when the sentential prompts first appeared on the screen until when participants pressed the button to type their continuation.

We predicted and confirmed that the proportion of pronouns referring back to NP1 or NP2 differs as a function of both IC type ($p<0.001$) and temporal location of the explanation clause ($p=0.02$). Simple effects tests looking at the effect of temporal location on each of the IC types reveal that while the present/past manipulation has no effect on NP2 verbs ($p=0.87$), the temporal manipulation was able to shift the pronoun biases for NP1 verbs ($p=0.04$) and non-IC verbs ($p=0.03$). While NP1 verbs show a clear bias toward NP1 reference in the present condition, this bias is weakened in the past condition. On the other hand, while non-IC verbs do not show a referential bias in the present, it shows a bias toward NP2 reference in the past condition. This result suggests that while external reasons do not have any temporal restrictions, internal reasons seem to require temporal overlap between cause and effect. Forcing non-overlap by means of a temporal adverb seems to lower the probability of an internal reason, causing IC biases to shift.

With respect to conceptualization time, we observed (i) a main effect of IC type ($p=0.001$): participants spent more time thinking about plausible explanations for events involving NP1 verbs than NP2 verbs; and (ii) a main effect of temporal location of the explanation clause ($p=0.001$): participants spent more time thinking about plausible explanations temporally located in the past than in the present. We also observed a marginally significant interaction between IC type and temporal location ($p=0.056$), indicating that while there is no significant difference in the conceptualization time of present and past explanations for NP2 verbs ($p=0.51$), there is a marginally significant difference for non-IC verbs ($p=0.055$), as well as a significant difference for NP1 verbs ($p=0.01$). These results point to the apparent incompatibility of internal reasons in the past: while participants spent an equal amount of time evoking present or past explanations for NP2 verbs, this was not the case for NP1 verbs, suggesting a high preference for temporal overlap between cause and effect for NP1 verbs.

3 Discussion

Our results lend support to the claim that there are fundamental differences on the semantic restrictions underlying causal relationships involved in implicit causality. There seem to be restrictions on temporal properties of causes and their effects: while externally-anchored reasons (NP2 verbs) do not seem to have any temporal constraints with the effects they induce, internally-anchored reasons (NP1 verbs) seem to prefer effects that are temporally proximal. Our results are incompatible with the view that IC biases are part of a verb's meaning: a verb's meaning should not change depending on the temporal location of its explanation. Our results support the view that IC biases are just probabilistic notions derived by abstracting over the likely causes of events.

References

Bott, Oliver, and Torgrim Solstad. Submitted. From verbs to discourse: A novel account of implicit causality. in C. Fabricius-Hansen *et al* (eds.): *Psycholinguistic approaches to meaning and understanding across languages*, Springer.

Brown, Roger, and Deborah Fish. 1983. The psychological causality implicit in language. *Cognition*. 14, 237--273.

Crinean, Marcelle, and Alan Garnham. 2006. Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes*, 21, 636--648.

Garvey, Catherine, and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry*. 5, 459--464.

Pickering, Martin, and Asifa Majid. 2007. What are implicit causality and consequentiality? *Language and Cognitive Processes*, 22, 780--788.

Solstad, Torgrim. 2010. Some new observations on 'because (of)', in M. Aloni *et al.* (eds.) *Logic, Language and Meaning*. Berlin: Springer, pp. 436--445.

Ordinary Kriging in Dialectology

Jack Grieve
Aston University
j.grieve1@aston.ac.uk

There are many situations that arise in dialectology where it is useful to be able to estimate the values of a linguistic variable at one or more unobserved locations. The standard approach to spatial interpolation in geography and other fields is known as ordinary kriging (Isaaks and Srivastava, 1989), although ordinary kriging in particular and interpolation in general has rarely been applied in dialectology (although see Grieve, 2013). This presentation will therefore introduce ordinary kriging and demonstrate how this geostatistical technique can be applied in dialectology by presenting a series studies that focus on regional variation in American English. In addition, the presentation will introduce the implementation of these techniques in R (see Bivand et al, 2008).

Ordinary kriging is a method for interpolating the value of a variable at an unobserved location based on the values of that variable at observed locations. Specifically, ordinary kriging estimates the value of a variable at an unobserved location by taking a weighted average of the values of the variable at observed locations, where these weights are based on both the distance separating the locations and the variogram for that variable, which is a function that describes the amount of spatial variability in the values of a variable measured over a series of locations (Isaaks and Srivastava, 1989). A variogram plots the variance between locations against the distance between locations, in essence providing a model of how the values of a variable change across space. Once a variogram has been estimated for a variable based on the values of that variable at observed locations, ordinary kriging can then be used to estimate the values of that variable at unobserved locations.

After introducing variogram analysis and ordinary kriging, a variety of different applications of ordinary kriging in dialectology will be demonstrated through analyses of regional linguistic variation in American English. In particular, the use of ordinary kriging for facilitating prediction, visualization, comparison, and aggregation of regional linguistic data will be discussed, based on a variety of American English datasets, including phonetic data from the Atlas of North American English (Labov et al., 2006), grammatical data from a corpus of written American English (Grieve et al., 2011), and lexical and phonological data from the Harvard Dialect Survey (Vaux, 2003).

First, the basic use of ordinary kriging to predict the values of linguistic variables at unobserved locations will be demonstrated through an analysis of a variety of individual linguistic variables.

Second, the use of ordinary kriging to estimate the values of a linguistic variable across an entire region at a very high level of resolution will be discussed. This is an especially powerful method for visualizing dialect data, as it allows for general patterns of regional variation to be mapped based on only the values of a relatively small number of known locations. For example, Figure 1 plots the values of an aggregated phonetic variable that was originally measured across 236 observation points after being interpolated across approximately 20,000 regularly spaced locations using ordinary kriging. In this way ordinary kriging can be used to plot isoglosses, dividing a region into sub-regions where the different values of the variable predominate.

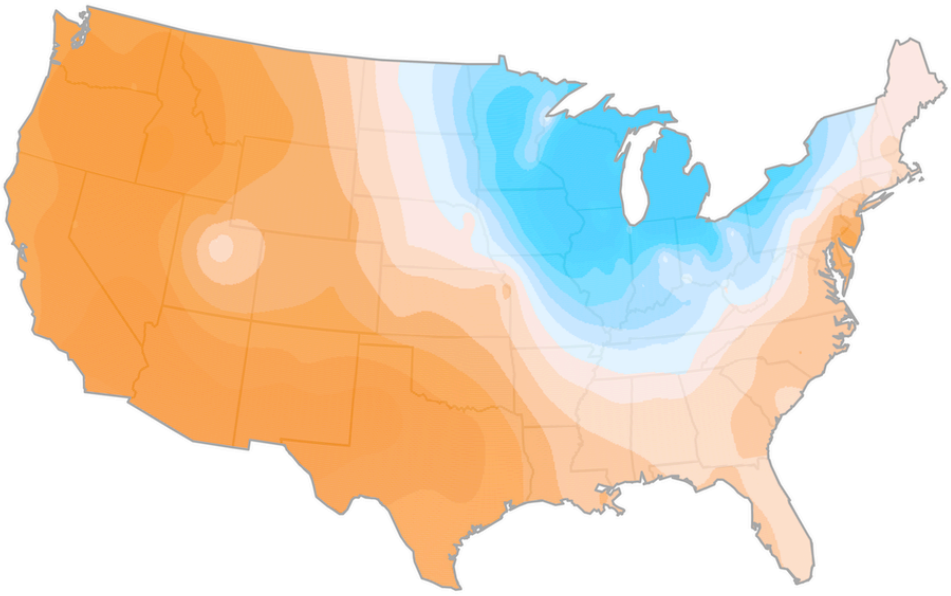

Figure 1: Example of Ordinary Kriging

Third, the use of ordinary kriging to facilitate the comparison of dialect maps that are based on different sets of locations will be demonstrated by interpolating maps from different dialect surveys over a regular grid of reference locations. These interpolated maps will then be correlated with each other in order to measure the similarity between the regional patterns identified in these various dialect surveys.

Finally, the use of ordinary kriging to facilitate the aggregation of dialect maps that are based on different sets of locations will be demonstrated by aggregating linguistic variables from different American dialect studies after each of the individual linguistic variables have first been interpolated over a regular grid of reference locations.

References

Bivand, Roger S., Pebesma, Edzer. J., and Gomez-Rubio, Virgilio. (2008). *Applied Spatial Data Analysis with R*. New York: Springer.

Grieve, Jack, Speelman, Dirk, and Geeraerts, Dirk. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23: 193–221.

Grieve, Jack. (2013). A statistical comparison of regional phonetic and lexical variation in American English. *Literary and Linguistic Computing*, 28: 82-107.

Isaaks, Edward H., and Srivastava, R. Mohan. (1989). *An Introduction to Applied Geostatistics*. Oxford, UK: Oxford University Press.

Labov, William, Ash, Sharon, and Boberg, Charles. (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

Vaux, Bert. (2003). American dialects. In: Steven Goldberg (Ed.) *Let's Go USA 2004*. Upper Saddle River, NJ: Prentice Hall.

(In)direct Causation Hypothesis Again: A Case Study of Chinese Analytic Causatives

Yanan Hu Dirk Geeraerts Dirk Speelman
KU Leuven
yanan.hu@student.kuleuven.be,
{dirk.geeraerts, dirk.speelman}@arts.kuleuven.be

The research examines the (in)direct causation hypothesis formulated first by Verhagen and Kemmer (1997), analyzed by Stukker (2005), falsified, mostly if not completely, by Speelman and Geeraerts (2009) from a Chinese perspective.

Theoretical Starting-point

A series of studies on the (in)direct hypothesis address the choice of *doen* and *laten* causative verbs in contemporary Dutch. Stukker (2005) associates *doen* with direct causation, and *laten* with indirect causation. Falsifying this hypothesis, Speelman and Geeraerts (2009) pursue a different one that *doen*, as a causative, "is an obsolescent form with a tendency towards semantic and lexical specialization." But more recently based on the (in)direct hypothesis again, Ni's thesis (2012) states that in Mandarin Chinese "*shi* is similar to *doen* in Dutch in that it is related to the inanimate entity as the causer part and it expresses the direct causation, and *rang* is related to the animate entity, just as *laten* in Dutch and it expresses the indirect causation." So this claim that Chinese *shi* and *rang* are the equivalents to Dutch *doen* and *laten* is so risky that needs to be tested.

Research Questions:

Starting with the assumption put forward by Ni (2012), we apply the statistical techniques developed in Speelman and Geerearts (2009) to specifically address the following questions:

(1) Does (in)direct causation hypothesis work for Chinese?

-- Do the factors related to the predictions derived from (in)direct causation hypothesis play a role in distinguishing Chinese analytic causatives *shi* and *rang*?

(2) If it does work for the Chinese case, how well does it work? Enough?

-- If (in)direct causation hypothesis does capture some difference between *shi* and *rang*, as Ni (2012) put, how significant is it? Is it an adequate reason for language users to choose either of them?

(3) Are there other possible scenarios for these two near-synonyms?

-- Is there any possibility that Chinese is another case, which doesn't settle for the (in)direct causation distinction but confirms the multivariate conception of the grammar suggested by Speelman & Geeraerts (2009)?

(4) What can we tell about (dis)similarities between Chinese and Dutch causatives?
-- After scrutinizing, can we still claim *shi* and *rang* are the equivalents of *doen* and *laten*? How (dis)similar are their distributions in the two languages?

Data and Methods:

The materials in this case study subsume two parts. First part is taken from Corpus Online (www.cncorpus.org) developed by the National Language Committee of China, which provides about 20 million characters of modern Chinese. We start from random sampling so that I get the comparable number of total observations, 4078 sentences (3261 cases of *shi*, 817 cases of *rang*) and then code these occurrences with the predictors assumed in the literature and related closely to (in)direct causation hypothesis (e.g. inanimateness of causer). Second part includes two corpora, Sheffield Corpus of Chinese (http://www.hrionline.ac.uk/scc/db/scc/index.jsp) and UCLA Chinese Corpus (http://www.lancs.ac.uk/fass/projects/corpus/UCLA/) 1st edition. So it has a time strata to itself and covers chronological spans of mandarin Chinese from 1100 BC to 2005 AD. But the size is relatively small, with 1764 sentences in total (shi 807, rang 957). And we code this part with more variational predictors, which may diverge from those (in)direct causation related ones in order to build a comparative model.

We implement binomial logistic regression and multiple correspondence analysis in the R project so as to statistically test significance and explaining power of those potential factors on the use of *shi* and *rang*. Last but not the least, we compare and contrast the newly acquired Chinese results with the Dutch ones.

Results and Interpretation:

The statistic shows the (in)direct causation hypothesis can tell some difference between Chinese causatives *shi* and *rang*. Although it is not unimportant, it's far from powerful enough to capture all the significant variation. It's safe for us to say it's a relatively minor way of taxonomy since only about 30% data has been explained by the (in)direct causation model. So there are plenty of factors which simultaneously draw the entire picture of Chinese causatives, at least the two main ones in the current study, for example, time periods, lexical fixation between causative auxiliaries and their causer or causee, and even some language-specific factors. Since both Chinese and Dutch causatives turn out to be complicated and beyond complete grip of (in)direct causation hypothesis, what we can do is to draw upon the overlaps of the previous Dutch research and our present study to show the (dis)similarities of their usages in the two languages.

Significance and Further researches:

This investigation is supposed to complement the existing researches on causatives, re-examine the (in)direct causation hypothesis from a Chinese standpoint, (dis)prove Ni's inference of the equivalence of *shi*, *rang* vesus *doen* and *laten*, unveil cross-linguistic correspondences or contrasts in the linguistic construal of causality. The present study is also part of a broader line of ongoing project in our research group.

References

Ni, Yueru. 2012. *Categories of Causative Verbs: a Corpus Study of Mandarin Chinese*. Utrecht: Utrecht University MA thesis.

Speelman, Dirk and Dirk Geeraerts. 2009. Causes for causatives: the case of Dutch 'doen' and 'laten'. In Ted Sanders and Eve Sweetser (eds.), *Causal Categories in Discourse and Cognition* 173-204. Berlin/New York: Mouton de Gruyter.

Stukker, Ninke. 2005. *Causality marking across levels of language structure*. PhD dissertation, University of Utrecht.

Verhagen, Arie and Suzanne Kemmer. 1997. Interaction and Causation: Causative Constructions in Modern Standard Dutch. *Journal of Pragmatics* 27. 61–82.

Judgment and production data in morphophonology: Converging sources of evidence

Vsevolod Kapatsinski, Amy Smolek, and Matthew Stave
University of Oregon
{vkapatsi;asmolek;mstave}@uoregon.edu

The relationship between acceptability judgment data and production data is a controversial topic in linguistic methodology. Some prominent researchers have suggested that one or the other of these sources of data should be privileged as providing the most direct window on the grammar (Labov, 1996, vs. Chomsky, 1965). More recent work has acknowledged that both types of data have their place in the study of grammar (Kepser and Reis, 2005). However, the arguments in favor of using acceptability judgments have hinged on the problem of data sparseness in studies of syntax: some constructions may be too infrequent to be observed even in a large corpus (Schütze, 2011). This issue does not often arise in (morpho)phonology, at least for commonly-studied languages. Neither does it arise in studies where a miniature artificial language is taught to participants in a laboratory (a focus of our work). Nonetheless we argue that acceptability judgments provide us with data that complement production data even when sparseness is not an issue.

A fundamental property of production is competition among alternative outputs. Faced with the novel verb fring, a speaker trying to produce the past tense form chooses to say *frung*, *frang*, *frought* or *fringed*. All of these alternative past tense forms compete with each other for production. Nonetheless, as Bybee and Slobin (1982) argue, *fringed* may be more like *frought* than it is like *frang*: both *fringed* and *frought* end in an alveolar stop. We show that judgment data can help capture similarity between alternative outputs: while production probabilities of all alternative outputs for a given input correlate negatively across speakers, judgments of similar outputs can show positive correlations (contra Albright and Hayes 2003). For instance, subjects who like *frought* as the past tense of fring may also judge fringed as being more acceptable than subjects who do not like frought (Kapatsinski 2007; 2012). These positive between-subject correlations suggest that the grammar contains a generalization that supports both input-output mappings (here, perhaps, that past tense forms end in an alveolar stop; Bybee and Slobin 1982): the subjects who assign a relatively high weight to this generalization like both outputs, while those who assign it a low weight dislike both outputs.

A second property of elicited production data is a bias against changing the input (Kapatsinski, 2012; Mitrović, 2012; Zuraw 2010). Participants may be unlikely to produce a stem change and yet judge the output of that change as being more acceptable than an alternative output that has not undergone the change (in Zuraw 2010, the same Tagalog subjects do not perform stop deletion but yet judge stop-less forms that result from it as being more acceptable than forms that retain the stops). We have replicated this finding with a miniature artificial language paradigm where adult English-speaking participants learn that either [p], [t], or [k] becomes [tʃ] before the plural suffix –a (palatalization). While participants rarely turn the stops into [tʃ] in production (Figure 1), they prefer plurals ending in [tʃa] to plurals ending in [ta], [pa], or [ka] in a judgment task (Figure 2). We also show that there is a bias against changing [p] compared to [t] and [k] (Figures 1 and 2) and that this bias is stronger in the production data (there is a significant three-way interaction between test modality, whether or not a consonant is supposed to be palatalized, and subject group; $z=2.37$, $p=.018$ based on a logistic mixed effects model with maximal random effects structure fit using the lme4 package in R; note that we can test this interaction because our judgment data, like production data, are binary). We suggest that a speaker producing an unknown (here, plural) tends to perseverate on gestures com-

prising the known (here, singular) form, which results in suppression of stem changes. Since production is an ecologically valid task faced by speakers every day, this bias against stem changes may explain why languages do not usually feature 'crazy' stem changes like *mail-membled* (Pinker and Prince, 1988). Nonetheless, judgment data are likewise informative about patterns of grammar change: stem changes that appear unproductive in elicited production data may nonetheless persist in languages (Kapatsinski, 2010; Köpcke and Wecker, 2013). This suggests that the unchanged forms favored in production are nonetheless disfavored in loanword adaptation, presumably because they are judged to be inferior to the changed forms (Zuraw 2000).

The data in Figure 2 illustrate another reason to suppress competition between outputs using a judgment task. Production data tell us that there is a bias against palatalizing [p]. However, they do not tell us whether this bias is because speakers dislike changing [p] into [tʃ(a)] or because they like [pa] more than they like [ka] or [ta]. Because [tʃa] competes with [pa], [ta], or [ka] for production, a preference for the former cannot be distinguished from avoidance of the latter. Judgment data allow us to distinguish the two: while participants do not learn to palatalize [p] as well as they learn to palatalize [k] and [t], they learn to dislike [pa], [ka], and [ta] equally well.

In conclusion, we suggest that judgment data complement production data in the study of morphophonology by 1) revealing similarities among alternative outputs, 2) distinguishing between attraction to an output and avoidance of a competitor, and 3) reducing perseveration on gestures comprising the input form.

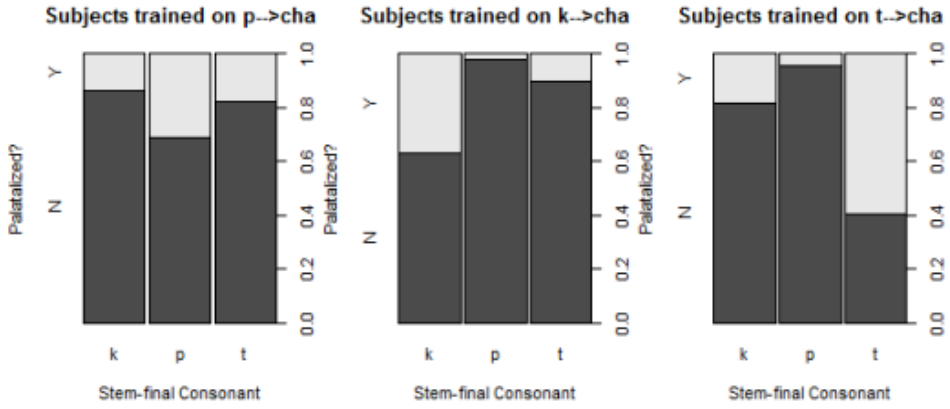

Figure 1: Elicited production data. Vertical axes show whether or not the consonant was palatalized (Y, shown in light, is "palatalized", N, in dark, is "not palatalized"). Participants exposed to labial palatalization before [a] palatalize [t] and [k] almost as much as they palatalize [p] whereas participants exposed to alveolar or velar palatalization palatalize [t] or [k] much more than other stops.

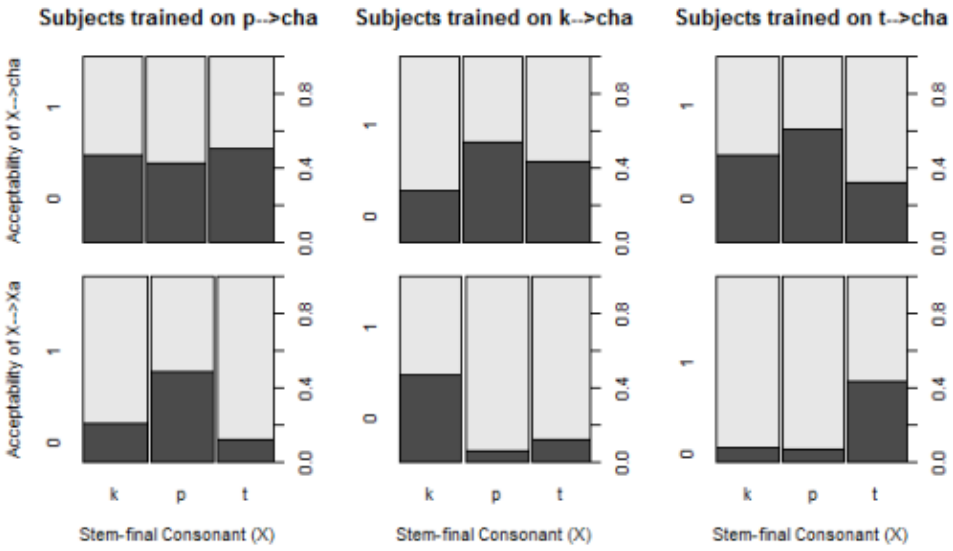

Figure 2: Judgment data. Dark parts of bars: 'this is the wrong plural form for this singular' responses. Light parts of bars: 'this is the right plural form for this singular' responses. Bottom row: after all kinds of training subjects learn to reject unchanged/non-palatalized stops before the palatalizing vowel [a], and they do it at equal rates.

References

Albright, Adam, & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90: 119--61.

Bybee, Joan, & Dan Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58: 265--89.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Kapatsinski, Vsevolod. 2012. What statistics do learners track? Rules, constraints or schemas in (artificial) grammar learning. In: Gries and Divjak (eds.), *Frequency effects in language learning and processing*, 53--82. Berlin: Mouton de Gruyter.

Kapatsinski, Vsevolod. 2010. Rethinking rule reliability: Why an exceptionless rule can fail. *Chicago Linguistic Society* 44: 277--91.

Kapatsinski, Vsevolod. 2007. To scheme or to rule: Evidence against the Dual Mechanism Model. *Berkeley Linguistics Society* 31: 193--204.

Kepser, Stephan, & Marga Reis. 2005. Evidence in linguistics. In Kepser & Reis (eds.), *Linguistic evidence*, 1--6. Berlin: Mouton de Gruyter.

Köpcke, Klaus-Michael, & Verena Wecker. 2013. The L2 acquisition of the German plural - Evidence for usage based models of language acquisition. Paper presented at DGfS 2013, AG7: Usage-based Approaches to Morphology, Potsdam, March 13-15.

Labov, William. 1996. When intuitions fail. *Chicago Linguistic Society* 32: 77--106.

Mitrović, Ivana. 2012. A phonetically natural vs. native language pattern: An experimental study of velar palatalization in Serbian. *Journal of Slavic Linguistics* 20: 229--68.

Pinker, Steven, & Alan Prince. 1988. On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition* 23: 73–193.

Schütze, Carson. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2: 206-21.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Ph.D. Dissertation, UCLA.

Sound change and hierarchical inference: Clarifying predictions of usage-based theory

Vsevolod Kapatsinski
University of Oregon
vkapatsi@uoregon.edu

Bybee (2001) identifies two word frequency effects in sound change. Reductive sound change is supposed to be due to repetition and begin in high-frequency words. As a result, Bybee predicts that the least reduced words should be the least frequent words. However, as Bybee also notes, high-frequency words are better able to resist analogical change due to imperfect acquisition of low-frequency lexical representations. For instance , whereas low-frequency English verbs are regular, having succumbed to –ed, many high-frequency verbs have retained their original past tense forms. Note that this analogical pressure is also in effect for articulatorily-motivated sound changes. This makes the prediction that reductive sound change should affect lowest-frequency words least questionable.

I implement the theory computationally in R and show that the theory actually predicts that reductive sound change should indeed affect high-frequency words first but once the change spreads sufficiently, low-frequency words are expected to fall in line, with some medium frequency words remaining exceptional.

Probability of reduction is treated as being due to 1) the overall tendency of the speaker or speakers we are studying to use the reduced variant, 2) the frequency of the word in which the variable occurs (whenever a word is used, its probability of being reduced is incremented), and 3) the identity of the word: some words are reduced more or less than their frequency would predict (Bybee 2002, Pierrehumbert 2002, Raymond & Brown 2012, Yaeger-Dror & Kemp 1992). Every generation of speakers reduces words in proportion to their frequency of use in speech (Bybee 2001, 2002, Pierrehumbert 2001). However, in L1 learning, each generation does not explicitly try to recover the function relating word frequency to probability of reduction. Rather the task of the learner is to learn to pronounce words correctly.

Following Labov (1969) and Pierrehumbert (2002), among others, I assume that as part of this process the learner acquires a probabilistic grammar of reduction, which specifies how often one picks a particular variant of a sublexical phonological structure in various contexts. This grammar allows the learner to, among other things, pronounce unfamiliar words and to adapt pronunciation to social context (as argued in Pierrehumbert 2002).

On this theory, neither the classical lexical diffusionist position ("every word has its own history", Schuchardt 1885), nor the Neogrammarian position ("sounds change", Osthoff & Brugmann 1878) are entirely correct, nor is Labov's (1981) compromise position where each change exhibits either lexical diffusion or Neogrammarian regularity. I propose that the learner blames neither only words nor only sublexical structures for the acoustics of a particular experienced token. Rather, blame for the perceived acoustics is apportioned on the basis of hierarchical inference (here implemented using a linear mixed effects model: the lme4 package in R, Bates 2013).

Thus the learner estimates overall probability of the reduced variant and the effects of individual words on reduction. The number of possible words is infinite and the number of observations per word is often insufficient to reliably estimate probability of reduction for an individual word. Therefore word identity is treated as a random effect. As a result, reduction probability estimates from individual words constrain each other: parameter estimates that are extreme outliers are brought into the fold, especially if those estimates are based on few observations (Gelman & Hill 2007) as is the case for low-frequency words.

Figure 1 shows that if this theory is correct, the frequency effect, monotonic in its online effect on production, will nonetheless be U-shaped in the lexicon, with low-frequency words being pulled in to not deviate from the overall mean.

The idea that lexical diffusion patterns are in part due to hierarchical inference predicts that once a sound change has progressed far enough, increased word frequency should increase the probability of reduction only for words that are frequent enough. The hypothesis allows us to also make more specific predictions for the variables influencing patterns of lexical diffusion across and within individuals. These include 1) the distribution of word frequencies for words that are potentially affected by the change due to meeting its structural description: if these words happen to be mostly low in token frequency and if the class of words potentially affected is numerous, the change is likely to exhibit weaker lexical diffusion effects, particularly with respect to the difference between low-frequency and medium-frequency words; and 2) the speakers' willingness to jump to conclusions about the behavior of a word on the basis of limited evidence, which may be affected by personality characteristics (for instance, local processing bias may cause one to focus on sublexical phonological structures rather than the more global lexical structures, Happé 1999, Yu 2010, leading speakers high on the autism spectrum to be likely to exhibit weaker lexical diffusion effects, again, especially for low-frequency vs. medium-frequency words where inferential biases are most important).

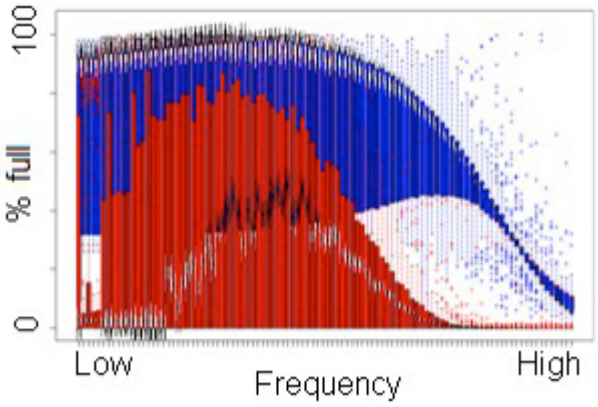

Figure 1: The effect of word frequency on probability of resisting reduction as sound change spreads through the lexicon. Generation 1 (blue) vs. 5 (red). Notches show 95% confidence intervals of the median.

References

Bates, Douglas. 2013. Linear mixed model implementation in lme4. http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation & Change* 14: 261--90.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge, UK: Cambridge University Press.

Gelman, Andrew, & Jennifer Hill. 2007. *Data analysis using regression and multilevel / hierarchical models*. Cambridge, UK: Cambridge University Press.

Happé, Francesca. G. 1999. Autism: Cognitive deficit or cognitive style? *Trends in Cognitive Sciences* 3: 216--222.

Labov, William. 1981. Resolving the Neogrammarian controversy. *Language* 57: 267--308.

Labov, William. 1969. Contraction, deletion and inherent variability in the English copula. *Language* 45: 715--62.

Osthoff, Hermann, & Karl Brugmann. 1878. *Morphoplogische Untersuchungen auf den Gebiete der indogermanischen Sprachen*. I. Leipzig.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In: Gussenhoven and Warner (eds.), *Laboratory Phonology* 7, 101--40. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee and Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137--57. Amsterdam: John Benjamins.

Raymond, William D., and Esther L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In: Gries and Divjak (eds.), *Frequency effects in language learning and processing*, 35--52. Berlin: Mouton de Gruyter.

Schuchardt, Hugo. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.

Yaeger-Dror, Malcah, and William Kemp. 1992. Lexical classes in Montreal French. *Language & Speech* 35: 251--93.

Yu, Alan C. L. 2010. Perceptual compensation for coarticulation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLOS One* 5(8): e11950.

Developing methods for the study of social emotions: *SHAME* in British and American English

Karolina Krawczak
Faculty of English, Adam Mickiewicz University, Poznań
Institut de langue et littérature anglaises, Université de Neuchâtel
karolina@wa.amu.edu.pl

1 Introduction

The present study is a corpus-based quantitative analysis of SHAME from a comparative perspective. It reveals how this socially rich **emotion concept** is construed in British English and in American English. The concept is operationalized through three lexemes instantiating it – *ashamed*, *embarrassed*, and *humiliated*. Their actual usage is examined along formal, semantic and sociolinguistic parameters. The study employs a multifactorial usage-feature analysis (Geeraerts *et al*. 1994, Heylen 2005, Gries 2003, 2006, Grondelaers *et al*. 2007, Glynn 2009, 2010, Divjak 2010). Accordingly, it is assumed that patterns of language use are indicative of an underlying conceptual and socio-cultural structure.

The analysis of actual language use across many communicative situations offers an insight into the **social** dimension of language and cognition. SHAME, originating from the subject's sensitivity to others' actual or potential criticism, is a truly social emotion (Wierzbicka 1992, 1999). It integrates the speaker's internal and external perspectives on a given situation. Wierzbicka's results serve as a basis of this usage-based study, upon which the intersubjective facets of the concept are elucidated.

2 Hypotheses

It is hypothesized, in accordance with Wierzbicka's (1992, 1999) introspection-based research, that *ashamed* will be linked to more serious atemporal causes originating in the experiencer's own properties or actions. *Embarrassed* and *humiliated*, on the other hand, are expected to be correlated more distinctly with the here and now of the situation engendering the emotion and with causes of lesser magnitude, i.e., causes whose effect on the subject is much more ephemeral. It is also proposed that *humiliated*, in particular, will be related to purely external causes. Lastly, one would expect cross-dialectal differences to be revealed with respect to the causes of the emotion given the differences between the two cultures.

3 Methodology

The method of corpus-driven semantic analysis entails the meticulous **manual annotation** for usage-features of large numbers of examples. In this study, 200 occurrences per lexeme/dialect, accompanied by extended context, were extracted from the British National Corpus and the Corpus of Contemporary American English (Davies 2008-). It is likely that style and register impact substantially upon the representation of SHAME emotions. For this reason, sociolinguistic variation is controlled for by restricting the sample to the fiction component of the corpora. In total, 1200 examples are tagged for a range of

formal and **semantic** variables, some of which are enumerated in Table 1. The features of the most critical factor regarding the cause of the emotion were based on prior research by Kövecses (1986, 1990), Tissari (2006), Fabiszak *et al.* (2007), and Krawczak (in press). This category is subjective in nature, representing an inherent limitation to the reliability of the results. However, following Glynn (2010), the study assumes that, with due care, subjective factors can be included in quantitative conceptual analysis. The current study does not employ tests for inter-rater agreement, but the subjective categories are operationalized with a list of 'test' questions. The feature-analysis reveals the conceptual structure of the lexemes, their onomasiological interrelations, as well as distinctive lectal profiles. The data are treated with **multivariate exploratory** and **confirmatory statistics** in the form of Correspondence Analysis (Glynn 2013) and Polytomous Logistic Regression (Arppe 2008), respectively. These methods enable the author to identify falsifiable patterns of language use and the intersubjective conceptual profiles of SHAME emerging in the two communities.

4 Results

The exploratory analysis presented in Figure 1 reveals three distinct clusters, each corresponding to one of the three lexemes relative to dialect. The visualization in the plot is reliable, accurately depicting 86% of the variation explained by the first two dimensions. The lexeme ***ashamed*** is shown to be closely associated with the violation of the social norm of emotional reaction, dubious social status, bodily causes, and with causes of an atemporal and internal (i.e., coming from within the experiencer) nature (e.g., examples 1, 2, 3). ***Embarrassed*** corresponds distinctly to personal insecurities, violation of the social norms of politeness and decency, inadequacy and causes that relate to the present situation (e.g., examples 4 and 5). Finally, ***humiliated*** forms a clear semantic profile relative to causes designating rejection and mistreatment (e.g., 6, 7, 8). The factor of dialect does not seem to contribute much to the structuring of the data, as irrespective of the language variant, the lexemes realizing the given emotion cluster together relative to the causes of the emotion, the temporal referential scope of the cause and its type.

However, the confirmatory method of Polytomous Logistic Regression (Table 2) demonstrates that there are subtle cross-dialectal differences between the lexemes. The goodness of fit of the model can be determined on the basis of the pseudo R^2 scores, with the McFadden R^2 at 0.2707 and the Nagelkerke R^2 at 0.6351505. The rule of thumb is that respectively 0.2 (McFadden 1979: 307) and 0.3 (Lattin *et al*. 2003: 486) are good values. Further diganostics show that the model has problems obtaining adequate precision and accuracy for US *embarrassed* and US *ashamed*. This susggests that these lexemes are less distinctive and may indicate that the US use of the lexemes lies between that of the two UK lexemes. Overall, the analysis shows that British occurrences of *ashamed* are predicted most clearly by atemporal and internal causes, as well as the shared type of the emotion. American uses of *ashamed* are dissociated from present causes. *Embarrassed*, for both British and American uses, is related to present causes, the difference being that the UK occurrences are disassociated from non-intentionality, while US *embarrassed* is predicted by the shared type of the emotion. Finally, *humiliated* is shown to be related to causes concerning mistreatment and rejection in British English, whereas the American occurrences are disassociated from the following factors: general and internal causes, dubious social status, inadequacy, violation of the norm of decency, financial causes, rejection, and shared type of the emotion. Some of these significant associations are illustrated in examples (1) to (8).

The present study confirms the hypotheses formulated above. It shows that SHAME corresponds more readily to morally underpinned causes (e.g., dubious social status), inherent qualities (emotional weaknesses, as revealed in violation of the norms concerning emotional reactions), and intentional behaviors. EMBARRASSMENT and HUMILIATION, in turn, are more importantly linked to the immediate context in which the emotion is experienced (the present time) and, therefore, to ephemeral causes.

Table 1. Annotation schema

Formal Factors	Features	Socio-semantic Factors	Features
Adjective Type	Attributive, Predicative	Emotion Cause	Bodily, Dubious Social Status, Failure, Financial, Inadequacy, Insecurity, Mistreatment, Norm Violation Decency, Norm Violation Emotional Reaction, Norm Violation Politeness, Rejection, Unprestigious Status
Experiencer Grammatical Pers.	1st Pers, 2nd Pers, 3rd Pers	Cause Intentionality	Intentional, Unintentional, Non-intentional
Sentence Tense	Present, Past, Future	Cause Temporal Scope	General, Present, Past
Sentence Modality	Qualified, Unqualified	Cause Type	Internal, External
		Emotion Type	Individual, Shared, Projected

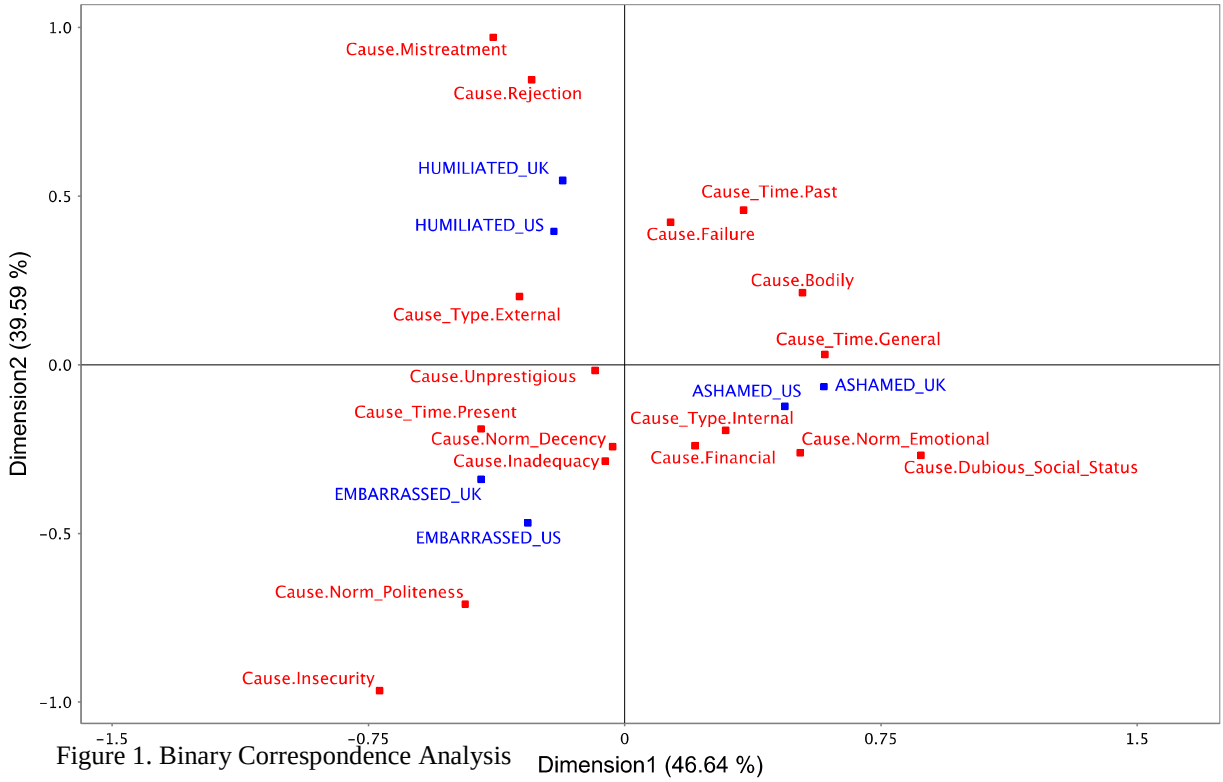

Figure 1. Binary Correspondence Analysis

Table 2. Polytomous Logistic Regression Analysis

Formula: Lexeme_Dialect ~ Cause_Type + Cause + Cause_Time + Emotion_Type + Intentionality

Log-odds:	UK ashamed	US ashamed	UK embarrassed	US embarrassed	UK humiliated	US humiliated
Cause: General	**0.9736**	(-0.541)	(0.9496)	(1.218)	(-0.4374)	**-1.218**
Cause: Present	(-0.5949)	**-1.933**	**2.284**	**2.673**	(-0.609)	(0.03522)
Cause: Internal	**1.694**	(0.4126)	(-0.6502)	(1.057)	(-0.6264)	**-1.513**
Dubious Social Status	(0.5057)	(-0.09135)	(0.9481)	(15.22)	(-0.5309)	**-2.586**
Failure	(0.1035)	(-1.718)	(-0.3613)	(16.76)	(1.637)	(-1.394)
Inadequacy	(-0.8705)	(0.2159)	(0.2438)	(17.14)	(1.117)	**-1.812**
Insecurity	(-0.5136)	(-0.3594)	(1.928)	(17.59)	(-14.93)	(-20.48)
Norm Violation: Decency	(-0.8895)	(0.5389)	(0.9244)	(16.2)	(1.303)	**-2.683**
Norm Violation: Emotional	(0.2599)	(0.01384)	(0.4214)	(16.57)	(0.9849)	(-19.15)
Norm Violation: Politeness	(-1.603)	(0.1211)	(1.393)	(16.94)	(0.7326)	(-19.06)
Status Loss: Financial	(-0.2392)	(-0.02772)	(-1.042)	(17.77)	(1.219)	**-2.722**
Status Loss: Mistreat	(-1.228)	(-1.285)	(-0.6556)	(0.3125)	**2.84**	(-1.601)
Status Loss: Rejection	(0.1668)	(-15.98)	(-0.5701)	(16.31)	**2.7**	**-2.14**
Status Loss: Unprestigious	(-1.285)	(-0.02888)	(0.9112)	(16.41)	(0.2976)	(-0.3649)
Emotion Shared	**1.549**	(-0.3072)	(0.01025)	**1.228**	(-0.6073)	**-1.892**
Cause: NonIntentional	(-0.6519)	(0.4503)	**-1.154**	(-0.1205)	(0.795)	(0.734)
Cause: Unintentional	(0.2911)	(-0.08921)	(0.395)	(-0.7433)	(-0.2246)	(0.5238)

R2.McFadden: 0.2707
R2.Nagelkerke: 0.6351505

Examples

(1) *"You said whatever you had to say to get elected. When I recall your memories, I feel **ashamed**." She looked directly at him. " Don't you?"* (*ashamed*: Dubious Social Status & Emotion: Shared)

(2) *I'm surprised to see tears glisten in the girl's eyes and she quickly turns away, as though **ashamed** to reveal weakness.* (*ashamed*: Social Norm Violation: Emotional)

(3) *They knew why Vice stopped playing Little League -- because the one time his drunk-ass father showed up to a game, he nearly beat the kid senseless for striking out. They knew why Vice always played shirts in pickup hoops -- because he was **ashamed** of the scars that the belt had left on his back.* (*ashamed:* General Time of the Cause & Internal Cause)

(4) *And for all his bulk, he was gentle in that olive tree bed, quiet, shy, **embarrassed** by his own needs.* (*embarrassed*: Insecurity, Present Time)

(5) *"Michael, you know this isn't how we handle the diabolically challenged now." Michael gaped at his friend. "What the hell kind of jive talk is that? That" he pointed with his sword toward the headless heap "is a demon." Gabriel barely glanced in that direc-*

*tion, then said in a low voice, almost as if he was **embarrassed**, "Don't say that. We don't call them that anymore." "Don't say demon? I sure as hell am not saying' diabolically challenged.' That's retarded." "No," Gabriel said. "Don't say' jive talking.' No one says that anymore. And you really shouldn't say' retarded' either. It's not PC."* (*embarrassed*: Social Norm Violation: Politeness & Emotion: Shared)

(6)*'They treat you like -- like shit,' Gloria said. 'I've never been so **humiliated**.'* (*humiliated*: Social Status Loss: Mistreatment)

(7) *'I think he feels **humiliated** as much as grief-stricken. For an Arab to be refused by his amour....'* (*humiliated*: Social Status Loss: Rejection)

(8) *She'd felt too **humiliated** about being heavy to buy maternity clothing when she was pregnant*. (*humiliated*: Non-Intentional Cause)

References

Arppe, Antti. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonymy*. [PhD dissertation]. Helsinki: University of Helsinki.

Davies, Mark. 2004-. *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). <http://corpus.byu.edu/bnc/>

Davies, Mark 2008-. *The Corpus of Contemporary American English: 450 million words, 1990-present*. <http://corpus.byu.edu/coca/>

Divjak, Dagmar. 2010. *Structuring the lexicon: A clustered model for near-synonymy*. Berlin: Mouton de Gruyter.

Fabiszak, Małgorzata, and Anna Hebda. 2007. Emotions of control in Old English: Shame and guilt. *Poetica* 66: 1-35.

Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema. 1994. *The Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2009. Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics, in: Evans and Pourcel (eds.): *New Directions in Cognitive Linguistics*, 77-106. Amsterdam: John Benjamins.

Glynn, Dylan. 2010. Testing the hypothesis: Objectivity and verification in usage-based Cognitive Semantics. In: Glynn and Fischer (eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches*, 239–269. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2013. Correspondence Analysis. An exploratory technique for identifying usage patterns. In: Glynn and Robinson (eds.), *Polysemy and synonymy. Corpus methods for cognitive semantics*. Amsterdam: John Benjamins.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of Particle Placement*. London: Continuum Press.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of to run. In: Gries and Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57-99. Berlin: Mouton de Gruyter.

Grondelaers, Stefan, Dirk Geeraerts, and Dirk Speelman. 2007. A case for a Cognitive corpus linguistics. In: González-Márquez, Mittelberg, Coulson and Spivey (eds.), *Methods in Cognitive Linguistics*, 149-169. Amsterdam: John Benjamins.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In: Kepser and Reis (eds.), *Linguistic Evidence: Empirical, theoretical and computational perspectives*, 241-264. Berlin: Mouton de Gruyter.

Kövecses, Zoltán. 1986. *Metaphors of Anger, Pride and Love. A lexical approach to the structure of concepts*. Amsterdam: John Benjamins.

Kövecses, Zoltán. 1990. *Emotion Concepts*. New York: Springer.

Krawczak, Karolina. In press. Shame and its near-synonyms in English: A multivariate corpus-driven approach to social emotions. In: Novakova, Blumenthal and Siepmann (eds.), *New Directions in Lexical Semantics and Discourse Organization*. Frankfurt am Main: Peter Lang.

Lattin, James M., Douglas J. Carrol, and Paul Edgar Green. 2003. *Analyzing Multivariate Data*. Pacific Grove, CA: Thomson Brooks.

McFadden, Daniel. 1979. Quantitative methods for analyzing travel behavior of individuals: Some recent developments. In: Hensher and Stopher (eds.), *Behavioral travel* modeling, 279-318. London: Croom Helm.

Tissari, Heli. 2006. Conceptualizing shame: Investigating uses of the English word shame, 1418-1991. In: McConchie, Timofeeva, Tissari, and Säily (eds.), *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis*, 143-154. Somerville, MA: Cascadilla Proceedings Project.

Wierzbicka, Anna. 1992. Defining emotion concept. *Cognitive Science* 16: 23-69.

Wierzbicka, Anna. 1999. Emotional universals. *Language Design* 2: 23-69.

The Relationship between Form and Meaning: Modelling Semantic Densities of English Monomorphemic Verbs

Aki–Juhani Kyröläinen[1] Kristina Geeraert[2]
[1]University of Turku [2]University of Alberta
akkyro@gmail.com geeraert@ualberta.ca

1 Background

Most words in language tend to be associated with multiple meanings. The relationship between these meanings has long been debated in theoretical linguistics (cf. Apresjan, 1974; Geeraerts, 1993; Tuggy, 1993; Gries, 2006). Polysemy is often regarded as one form having related senses, such as Lakoff's (1987) discussion of *over*, where *over* can mean 'above' as in *the picture is over the sofa* but 'across' in *he drove over the bridge*. In contrast, homonymy is characterized as two separate forms having unrelated meanings, as in *bank* 'financial institution' and *bank* 'edge of a river'. However, little is known about the factors which influence semantic expansion, or how semantic densities (i.e. sense and meaning relations) are stored or processed in the mental lexicon. We investigate two questions in this study: first, what set of factors influence the modulation of semantic densities; and second, how do these factors relate to semantic processing in lexical decision and naming tasks.

Numerous approaches have been proposed to account for semantic densities. In usage-based models, the senses are assumed to be extended from a prototypical meaning (Lakoff, 1987); while in formal approaches, these are represented by rules and decomposition (Pustejovsky, 1995). Other approaches propose models containing single and complex semantic representations (Ruhl, 1989), or where the senses form separate representations sharing only the phonological form (Klein and Murphy, 2001). In contrast, recent experimental studies show that polysemous and homonymous forms are processed differently (cf. Beretta et al., 2005; Pylkkänen et al., 2006). Additionally, it has been shown that semantic density is positively correlated with frequency in German (Köhler, 1986) and the (ir)regularity of the verb in three Germanic languages (Baayen and Moscoso del Prado Martín, 2005). Finally, Shillcock et al. (2001) have shown that phonological and semantic distances are positively correlated. This study investigates correlations between semantic densities, phonological forms, and other linguistically influential variables based on these previous findings, using English monomorphemic verbs.

2 Methodology

We estimate the network structure of the lexicon using phonological neighbours, defined as a one-phoneme difference (e.g. *go* is neighbours with *show* and *tow*). The interaction between phonology and semantics has been identified in previous studies (cf. Shillcock et al., 2001; Bergen, 2004). Bergen, for example, shows that phonaestemes have a psychologically real semantic component: the *gl* in *glean, glow, glisten,* and *glitter* meaning 'light/vision', or the *sn* in *sneeze, snout, snot,* and *snore* meaning 'mouth/nose'.

We extracted approximately 2,300 monomorphemic verbs from the English Lexicon Project (Balota et al., 2007), along with their mean reaction times (RT) in a lexical decision (LD) and a naming (NMG) task. The number of synsets was extracted from WordNet (Miller et al., 1990) in order to estimate the semantic density of each word. We creat-

ed a phonological network in which each node is represented by a phonological form (cf. McClelland, 1987). Lastly, we utilize graph theory to measure the complexity and structure of this network (Vitevitch, 2008), specifically: Degree (i.e. the number of neighbours), Clustering Coefficient (i.e. whether the neighbours are neighbours), and Closeness (i.e. a measure of the average distances from one node to every other node).

To determine the influence of these variables, additional variables, already known to be correlated with semantic processing, were also included. First, Verb Type indicates the (ir)regularity of the verb. Second, we created a set of variables to account for possible ambiguities. One variable, Homonymy, is a binary variable indicating whether the meanings associated with a form are homonymous or not, as determined by the Oxford English Dictionary and Wordsmith. Two variables accounted for possible part-of-speech ambiguity: Verbiness, the ratio of the synsets as a verb relative to all synsets of the word; and Noun-to-Verb Ratio, the frequency of a word as a noun relative to the frequency of it as a verb, based on subtitle frequencies. Additionally, four variables were used to account for differences in usage: Frequency and Dispersion measured distributional differences; and Local and Global Semantic Neighbours measured differences in co-occurrences (cf. Durda and Buchanan, 2006). Finally, the Mean Bigram Frequency (i.e. orthographic letter) was included to control for orthography. We modelled the semantic densities, as well as the RTs, using generalized additive models (Wood, 2006). RTs were included to estimate the contribution of these variables in processing.

3 Results

The Semantic Density Model models the number of synsets for each word and has 65% of the deviance explained. The statistically significant smooth functions are summarized in Table 1, along with a Delta AIC value, which measures the importance of the variables in each model. After controlling for other predictors, the model shows a small but significant correlation between phonology and semantics, modulated by the interaction of Degree and Closeness. This suggests that a word has a more dense semantic structure when it has more phonological neighbours and a shorter average distance to all other nodes in the network.

The importance of these functions in processing was modelled with the RTs. The residuals from the Density Model were extracted and used as a new variable in these models to control for the influence of semantic density. We used the residuals of the Density Model in order to avoid collinearity with other predictors. The LD Model has 59% of the deviance explained, compared with 44% in the NMG Model. These models reveal that the majority of functions in Table 1 also influence processing, with Degree and Closeness both facilitating processing in these tasks.

In sum, these results suggest that semantic density is, at least partly, correlated with the structure of the mental lexicon (cf. Shillcock et al., 2001). Furthermore, the models indicate that phonological and semantic representations share general organizational principles. The phonological structure of a word connects it to other similar words in the network. Highly dense regions indicate an increased phonological regularity, such as *run* and *go*, in contrast to *wield*, and this strengthened connectivity facilitates both semantic expansions and processing. Thus, the network structure of the lexicon contains a complex interaction between rich representations of phonology and semantics.

	Density Model	LD Model	NMG Model
Function	Δ AIC	Δ AIC	Δ AIC
Frequency, Dispersion by Verb Type	728	416	145
Degree, Closeness	11	15	30
Local SN[a], Global SN by Homonymy	399	32	14
Clustering Coefficient, Verbiness	201	8	2
Mean Bigram Frequency	26	4	14
Noun-to-Verb Ratio	202	NA[b]	NA
Semantic Density, Noun-to-Verb Ratio	NA	22	3

Note. $\Delta < 2$: substantial evidence for variable exclusion; $\Delta > 3 < 7$: considerably less support; $\Delta > 10$: very unlikely

[a] SN = Semantic Neighbour. [b] NA = Not Available

Table 1: Estimated Smooth Functions and Variable Importance in the Models

References

Apresjan, Jurij D. 1974. Regular Polysemy. *Linguistics,* 142(12): 5–32.

Baayen, R. Harald, and Fermín Moscoso del Prado Martín. 2005. Semantic Density and Past-Tense Formation in Three Germanic Languages. *Language*, 81(3): 666–698.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39(3): 445–459.

Beretta, Alan, Robert Fiorentino, and David Poeppel. 2005. The Effects of Homonymy and Polysemy on Lexical Access: An MEG Study. *Cognitive Brain Research,* 24(1): 57–65.

Bergen, Benjamin K. 2004. The Psychological Reality of Phonaesthemes. *Language*, 80(2): 291–311.

Geeraerts, Dirk. 1993. Vagueness's Puzzles, Polysemy's Vagaries. *Cognitive Linguistics*, 4(3): 223–272.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many senses of 'to run'. In: Gries and Stefanowitsch (eds.), *Corpora in Cognitive Linguistics,* 57–99. Berlin: Mouton de Gruyter.

Durda, Kevin, and Lori Buchanan. 2006. *WordMine2 [Online]*. Available at: http://web2.uwindsor.ca/wordmine.

Klein, Devorah E., and Gregory L. Murphy. 2001. The Representation of Polysemous Words. *Journal of Memory and Language*, 45(2): 259–282.

Köhler, Reinhard. 1986. *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

McClelland, James L. 1987. The case for interactionism in language processing. In: Coltheart (ed.), *Attention and Performance XII: The Psychology of Reading,* 3–36. Hillsdale: Lawrence Erlbaum Associates.

Miller, George A., Richard Beckwith, Chistiane Fellbaum, Derek Gross, Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography,* 3(4): 235–244.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge: MIT Press.

Pylkkänen, Liina, Rodolfo Llinás, and Gregory L. Murphy. 2005. The Representation of Polysemy: MEG Evidence. *Journal of Cognitive Neuroscience,* 18(1): 97–109.

Ruhl, Charles. 1989. *On Monosemy: A Study in Linguistic Semantics*. Albany: State University of New York Press.

Shillcock, Richard, Simon Kirby, Scott McDonald, and Chris Brew. 2001. Filled pauses and their status in the mental lexicon. In: *Proceedings of the 2001 Conference of Disfluency in Spontaneous Speech*, 53–56.

Tuggy, David. 1993. Ambiguity, Polysemy, and Vagueness. *Cognitive Linguistics,* 4(3): 273–290.

Vitevitch, Michael S. 2008. What Can Graph Theory Tell Us about Word Learning and Lexical Retrieval? *Journal of Speech, Language, and Hearing Research,* 51(2): 408–422.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman and Hall/CRC Press.

Thematic Roles and Semantic Space: Insights from Distributional Semantic Models

Gabriella Lapesa
University of Osnabrück
glapesa@uos.de

Stefan Evert
FAU Erlangen Nürnberg
stefan.evert@fau.de

The goal of this work is to use Distributional Semantic Models (Sahlgren, 2006; Turney, 2010) to get insights into the nature of thematic roles. In particular, we investigate whether the semantic representation produced by Distributional Semantic Models (henceforth, DSMs) is sensitive to effects of typicality involving thematic roles, and we quantify their relative prominence in the semantic representation encoded in the distributional space. Corpus-based modeling of selectional preferences and thematic fit is a well established field of research (see Erk et al., 2007 and references therein). What is peculiar to our approach is its attempt to model thematic fit data without taking into account syntactic relations, on the basis of distributional relatedness in bag-of-words DSMs. In this abstract we will show that (a) DSMs that make no use of syntax show good performances in a task related to selectional preference and (b) that the distribution of DSMs' performance across thematic relations shows patterns which are compatible with some general assumptions in theoretical linguistics.

1 Data and Models

Our computational simulation is based on experimental items from two studies which investigate event knowledge effects in semantic priming:

- Ferretti et al. (2001) showed that verbs facilitate the processing of nouns denoting prototypical participants in the depicted event (in the role of AGENT, PATIENT, INSTRUMENT), and of adjectives denoting features of prototypical participants (PATIENT FEATURE). The thematic role LOCATION did not show any priming effect.
- McRae et al. (2005) showed that nouns facilitate processing of verbs denoting events in which they are prototypical participants (in the role of AGENT, PATIENT, INSTRUMENT, LOCATION).

The set of stimuli from these two studies constitutes the gold standard of our evaluation task. Table 1 reports the number of triples for every thematic relation in the dataset and one example triple for each relation.

DSMs are evaluated in a classification task: given a target (e.g., *interview*) and the corresponding pair of primes in the dataset (*reporter* and *carpenter*, for the thematic role AGENT), we measure DSMs' accuracy in picking up the congruent prime on the basis of semantic distance. We expect the vectors for prototypical thematic role fillers to be closer to the respective verbs than the non-prototypical ones; in parallel we expect verbs to be closer to their prototypical fillers than to non-prototypical ones. Verbs and prototypical fillers co-occur, therefore, they occur in similar contexts. The reason why we expect verbs and prototypical fillers to be found closer in the semantic space is the presence of a shared a topic, namely, the event.

Our research consists of a large-scale evaluation of DSMs and their parameters. Evaluated parameters are:

Dataset	Relation	N	Prime$_c$	Prime$_i$	Target
Verb-Noun	AGENT	28	Pay	Govern	Customer
	PATIENT	18	Invite	Arrest	Guest
	PATIENT FEATURE	20	Comfort	Hire	Upset
	INSTRUMENT	26	Cut	Dust	Rag
	LOCATION	24	Confess	Dance	Court
Noun-Verb	AGENT	30	Reporter	Carpenter	Interview
	PATIENT	30	Bottle	Ball	Recycle
	INSTRUMENT	32	Chainsaw	Detergent	Cut
	LOCATION	24	Beach	Pub	Tan

Table 1: Experimental datasets from Ferretti et al. (2001) and McRae et al. (2005): number of pairs per relation and example stimuli

- Source corpus: BNC, WaCkypedia_EN, Wp500[1], UkWaC, and a combination of BNC, Wackypedia_EN, and UkWaC;
- Context window: 2, 5 and 15 words to the left and to the right of the target;
- Use of part-of-speech information: no part of speech information, part of speech information on the target, part of speech information on both targets and features;
- Scoring measure: frequency, simple log-likelihood, Mutual Information, t-score, z-score, Dice coefficient;
- Vector transformation: no transformation, logarithmic, sigmoid and root transformation;
- Distance measure: cosine, euclidean and manhattan distance;
- Dimensionality reduction: no dimensionality reduction, Random Indexing (1000 dimensions) and Randomized Singular Value decomposition (300 dimensions);
- Index of distributional relatedness: distance in the semantic space, backward rank (rank of prime in the neighbors of the target), forward rank (rank of target in the neighbors of the prime), average of backward and forward rank[2].

In total, 38,880 models were computed for all possible combinations of these parameters.

2 Results

Table 2 shows range and mean accuracy achieved by the DSMs for each thematic role. Results are reported for two indexes of distributional relatedness, namely *distance* and *forward rank* (position of the target in the ranked neighbors of the prime).

Dataset	Relation	Distance		Forward rank	
		Range	M	Range	M
Verb-Noun	AGENT	43-100	79.3	39-100	85.6
	PATIENT	44-100	83.4	50-100	87.8
	PATIENT FEATURE	35-95	72	40-100	81.2
	INSTRUMENT	42-100	80.2	38-100	82.6
	LOCATION	30-96	73.6	42-100	82.9
Noun-Verb	AGENT	40-100	77.1	47-100	87.5
	PATIENT	47-100	85.6	60-100	93.6
	INSTRUMENT	40-100	75.4	47-100	87.6
	LOCATION	42-96	79.4	46-96	85.2

Table 2: Identification of consistent prime on the basis of distributional relatedness

[1] A subset of WaCkypedia_EN, composed by the first 500 words of each article.

[2] The introduction and evaluation of this parameter has many implication for cognitive modeling, as rank can capture directionality in priming effects. We will not tackle this issue in this abstract for reasons of space.

First of all, the high performance achieved on all the relations shows that selectional preference is indeed a matter of topic. Even if we are not claiming that syntax does not play any role in selectional preference, the fact that such a high performance is achieved without using syntactic information suggests that verbs and their prototypical fillers can be interpreted as cues to event knowledge (for a review of this claim and of its consequences for lexical theories, see Elman, 2009).

A comparison between mean accuracies allows to rank thematic relations with respect to the robustness of the typicality effects shown by DSMs. The results of such ranking of thematic roles for the two datasets are:

- Ferretti et al. (2001)
  - Distance: PATIENT>INSTRUMENT>AGENT>LOCATION>PATIENT FEATURE
  - Forward rank: PATIENT>AGENT>LOCATION>INSTRUMENT>PATIENT FEATURE
- McRae et al. (2005)
  - Distance: PATIENT>LOCATION>AGENT>INSTRUMENT
  - Forward rank: PATIENT>INSTRUMENT>AGENT>LOCATION

These rankings reflect the relative saliency of thematic roles as event features in the DSM semantic space. They may also be interpreted in the light of distinctions commonly assumed in theoretical linguistics, namely between *arguments* and *adjuncts* and between the *internal* and *external* argument. In particular, the ranking PATIENT>AGENT>LOCATION>INSTRUMENT reported above may be mapped onto the scale of the syntactic proximity to the verb, the internal argument (e.g., THEME or PATIENT) being the closest to it, followed by the external argument (e.g., AGENT or CAUSE) and adjuncts (e.g., INSTRUMENT or LOCATION).

We display some of the best performing models for each syntactic relation in table 3 (index of distributional relatedness: forward rank). For each relation, we report the number of models that achieved the best accuracy and we specify one of the best models: this choice should not be considered representative of general trends of performance. Such trends need to be evaluated with a different type of analysis, given the high number of parameter combinations involved in our study.

Dataset	Relation	Best Model								
		Acc	N	Corpus	Win	Pos	Score	Trans	Dist	Dim.Red
V-N	AGENT	100	11	wacky	5	targ+feat	freq	none	man	ri
	PATIENT	100	825	wacky	15	target	MI	none	cos	ri
	PATIENT FEAT.	100	4	wacky	5	targ+feat	freq	none	cos	ri
	INSTRUMENT	100	22	joint	15	targ	freq	none	man	rsvd
	LOCATION	100	127	joint	15	no pos	t-sc	log	cos	none
N-V	AGENT	100	643	bnc	15	none	s-ll	none	euc	none
	PATIENT	100	2357	ukwac	5	targ+feat	freq	log	cos	none
	INSTRUMENT	100	302	ukwac	15	none	freq	none	cos	rsvd
	LOCATION	95.8	504	joint	15	none	s-ll	none	cos	none

Table 3: Identification of consistent prime on the basis of distributional relatedness, forward rank: best accuracy (*Acc*), number of models that achieved best accuracy (*N*), the set of parameters defining one of the best models

3 What we will present

This abstract sketches the general features of our study. In the presentation we will provide further details concerning the analysis of distribution of accuracy per thematic relation, and we will present an evaluation of the impact of the different parameters on the performance of the models.

References

Elman, Jeff L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547-582.

Ferretti, Todd, Ken McRae and Anne Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4), 516-547.

Erk, Katrin, Sebastian Padó and Ulrike Padó. 2007. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4), 723- 763.

McRae, Ken, Mary Hare, Jeff L. Elman and Todd Ferretti. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7), 1174-1184.

Sahlgren, Magnus. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high- dimensional vector spaces*. Unpublished doctoral dissertation, University of Stockolm.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141-188.

We don't all *think* exactly alike: empirical evidence for cross-linguistic lexical contrast

Mildred Lau[1,2] Antti Arppe[2]
[1]University of Eastern Finland [2]University of Alberta
{mmlau, arppe}@ualberta.ca

Language learners as well as translators frequently consult dictionaries for both purposes of stylistic variation and finding the best word to express a particular concept. However, dictionaries and thesauri rarely give explicit information on which synonym is most appropriate for a given context (Partington 1998: 29). For example, Table 1 shows part of one such English translation definition for the Finnish verb *ajatella*, representing various word senses.

ajatella	
1.	think, cogitate
2.	(pohtia) reflect/meditate (on), ponder, turn over in the mind, contemplate
3.	(käyttää aivoja) use your mind/wits, apply the mind
4.	(harkita) deliberate
5.	(hautoa) dwell on, brood on/over

Table 1: First five senses from translation dictionary definition of *ajatella* (Rekiaro & Robinson 2005)

How in practice does a native Finnish-speaking learner of English come to learn when and which of the provided synonyms is an appropriate translation for *ajatella*? According to the distributional hypothesis (Harris 1954), any difference in meaning, including sense, is accompanied by differences in context. Therefore, learning to differentiate between when to use each synonym is a matter of learning the distributional characteristics of each synonym. From prior studies we know that these distributional characteristics, as evident in natural linguistic usage, are quite extensive and complex (e.g. Divjak and Gries 2006; Arppe 2008). Importantly, we also have experimental evidence that native speakers are aware of these distributional characteristics in a consistent manner, even though they might not be able to accurately articulate them (e.g. Arppe and Järvikivi 2007)—but what about language learners at various degrees of proficiency? In order to study how such differences in contextual associations between translation equivalents are acquired, the first step is to comprehensively analyze and make explicit the contextual features pertaining to a set of possible translation equivalents in two languages, which has become possible with comprehensive usage data available in native speaker corpora.

As an example case, we contrast near-synonymous verbs for *think* in two genetically unrelated languages, namely English and Finnish. The English verbs selected for study, THINK, CONSIDER, REFLECT, and PONDER, were chosen by examination of lexicographic descriptions of verbs expressing continued thought in the *Oxford English Dictionary Online* (2013), in particular the synonymous verbs in each relevant definition. The four selected verbs cover a similar range of meaning, based on pairwise similarities between their definitions, and they each have more than 500 occurrences in the 100 million-word British National Corpus. The English analysis involved 1024 corpus lines randomly selected, near-equally distributed among the four verbs (Lau 2013). The Finnish verbs AJATELLA, MIETTIÄ, POHTIA, and HARKITA were selected based on their synonyms and usage examples given in the authoritative dictionary *Suomen kielen perussanakirja* (Haarala et al. 1997), and the research corpus comprised 3.7 million words from daily newspapers and Internet discussion groups, resulting in a total of 3404 instances of the four Finnish verbs in proportion of their frequency within the corpus (Arppe 2008).

Each sentence in the two research corpora was annotated for morphosyntactic, semantic, and functional features of the thinking verb therein and its arguments. Morphosyntactic features annotated include verbal inflection, the grammatical person of the thinker and types of nominal clauses as arguments; semantic features were applied to nominal arguments and adjuncts to the verb. Nominal arguments were semantically categorized using the 25 headers of lexicographer files in WordNet as a starting point. The distributions of the occurrences of the various contextual feature classes among the thinking verbs per the two languages were cross-tabulated in the R statistical computing environment (R Development Core Team 2013), and subsequently scrutinized using relatively simple but nevertheless very informative univariate statistical methods, namely the χ^2 statistic with standardized Pearson residuals (Arppe 2008, 2012), in order to determine which features display statistically significant differences either in favor of, or against, the occurrence of the thinking verbs. Tables 2 and 3 illustrate the preferences and dispreferences of the semantic subclassifications pertaining to the agent performing cognition for the four thinking verbs in English and Finnish, respectively, based on the research corpora. Importantly, no two verbs display identical profiles for preferred and dispreferred agent types, neither within nor between the languages.

Lexeme \ Agent	**None**	**Individual**	**Group**	**Generic**	**Other**
THINK	0	+	–	–	0
CONSIDER	+	–	+	0	0
REFLECT	–	+	0	+	0
PONDER	–	+	0	0	0

Table 2: Distribution of the semantic types of agents among the studied English *think* verbs
(Notation: +: positive coefficient; –: negative coefficient; 0: coefficient not statistically significant)

Lexeme \ Agent	**None**	**Individual**	**Group**	**Other**
AJATELLA	0	+	–	0
MIETTIÄ	–	+	–	0
POHTIA	+	–	+	+
HARKITA	+	–	+	0

Table 3: Distribution of the semantic types of agents among the studied Finnish *think* verbs
(Notation: +: positive coefficient; –: negative coefficient; 0: coefficient not statistically significant)

In a similar fashion, we can scrutinize and contrast all of the types of contextual features and their associations with the thinking verbs in both languages, for which the positive preferences are cross-tabulated in Table 4. Crucially, it is apparent that none of the verbs match one-to-one between English and Finnish: indeed, the positive associations for any one verb are mapped across multiple verbs in the other language, although the "best" match is between THINK and MIETTIÄ, with six positive associations in common. Both CONSIDER and POHTIA are marked in that they are the only studied verb in their respective languages with characteristic associations of group agents and the passive voice.

Finnish \ English	THINK	CONSIDER	REFLECT	PONDER
AJATELLA	Agent: Individual Agent: 1st person Adjunct: Negation	Agent: 2nd person Adjunct: Frame	Agent: Individual Theme: *that*-clause Adjunct: Manner	Agent: Individual
MIETTIÄ	Agent: Individual Agent: 1st person Agent: 2nd person Adjunct: Frequency Mood: Imperative Function: Quotative	Agent: 2nd person Theme: Communication Mood: Imperative	Agent: Individual Function: Quotative	Adjunct: Duration
POHTIA	–	Agent: Group Voice: Passive	Agent: 3rd person Theme: Attribute	Theme: Cognition
HARKITA	Adjunct: Frequency	Theme: Act	–	–

Table 4: Common positive contextual associations between the selected English and Finnish *think* verbs

Overall, the possible contextual particulars associated with thinking are distributed differentially among synonym sets within individual languages. That is, speakers of different languages compartmentalize the action of thinking in distinct ways that become evident in usage data. Achieving communicative competence in a new language would involve learning to re-map the various contextual associations among the synonym set in the first language to the equivalent set in the new language.

The present work demonstrates the complexity and variation in the way languages use synonyms to articulate subtly different presentations of a situation or event (Divjak and Gries 2006), in this case, thinking. Comprehensively understanding the contextual relationships among and between a set of synonymous words in these two languages is necessary for us to proceed to experimental studies with Finnish-speaking learners of English at various degrees of proficiency to evaluate the role of such contextual contrasts in second language acquisition. Ultimately, such research will add to our insight on how language is processed and represented in our minds.

References

Arppe, Antti. 2008. *Univariate, Bivariate and Multivariate Methods in Corpus-Based Lexicography: A Study of Synonymy*. (Publications of the Department of General Linguistics, University of Helsinki 44). PhD Dissertation, University of Helsinki.

Arppe, Antti. 2012. Package 'polytomous': Polytomous Logistic Regression for Fixed and Mixed Effects. Version 0.1.4. The R Project for Statistical Computing. URL: http://www.r-project.org/

Arppe, Antti and Juhani Järvikivi. 2007. Every Method Counts: Combining Corpus-based and Experimental Evidence in the Study of Synonymy. *Corpus Linguistics and Linguistic Theory* 3 (2): 131–159.

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Divjak, Dagmar and Stefan Th. Gries. 2006. Ways of Trying in Russian: Clustering Behavioral Profiles. *Corpus Linguistics and Linguistic Theory* 2 (1): 23–60.

Haarala, Risto and Marja Lehtinen (eds.). 1997. *CD-Perussanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 94. Helsinki: Edita.

Harris, Zellig. 1954. Distributional structure. *Word* 10 (23): 146–162.

Lau, Mildred. 2013. *Contextual Associations of Thinking Verbs: A Corpus-Based Investigation of English*. BA(Honors) Thesis, University of Alberta.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam & New York: John Benjamins.

Oxford English Dictionary Online. 2013. Oxford: Oxford University Press. URL: http://www.oed.com/ (accessed 16 March 2013).

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. URL: http://www.R-project.org/

Rekiaro, Ilkka and Douglas Robinson. 2005. *Suomi–Englanti–Suomi -Sanakirja* (4th ed.). Jyväskylä: Gummerus.

Exploring Word Order Universals: a Probabilistic Graphical Model Approach

Xia Lu
University at Buffalo
xialu@buffalo.edu

1 Introduction

Previous statistical methods in the research of word order universals have yielded interesting results but they have to make strong assumptions and do considerable amount of data preprocessing to make the data fit the statistical model (Greenberg, 1963; Hawkins, 1983; Dryer, 1989; Nichols, 1986; Justeson & Stephens, 1990). Recent studies using probabilistic models are much more flexible and can handle noise and uncertainty better (Daume & Campbell, 2007; Dunn et al., 2011). However these models still rely on strong theoretic assumptions and heavy data treatment, such as using only two values of word order pairs while discarding other values, purposefully selecting a subset of the languages to study, or selecting partial data with complete values. In this paper we introduce a novel approach to use a probabilistic graphical model to study word order universals.

2 Method

There are two advantages of using probabilistic graphical model to study word order universals. First the graphical structure can reveal much finer structure of language as a complex system. We assume there is a meta-language that has the universal properties of all languages in the world. We want a model that can represent this meta-language and make inferences about linguistic properties of new languages. This system is composed of multiple sub-systems such as phonology, morphology, syntax, etc. which correspond to the subfields in linguistics. In this paper we focus on the sub-system of word order only. The other advantage of PGM is that it enables us to quantify the relationships among word order features. A PGM model for word order subsystem encodes a joint probabilistic distribution of all word order feature pairs. Using probability we can describe the degree of confidence about the uncertain nature of word order correlations.

We choose DAG structure as our current model framework. Regarding the sampling problem we do not make any assumption about the i.i.d property of the language samples and propose two models: one is FLAT, which assumes samples are independent and identically distributed (i.i.d.); the other is UNIV, which takes care of the possible dependencies among the samples. By comparing the predictive power of these two models we hope to find one that is closer to the real distribution.

There are two big problems in learning DAG structure for the FLAT model. One is caused by large number of missing values. Because EM method for structures from incomplete data takes very long time to converge due to the large parameter space of our model, we decided to use imputation method to handle the missing data problem (Singh, 1997). The other difficulty is caused by limited data. To solve this problem we used model averaging by using bootstrap replicates (Friedman et al., 1999). To solve the problem of dependence among the languages in learning DAG structure for the UNIV model, we take an incremental and divide-and-conquer approach. Using clustering algorithm we identified five clusters in the WALS data. In each cluster we picked *1/n* of the data and combine them to make a subset. In this way we can have *n* subsets of data which have decreased degree of dependencies among the samples. We learn a structure for each subset and fuse the *n* graphs into one single graph.

The DAG structures for the two models are shown in Figure 1.

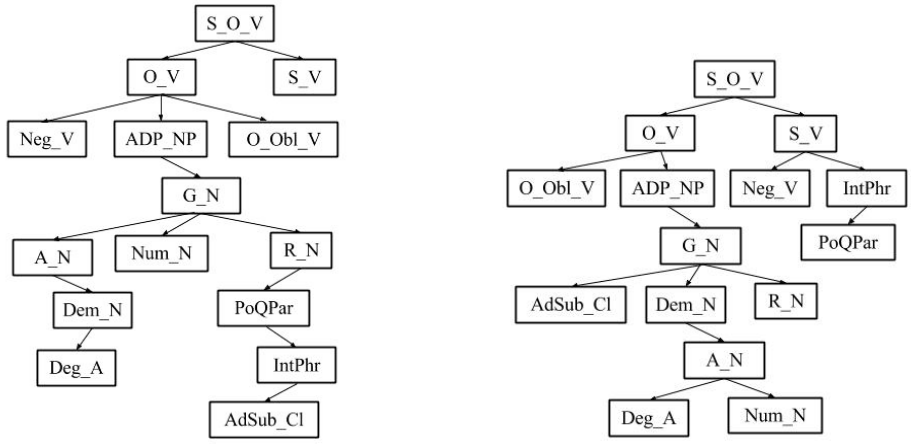

Figure 1. DAG structures of the two models (left: FLAT right: UNIV)

3 Quantitative Analysis of Results

The word order universal results are difficult to evaluate because we do not know the correct answers. Nonetheless we did a quantitative evaluation following Daumé III and Campbell (2007)'s method. The results are shown in Figure 3.

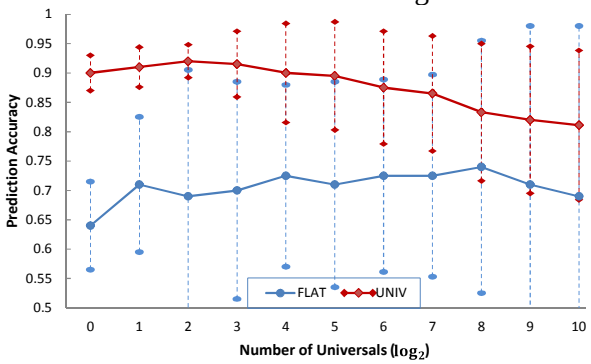

Figure 2. Results of Quantitative Evaluation

As we can see the predictive power of the UNIV model is much better than that of the FLAT model. The accuracy of our both models is lower than those of Daumé III and Campbell's. But this does not mean our models are worse considering the complexity in model learning. Instead our UNIV model shows steady accurate prediction for the top ten universals and has more stable performance compared with other models.

4 Qualitative Analysis of Results

We also did qualitative evaluation through comparison with the well-known findings in word order correlation studies: those of Greenberg's, Dryer's, and Daumé III and Campbell's. For Universal 2, 3, 4, 5, 10, 18 and 19, our results conform to Greenberg's (see Table1 in Appendix A). But for others there are discrepancies of different degrees. In comparison with Dryer (1992)'s work (see Table2 in Appendix A), we noticed in our results there is an asymmetry in terms of V_O's influence on other word order pairs, which was not discussed in previous work. In the correlated pairs, only ADP_NP and G_N show bidirectional correlation with O_V while PoQPar becomes a non-correlated pair. In the non-correlated pairs, Dem_N becomes a correlated pair and other pairs also show correlation of weak strength. Most of our results therefore do not confirm Dryer's findings.

We also compared the probabilities of single value pairs of the top ten word order universals with Daumé III and Campbell's results which are shown in the following figure:

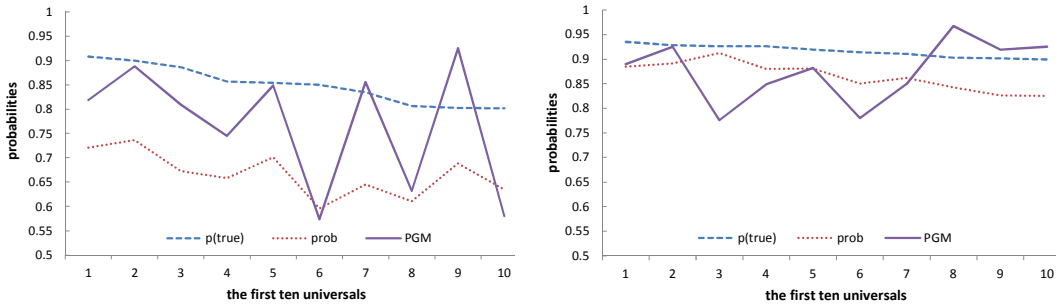

Figure 3. Compare with Daumé III and Campbell's models (left: HIER right: DIST)

It is hard to tell which model does a better job just by doing comparison like this. Daumé III and Campbell's model computes the probabilities of 3442 feature pairs separately. Their model with two values as nodes does not consider the more complex dependencies among more than two features. Our model provides a better solution by trying to maximize the joint probabilities of all word order feature pairs.

5 Inference

Besides discovering word order universals, our model can reveal more properties of word order sub-system through various inference queries such as inferring the probabilities of having each value of unobserved features given observed values; using MAP query to find the combination of values which has the highest probability when knowing one or more values of some features; and calculating the likelihood of a language in terms of word order properties.

6 Conclusion

Probabilistic graphic modeling provides solutions to the problems we noticed in the previous studies of word order universals. By modeling language as a complex system we shift our attention to the language itself instead of just features. Using PGM we can infer properties about a language given the known values and we can also infer the likelihood of a language given all the values. In the future if we include other domains, such as phonology, morphology and syntax, we will be able to discover more properties about language as a whole complex system.

References

Daumé, H., & Campbell, L. 2007. A Bayesian model for discovering typological implications. In *Annual Meeting –Association For Computational Linguistics* (Vol. 45, No. 1, p. 65).

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in Language 13*, 257 – 292.

Dryer, M. S. 1992. The Greenbergian word order correlations. *Language*, 81-138.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature 473*. 79–82.

Friedman, N., Nachman, I., & Peér, D. 1999. Learning bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 206-215). Morgan Kaufmann Publishers Inc.

Greenberg, J. H. 1963. *Some universals of grammar with particular reference to the order of meaningful elements*. In Universals of Language, J. H. Greenberg, Ed. MIT Press, Cambridge, MA, 73-113.

Hawkins, John A. 1983. *Word Order Universals.* Academic Press, 1983.

Justeson, J. S., & Stephens, L. D. 1990. Explanations for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists* (Vol. 3, pp. 2372-76).
Nichols, J. 1986. Head-marking and dependent-marking grammar.Language, 56-119.
Singh, M. 1997. Learning Bayesian networks from incomplete data. In *Proceedings of the National conference on Artificial Intelligence* (pp. 534-539). JOHN WILEY & SONS LTD.

Appendix: Comparison with others' work

Universals	**Dependencies**	**UNIV**
U2: ADP_NP<=>N_G	POST->GN PRE->NG GN->POST NG->PRE	83.59 70.29 78.45 81.91
U3: VSO->PRE	VSO->PRE	74.41
U4: SOV->POST	SOV->POST	85.28
U5: SOV&NG->NA	SOV&NG->NA	68.95
U9: PoQPar<=>ADP_NP	Initial->PRE Final->POST PRE->Initial POST->Final	41.87 49.67 15.80 31.73
U10: PoQPar<=> VSO	all values of PoQPar: VSO below 10%	below 10%
U11: IntPhr->VS	Initial->VS	24.12
U12: VSO->IntPhr	VSO->Initial SOV->Initial SOV->Not_Initial	50.54 28.52 60.41
U17: VSO->A_N	VSO->A_N	24.86
U18&19: A_N<=>Num_N<=>Dem_N	AN->NumN AN->DemN NA->NNum NA->NDem	68.86 73.74 61.74 61.00
U24: RN->POST (or AN)	RN->POST RN->AN	65.73 29.23

Table 1. Comparison with Greenberg's work

OV	**UNIV**	**VO**	**UNIV**
correlated pairs			
ADP_NP(POST)	90.48	ADP_NP(PRE)	82.72
G_N(GN)	79.38	G_N(NG)	61.49
R_N(RN)	19.66	R_N(NR)	75.17
PoQPar(Final)	31.89	PoQPar(Initial)	15.79
AdSub_Cl (Final)	20.90	AdSub_Cl (Initial)	49.22
IntPhr(Not_Initial)	58.74	IntPhr(Initial)	34.36
non-correlated pairs			
A_N(AN)	29.48	A_N(NA)	65.00
Dem_N(Dem_N)	52.27	Dem_N(N_Dem)	54.25
Num_N(NumN)	41.6	Num_N(NNum)	49.25
Deg_A(Deg_A)	43.48	Deg_A(A_Deg)	38.44
Neg_V(NegV)	48.06	Neg_V(VNeg)	25.13

Table 2. Comparison with Dryer's work

An exploratory approach to transitivizing morphemes in French

Nicolas Mazziotta Fabienne Martin
University of Stuttgart
nicolas.mazziotta@ulg.ac.be fabienne.martin@ling.uni-stuttgart.de

The transitivizing effect of affixes has been well documented for many Germanic languages (e.g. Lieber and Baayen 1993 for Dutch). For Romance languages like French, a few case studies suggest that some affixes also have a transitivizing effect, but they are all restricted to small subsets of the verbal lexicon (Junker, 1987; Aurnague and Plénat, 2007). In this contribution, we study all French affixes that have been claimed to display this effect, namely *a-*, *é-*, *en-*, *dé-*, *-iser*, *-ifier*, and take into account the entire database *Les Verbes Français* (Dubois and Dubois-Charlier 1997, henceforth "LVF"). In its XML version (henceforth "eLVF", see Hadouche and Lapalme 2010), the LVF describes the uses of 12.310 French verbs.

1 Methodology

1.1 Statistical technique

We make use of statistical exploratory techniques to extract major tendencies wrt affixation and transitivity (Lebart et al. 1998 for technical details). For French, these techniques have to our knowledge almost never been applied at the interface between morphology, syntax and semantics, and one of our goals is to show that they can be profitably used in this domain, by letting relevant generalizations and research questions emerge from a dataset that is too large to be processed intuitively. Our analyses use multiple correspondence analysis (henceforth "MCA"). MCA relies on a technique of decomposition of the information contained in the data into multiple dimensions. Focusing on the most important (information-wise) dimensions, MCA projects association tendencies contained in the data on a plane (showing simultaneously two dimensions). Each axis corresponds to a dimension and reveals an opposition. The further the points are located from the 0 coordinate, the stronger their opposition with points on the other side of this coordinate. Similar values agglomerate in a common area. The meaning of each opposition has to be interpreted according to the observed data.

For space reason, we focus on the most explicit MCA planes in this abstract.

1.2 Data preparation

The following ten variables are defined for each verbal entry (24.963 individuals, excluding auxiliaries and verbs without finite forms from the dataset): **HS** ∃ human/animal subject (0 = no, 1 = yes); **CL** Semantic class (14 values); **TS** ∃ a (direct) transitive construction (0, 1); **PS** ∃ a reflexive construction (0, 1); **NS** ∃ an indirect transitive construction (0, 1); **AS** ∃ an intransitive construction (0, 1); **CJ** infinitive inflection (1 = *-er*, 2 = regular *-ir*, 3 = other); **PXV** prefix (0 = none, *a-*, *dé1-*, *dé2-* non-productive, *é-*, *en-*); **SXV** suffix (0 = none, *-ifier*, *-iser*); **BPOS** base POS (0 = no derivation, A = adj., N = noun, V = verb).

Since the eLVF does not contain any morphological description except for the verbal inflection class, PXV, SXV and BPOS variables are output by DeriF Namer (2009), a morphological analyser. This provides a description of the morphological processes involved in

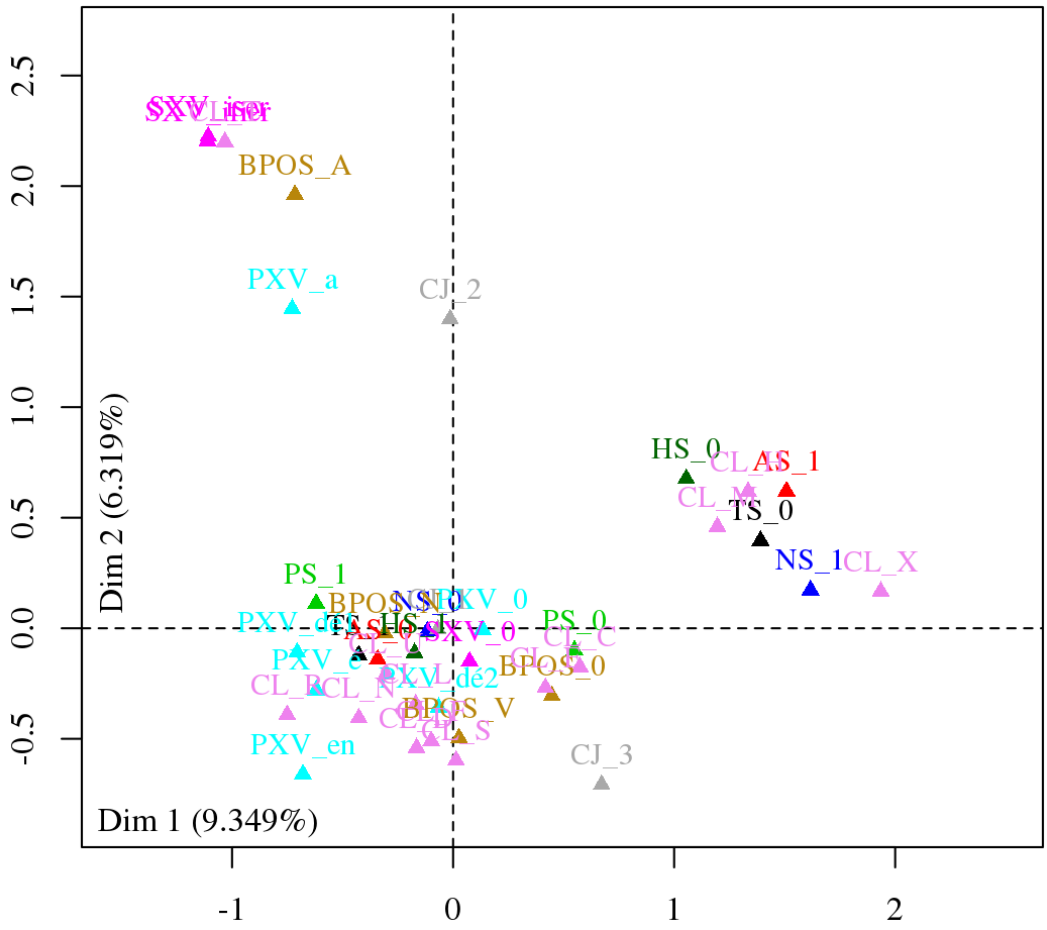

Figure 1: MCA on all verbs

the construction of the lexeme. We restricted the morphological analysis to the last step of the derivation process if the latter is complex.

2 Statistical Analyses

2.1 Transitivity and affixation

It can be easily observed that the horizontal axis of our MCA depicts a clear opposition (fig. 1): 1. affixes and deadjectival/denominal lexemes (variables are on the left side of the 0 coordinate); 2. TS_0, AS_1, NS_1 (cluster on the right).[1] This broadly suggests that affixation is correlated not to transitivity, but to the *absence* of intransitive or indirect constructions: it has a 'desintransitivizing' rather than a transitivizing effect.

The vertical axis parts deadjectival verbs that are change of state verbs (CL_T) from other verbs. Nearly all prefixes contrast with *-iser* and *-ifier* suffixes, but *a-* verbs of the *-ir* inflection class (CJ_2) pattern with these suffixes (they are located in the same area on the top of the plane). This probably reflects a stronger association between deadjectival verbs and *-ifier, -iser* and *a-*.

Removing deadjectival verbs from MCA does not change the plane or the contributions (not shown here). This suggests that the absence of intransitivity is primarily correlated with the use of affixes (rather than with the presence of an adjectival base), since affixed verbs tend to lack an intransitive reading even when not deadjectival.

[1] Contribution of all the variables on this axis are all very strong. Chi-square tests between each categories of these variables yield p-values under 10^{-10} for nearly each association.

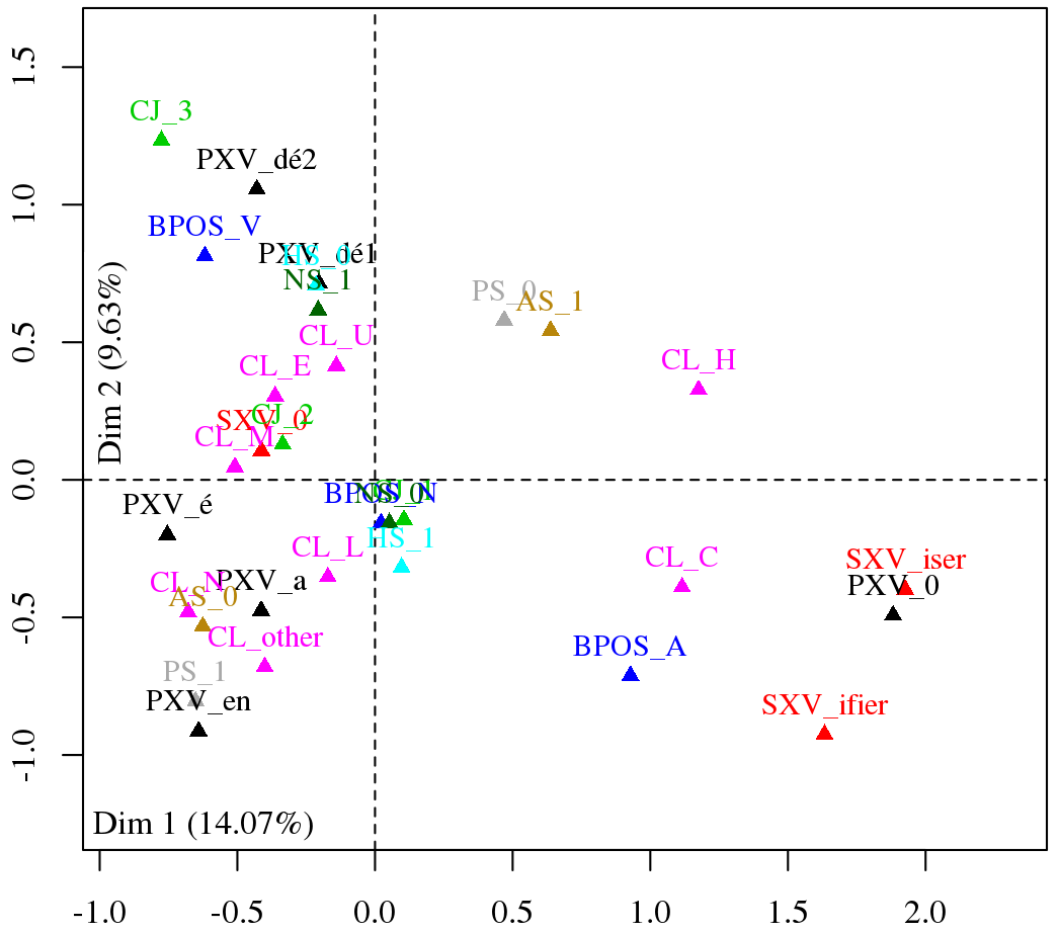

Figure 2: MCA on all non-transitive affixed verbs

2.2 Non-transitive affixed verbs

The MCA on non-transitive affixed verbs (fig. 2) shows how exceptions to the association between transitivity and affixation behave. Affixes are here associated with specific values of other variables. *Suffixes* appear in unergative behavior-related verbs (CL_H and CL_C), also known as performative/similative verbs, cf. e.g. *diplomatiser*. This contrasts with fig. 1, where suffixes are associated with unaccusative/anticausative verbs (CL_T). On the other hand, *prefixes* are associated with inanimate subjects (HS_0) as well as with movement verbs (CL_M, CL_E). Prefixes do not share a homogeneous behavior. *En-* and *a-* appear with regular verbs with a reflexive scheme (PS_1) and reject intransitivity (AS_1); this association does not hold for *dé1-* and *dé2-*, that are linked with a less interpretable cloud of variables.

3 Conclusion

These results lead to further discussion and research questions. Firstly, although the statistical analysis of the data confirms that affixation correlates with absence/loss of intransitivity (rather than presence/acquisition of transitivity), exceptions show that affixes cannot be described as mere transitivizers, as previously proposed by some authors. Additionally, the analyses reveal that affixes differ from each other wrt their association with the other variables taken into account (like the verbal class), but also that the degree of (dis)similarity between affixes varies with the construction used (transitive or intransitive): in intransitive constructions, *-ifier/-iser* verbs strongly differ from *é, a-* and *en-* verbs in that the former are often unergative while the latter are change of state verbs, but in transitive constructions, *a-*

is close to *-iser/-ifier* verbs.

Finally, the analysis also let emerge an interesting generalization about the two prefixes *dé1-/dé2-* (negative or intensive, cf. Namer 2009): they behave similarly in intransitive constructions but are clearly different in transitive ones, which might be related to the fact that the former is an external affix, while the latter is internal (Di Sciullo, 1997). This difference confirms that it is relevant to distinguish the two prefixes in the morphological decomposition, as in Namer 2009.

References

Aurnague, M. and M. Plénat (2007). *Contraintes sémantiques et dérivation en é- : attachement habituel, naturalité et dissociation intentionnelle*. Number 16 in Carnets de Grammaire.

Di Sciullo, A.-M. (1997). Prefixed-verbs and adjunct identification. In A.-M. Di Sciullo (Ed.), *Projections and Interface Conditions*, pp. 52–73. New-York/Oxford: Oxford University Press.

Dubois, J. and F. Dubois-Charlier (1997). *Les verbes français*. Larousse-Bordas.

Hadouche, F. and G. Lapalme (2010, dec). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages 10*(179-180), 193–220.

Junker, M.-O. (1987). Transitive, intransitive and reflexive uses of adjectival verbs in french. In J.-P. Montreuil (Ed.), *Advances in Romance Linguistics*, pp. 189–199. Foris Publications.

Lebart, L., A. Morineau, and M. Piron (1998). *Statistique exploratoire multidimensionnelle*. Paris: Dunod.

Lieber, R. and H. Baayen (1993). Verbal prefixes in Dutch: A study in lexical conceptual structure. In G. Booij and J. van Marle (Eds.), *Yearbook of Morphology*, pp. 51–78. Dordrecht: Kluwer.

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues. L'analyseur DériF*. Paris: Lavoisier.

Resultativity and the decline of preverbal *ge-* from Old to Middle English

Thomas McFadden
Universitetet i Tromsø
thomas.mcfadden@uit.no

The prefix *ge-* played an important though still poorly understood role in the verbal system of Old English, appearing on the main verb in roughly 25% of all clauses. In late OE and over the course of ME, however, its frequency dropped precipitously, and it was ultimately lost (Table 1). A clear understanding of developments of this kind cannot be gleaned from even the most careful reading of texts. And given the historical nature of the task, it is of course impossible to obtain any help from the intuitions of native speakers. In cases like this, a quantitative investigation of the patterns in the texts is the only way to proceed. In this talk I will thus present results of a corpus study of the use and disappearance of the prefix based on the YCOE (Taylor et al., 2003) and the PPCME (Kroch and Taylor, 1999). This investigation – e.g. the sorts of things I searched for in the corpus to see how they correlate with the distribution of the prefix – has been guided from the start by theoretical considerations based on earlier work on similar prefixes in other languages. Taking off from that work, I will argue that, in the OE system, *ge-* indicated a kind of resultativity (van Kemenade and Los, 2003), being the default realization of the *res* head which is one of the central components of the decomposition of (verbal) predicates proposed by Ramchand (2008). Evidence for this analysis includes the following. Unlike its modern German cognate, which has simply become a part of the past participle, OE *ge-* can appear productively on all verb forms, but it appears with far lower frequency on the present participle, typically used in non-resultative contexts, and with far higher frequency on the past participle, typically used in resultative ones (Table 2). The prefix also has an especially high frequency with achievement and accomplishment verbs like *niman* 'take', *halgian* 'hallow' and *hælan* 'heal', a much lower one with activity verbs like *sprecan* 'speak', *secgan* 'say' and *gān* 'go', and is essentially unattested with statives like the pre-modals, *bēon/wesan* 'be' and *habban* 'have' (Table 3). An analysis in terms of the *res* head is favored over other resultative analyses by *ge*'s rather low frequency with (surely resultative) consumption verbs like *drincan* 'drink' and *etan* 'eat', since, according to Ramchand's analysis, they derive their resultative interpretation from means that don't involve a *res* head. Furthermore, the otherwise problematic fact that *cuman* 'come' almost never appears with the prefix can be accounted for in Ramchand's system if *cuman*, being obligatorily resultative, 'spans' to spell out the *res* head itself, taking precedence over the default *ge*.

Such an analysis of *ge-* in OE also offers insights into its development in ME. While it shows a steady decline, this trend is not uniform across environments. It proceeds rather differently in the two places where the prefix is most common — perfects and in passives (Table 4). The frequency of *ge-* is comparable in the two in period M1, but while the subsequent drop is fairly smooth over the next three periods in the passive, in the perfect the frequency remains stable into M2, before dropping suddenly in M3. We can actually make sense of this development if we consider the resultative analysis of *ge-* being proposed here in the light of McFadden and Alexiadou (2010)'s findings on the development of the perfect in ME. They show that in OE and early ME, the periphrastic perfect was only used with a perfect-of-result reading, and thus could only be built on resultative predicates. Starting in period M3, however, a new experiential — crucially non-resultative — use of the perfect with *have* arose. This accounts for why the *ge-* was so common in the perfect in the early periods, if as proposed here it was the default morphophonological realization of the

underlying resultative structure. It also predicts the sudden drop in the frequency of *ge-* in perfects, precisely in period M3, due to the influx of the new experiential perfect. This placed no resultativity requirement on the predicates it was built on, thus did not favor *ge-* the way the old resultative perfect had. Indeed, as Table 4 shows, the marked decrease in the percentage of perfects with *ge-* in that period results not from a decrease in instances of *ge-* but from a sudden increase in the total number of perfects, as expected. The old perfect-of-result, which favored *ge-*, continued to be used at similar rates as before, but it was swamped by the new experiential one, which did not favor *ge-*, much as McFadden and Alexiadou found that the purely resultative *be*-perfect was swamped by the resultative-or-experiential *have*-perfect in the same period. I will argue that this understanding of developments will allow us to factor out the effects of the changes in the perfect system so as to isolate the changes that were peculiar to *ge-* itself and led to its disappearance.

Period	***ge-***	**no-pref**	**total**	***ge-%***
M1	2297	30190	32487	7.07%
M2	989	16850	17839	5.54%
M3	1106	58519	59625	1.85%
M4	162	31614	31776	0.51%

Table 1: Decline of *ge-* in ME

Form	**ge-**	**no**	**% ge-**
Pres. Ptc.	107	1493	6.7
Finite	23723	102434	18.8
PPP	11504	1418	89.0

Table 2: *ge-* by verb form in OE

Verb	**Gloss**	**ge-**	**no**	**% ge-**
(pre-)modals		0	2575	0.0
bēon/wesan	'be'	1	30127	0.0
habban	'have'	13	5053	0.3
cuman	'come'	29	4687	0.6
drincan	'drink'	17	779	2.1
etan	'eat'	26	538	4.6
gān	'go'	128	1927	6.2
secgan	'say'	288	3783	7.1
sprecan	'speak'	90	1134	7.4
niman	'take'	1431	1265	53.1
halgian	'hallow'	392	108	78.4
hælan	'heal'	626	110	85.1

Table 3: *ge-* by lexical verb in OE

	Perfect				**Passive**			
Period	***ge-***	**no-pref**	**total**	***ge-%***	***ge-***	**no-pref**	**total**	***ge-%***
M1	437	424	861	50.75%	967	1222	2189	44.18%
M2	217	265	482	45.02%	352	1096	1448	24.31%
M3	213	1891	2104	10.12%	691	4730	5421	12.75%
M4	10	1247	1257	0.80%	85	3136	3221	2.64%

Table 4: Decline of *ge-* in perfects and passives in ME

References

Kroch, A. and A. Taylor (1999). Penn-Helsinki parsed corpus of Middle English, 2nd ed. Univ. of Pennsylvania.

McFadden, T. and A. Alexiadou (2010). Perfects, Resultatives, and Auxiliaries in Earlier English. *Linguistic Inquiry 41*(3), 389–425.

Ramchand, G. (2008). *Verb meaning and the lexicon: A first phase syntax*. Cambridge: Cambridge University Press.

Taylor, A., A. Warner, S. Pintzuk, and F. Beths (2003). York-Toronto-Helsinki parsed corpus of Old English prose. University of York.

van Kemenade, A. and B. Los (2003). Particles and prefixes in Dutch and English. In

G. Booij and J. van Marle (Eds.), *Yearbook of Morphology*, pp. 79–118. Dordrecht: Kluwer.

Lexical category distribution in a spontaneous speech corpus of Brazilian Portuguese

Heliana Mello[1,2] Flávio C. Coelho[2] Crysttian A. Paixão[2] Renato R. Souza[2]
Tommaso Raso[1]
UFMG[1] Fundação Getúlio Vargas[2]
{hmello, traso}@ufmg.br,{fccoelho, crysttian.paixao, rrsouza}@fgv.br

This paper discusses the distribution of lexical categories in the C-ORAL-BRASIL corpus (Raso & Mello, 2012), a spontaneous speech corpus of informal Brazilian Portuguese. The corpus is comparable in architecture and segmentation criteria with the four C-ORAL-ROM corpora (Cresti & Moneglia, 2005). C-ORAL-BRASIL presents 75% of private/familiar texts and 25% of public texts; for each context 1/3 of texts are monologues, 1/3 dialogues and 1/3 conversations. Each individual text is sized at about 1,500 words. The corpus is text to speech aligned through the WinPitch software. Its main goal is to document diaphasic variation with the widest range of different communicative situations. Segmentation of the speech flow is done through prosodic criteria resulting in utterances and tone units. Utterances, defined as the smallest pragmatically interpretable unit, end with a prosodic break perceived as conclusive, while tone units – which make up utterances - end with a non- terminal break. Transcriptions were made with quasi-orthographic criteria that aimed at representing grammaticalization and lexicalization phenomena in speech, while also attempting to maintain easy readability of the texts and consistency in transcribers' perceptions.

C-ORAL-BRASIL was entirely POS tagged and syntactically marked using the PALAVRAS parser (Bick, 2012), especially adapted to deal with the particularities of this corpus. Therefore, keeping consistency in relation to the segmentation of the speech flow, the analytical unit for POS and syntactic tagging was each tone unit within a given utterance. Tone unit boundaries, marked through prosodic breaks, were read as punctuation by the parser. Hence, a semicolon was adopted as the equivalent to the (//) terminal breaks (alternating with '...' for interruptions), and a comma for the non-terminal breaks (/) The tagging and parsing show that the morphological and syntactic organization of speech is quite different from that of written texts since the latter is organized around clauses and sentences, while the former relies on information units and utterances that may or may not hold a predication, therefore making meaning production depended not only on contextual clues but on prosodic patterns as well. Additionally, not all information units behave syntactically in the same way since utterances are made up of both textual and dialogical units. The former carry the burden of providing the core meanings codified syntactically and semantically by the illocution carried by the utterance. The later, on the other hand, are responsible for keeping the communication channels open between the interlocutors.

Previous work analysing the informational organization of the three different textual types in C-ORAL-BRASIL, that is, monologues, dialogues and conversations, along with the grouping in the familiar/private versus public environments, has shown that monologues tend to be richer in complex textual units, followed by dialogues and conversations. This points out to the straighforward hypothesis that the more informationally elaborate a text type is, the heavier it will be on lexical content, since the bulk of semantic meaning is portrayed by Nouns, Adjectives, Adverbs and Verbs.

In order to test this hypothesis, we have extracted all the words POS tagged as Nouns, Adjectives, Adverbs and Verbs from the corpus correlating the score of lexical categories with the interaction type by using an R script. The resulting clustering can be seen in the dendrogram portrayed in figure 1 below.

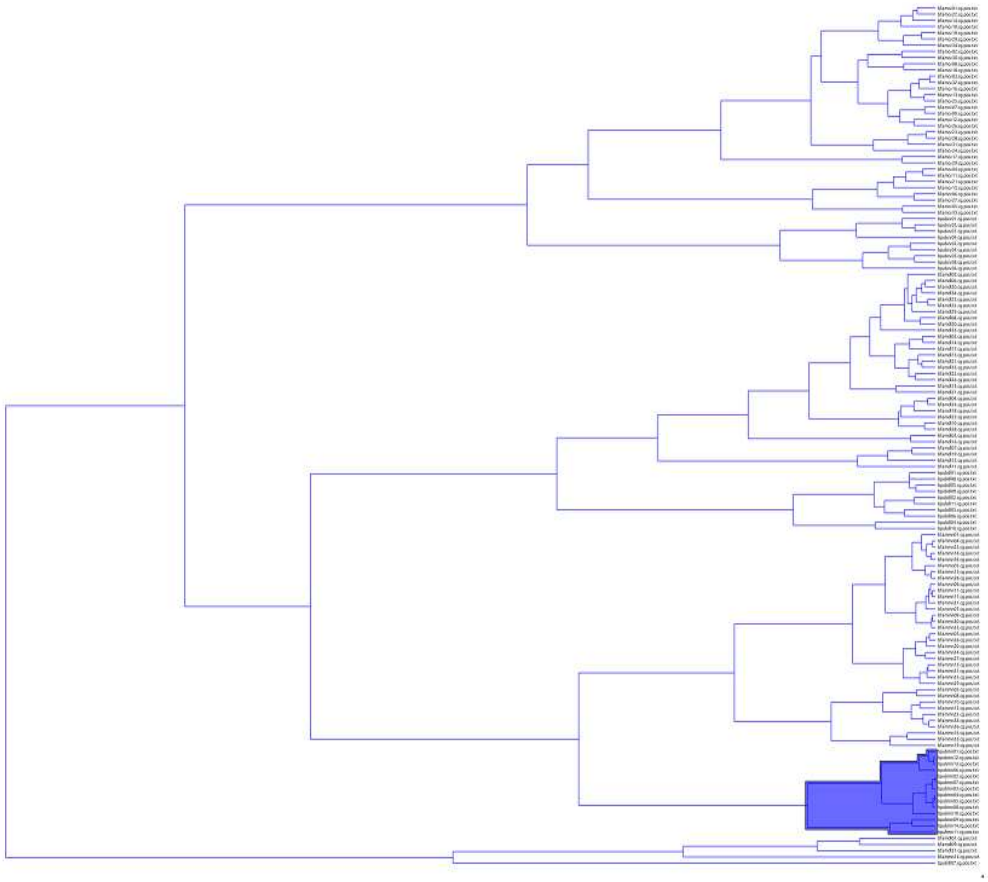

Figure 1: Lexical category distribution per text type

As predicted, the dendrogram shows that clades lead to the clustering of similar textual types, whereby the major groupings coincide with conversations as one branch of a clade while dialogues and monologues are gathered together in the other branch. Consistently with the initial hypothesis, dialogues and monologues are separate branches of the same clade. There is only a small portion of texts that behave anomalously (lower branch in first level clade). These (4 dialogues and 1 monologue, out of 139 texts) require qualitative inspection in order to be best described.

The quantitative study reported in this paper has demonstrated that there is a strong association between textual type and lexical category density in spoken Brazilian Portuguese. Monologues are the textual type that exhibit the most elaborate and extensive textual information unit organization. This is also the case for its lexical category organization. Following down on information structuring complexity scaling, there come dialogues and lastly, conversations. Coherently, the quantitative results demonstrate that as far as lexical category distribution is concerned, dialogues average closer to monologues than they do to conversations, supporting the view that conversions are more oriented towards pragmatic organization than morphosyntax.

References

Bick, Eckhard. 2012. A anotação gramatical do C-ORAL-BRASIL. In Raso. T. & Mello, H. (eds.), 223-254.

Cresti, E., Moneglia, M. (eds.) 2005. *C-ORAL-ROM: Integrated Reference Corpora For Spoken Romance Languages*. Amsterdam: John Benjamins.

Raso, T., Mello, H. (eds.) 2012. *C-ORAL-BRASIL I: Corpus de Referência do Português Brasileiro Falado Informal*. Belo Horizonte: Editora UFMG.

Distribution of modality markers in Brazilian Portuguese spontaneous speech

Heliana Mello[1,2] Flávio C. Coelho[2] Crysttian A. Paixão[2] Renato R. Souza[2]
UFMG[1] Fundação Getúlio Vargas[2]
hmello@ufmg.br, {crysttian.paixao, fccoelho, rrsouza}@fgv.br

Modality in speech can be taken to be the speaker's evaluation of an uttered locutive material. However, defining this category precisely is a difficult task due to different factors: (a) in its study tradition, modality has been the subject matter of both logical and natural language studies, which brought about a methodological maze not always productive for the research on its actual linguistic use; (b) this category interrelates with a number of grammatical phenomena such as tense, aspect and mood, prosody, information organization, among others; and (c) the concept of modality itself overlaps those as attitude, illocution and emotion in much of the literature. This paper explores the semantic notion of modality, indexed through lexical and grammatical items. The analysis is carried based on data extracted from a Brazilian Portuguese spontaneous speech corpus, the C-ORAL-BRASIL. For our purposes, modality is taken to be typologically distributed into three major categories: deontic, dynamic and epistemic. Deontic meanings relate to necessity, while epistemic meanings cover degrees of certainty and possibility, and dynamic meanings are associated to ability.

The building of C-ORAL-BRASIL took into account the utterance unit and its subinformational units, as proposed in the Language Into Act Theory (Cresti, 2000). The utterance is taken to be the minimal linguistic unit carrying pragmatic interpretability. Uterrances are made up of textual and dialogical units. Textual units carry the bulk semantic and syntactic meanings effecting communication, while dialogical units keep the communication channel open between interlocutors. The C-ORAL-BRASIL (Raso & Mello, 2012) is the fifth branch of the C-ORAL-ROM (Cresti & Moneglia, 2005), a comparable corpus representative of the four main European Romance languages (Italian, French, Spanish and European Portuguese), prosodically segmented into utterances and tone units. The corpus offers sound and transcription files, besides text-to-speech aligned files. Alignments were carried with the software WinPitch — allowing, at the same time, for the examination of sound, spectrogram and text.

In this paper we have looked at a balanced and informationally tagged subcorpus of C-ORAL-BRASIL. The subcorpus is comprised by 20 texts and 31,465 words, reproducing the architecture of the original corpus, i.e., there is a contextual partition between public and private environments, and a tripartite textual classification represented by monologues, dialogues and conversations. Modality indexes were manually annotated and classified, and later extracted through an R script. The distribution of modality indexes was evaluated in relation to frequency of tokens vis-à-vis types, frequency of modal typology in relation to lemma and frequency of modal indexes in relation to information units.

The quantitative analysis shows that there were 74 modal types distributed in 1,155 utterances, which represent 21% of the overall utterance number in the subcorpus analyzed. As for modal typology, there was a clear tendency towards epistemic modality, which accounts for nearly 80% of all modal tokens in the sample.

Modal indexes were unequally distributed among textual information units, with a clear tendency towards concentration in the core utterance unit which is the Comment. Besides the Comment unit, Bound Comments and Multiple Comments also exhibited clustering effects in modal indexes distribution. There were no modal occurrences in dialogical units.

The findings can be visualized through the graphs in figures 1 and 2, which show respectively the association between lemmas and modal categories and lemmas and information units. The graphs demonstrate the association strength between different notes connected through vectors. Therefore, the more vectors associated to a given node, the more representative that node is in relation to the analyzed parameter. Graphs have algorithms that calculate the weight terms have in a given document in relation to certain parameters. In our study, modal indexes were correlated both with their typological categorization and their distribution in information units.

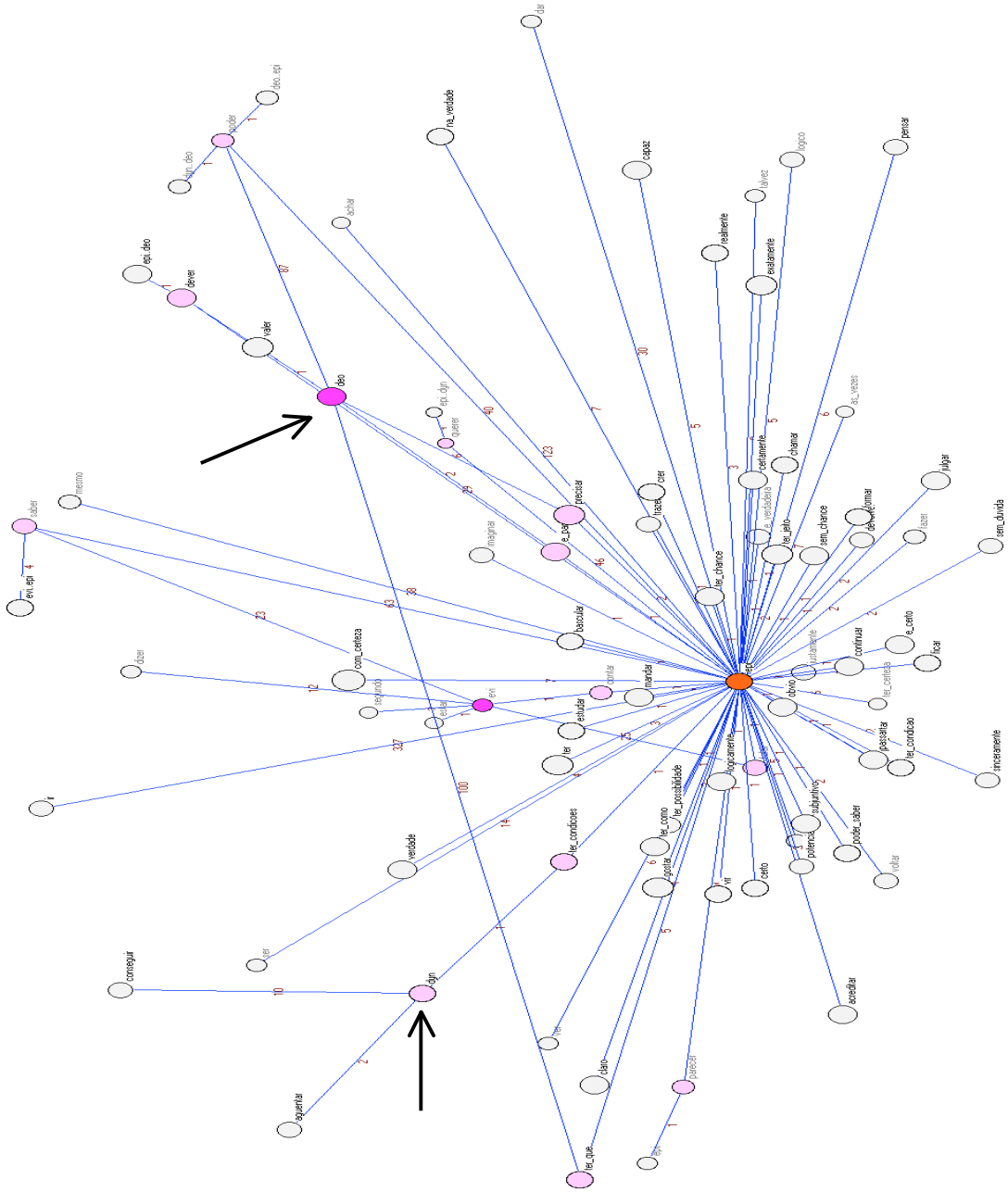

Figure 1: Modal lemmas distribution in modal categories

The graph in figure 1 has as its central node, in orange, epistemic modality. It is evident that epistemic modality has a much higher association rate to diverse modal indexes than

deontic and dynamic modality, represented in the graph through the light and dark pink nodes respectively.

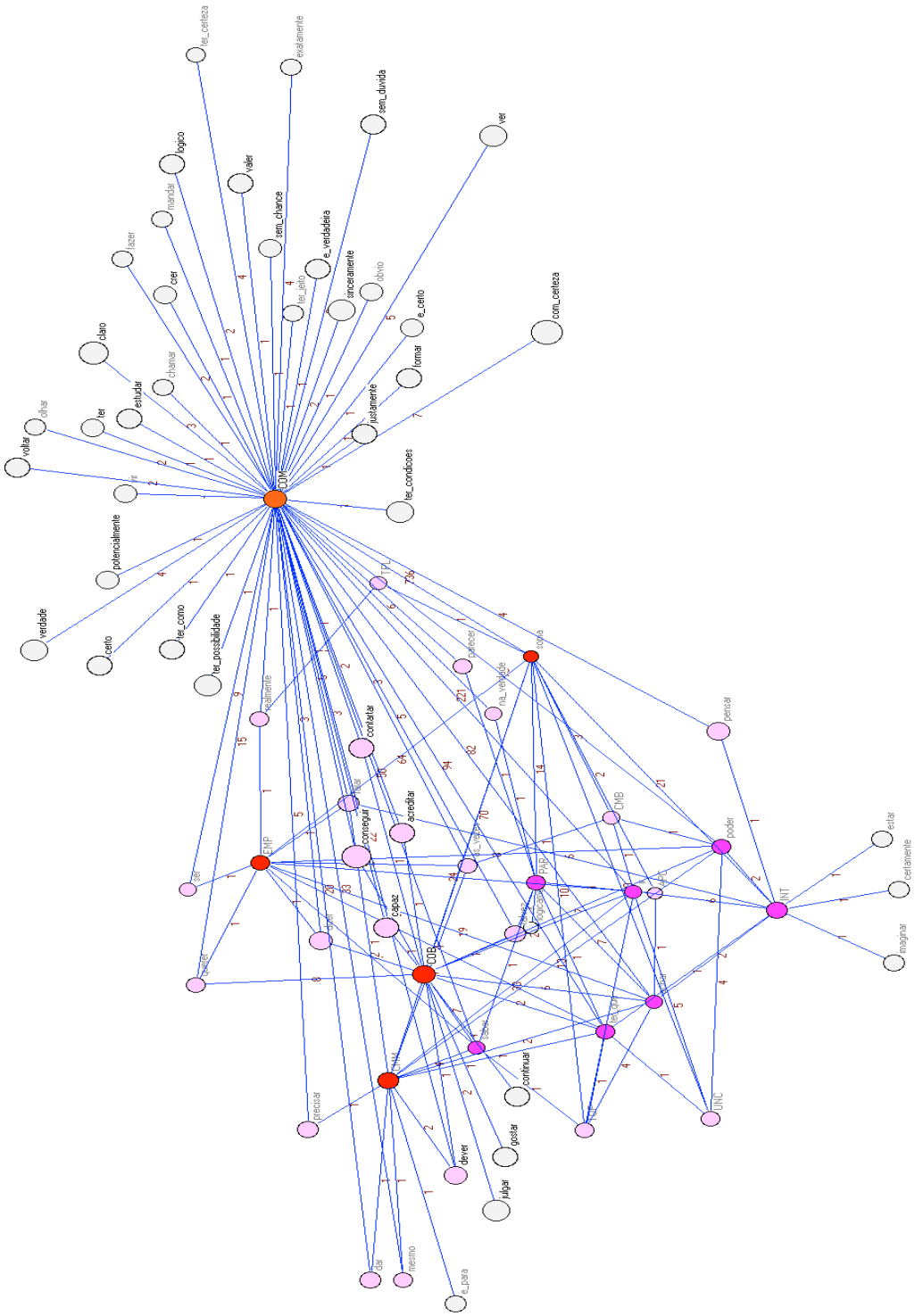

Figure 2: Modal lemmas distributed in information units

The graph in figure 2 shows the association strength between diverse modal indexes and the simple Comment information unit to be the predominant one in the data analyzed. The Comment is represented by the nuclear orange node in the graph. Other types of complex

Comment units, such as Multiple Comments and Bound Comments represented in the graph by red nodes, also present noteworthy association rates.

This study has shown that modal indexes are not chaotically distributed in spontaneous speech. We have found that there are strong correlations between modal indexes and specific modal values, with epistemic modality having the highest association score rate, as well as strong modal indexes distribution patterns in relation to information units, with the Comment unit scoring the highest.

References

Cesti, Emanuela. *Corpus di Italiano parlato*. 2000. Firenze: Accademia della Crusca. Vol 1.

Cresti, Emanuela, Moneglia, Massimo. (eds.) 2005. *C-ORAL-ROM: Integrated Reference Corpora For Spoken Romance Languages*. Amsterdam: John Benjamins.

Raso, Tommaso, Mello, Heliana. (eds.) 2012. *C-ORAL-BRASIL I: Corpus de Referência do Português Brasileiro Falado Informal*. Belo Horizonte: Editora UFMG.

How valuable are our judgments? Towards a better understanding of metalinguistic judgment data

Maria Mos Véronique Verhagen
Tilburg University
{maria.mos,v.a.y.verhagen}@tilburguniversity.edu

Metalinguistic judgments form an important and oft-used used type of data in linguistics. Their usefulness lies in that judgments are behaviors that bear on the cognitive systems that subserve language. Therefore, judgments can be used as evidence for making inferences about these cognitive systems. It is even argued that judgment data play a *crucial* role in linguistic investigation because they provide quantitative information not readily available from other kinds of data (Schütze and Sprouse, to appear). The use of such judgments as a source of linguistic evidence, however, is regularly called into question. An often-raised objection is that speakers' access to their experience of language is random and incomplete, thus challenging the reliability of judgments, especially when these judgments are not binary choices but involve gradient acceptability (Cook, 1998, in McGee, 2009).

This debate challenges us to try and further the understanding of judgment data as quantifiable measurements of linguistic knowledge. Our study contributes to this in three ways: by looking at the effects of the presence and absence of context, by investigating the test-retest reliabilities and by examining to what extent judgments can be predicted from stimuli's frequency.

The focus in this study is on Dutch multiword units (prepositional phrases, more specifically), and the judgment data concern perceived familiarity of these units, which bears on the degree to which the word strings are stored (or: entrenched). It is subject of discussion whether or not the degree of entrenchment of multiword units involves linguistic processes which are too deeply embedded for introspection. Biber et al. (1996, p.120), for example, believe that, "intuitions regarding lexical associations are often unreliable and inaccurate".

Possibly, intuitions regarding entrenchment levels seem unreliable because the context in which the stimuli are presented varies widely across experiments. Generally, linguistic processing studies present the stimuli as isolated word strings (e.g. Arnon and Snider, 2010; Tremblay and Baayen, 2009), while in judgment tasks it is common to embed the target phrases in a meaningful sentence. Context is likely to influence the way in which the words are processed (see, for example, Camblin et al., 2007). By presenting the same stimuli both with and without a sentential context, we examine the ways in which context may affect judgments.

Our study is also innovative in that it examines how stable judgments are during a period of two to three weeks, the importance of which is stressed by Labov:

> Linguists are building on sand until they can answer basic questions: what are the test-retest reliabilities of judgments of grammatical acceptability? Under what conditions do introspections match speech production? What are the sources of bias? Many hundreds of authors have published articles based on introspective data, but only a half dozen have been concerned with this issue. (Labov, n.d.)

We collected the familiarity judgments by means of a Magnitude Estimation task (Bard et al., 1996). This is a specific kind of judgment task allowing participants to distinguish as many different grades of familiarity as they feel are relevant. Participants were asked to rate 44 Dutch simple prepositional phrases (PPs), such as *op school* ('at school') and *met de hond* ('with the dog').

For each phrase, frequency data were collected using the Corpus of Spoken Dutch (CGN). The same corpus was used to determine prototypical contexts for the target phrases. Sentences were constructed with these prototypical elements and the target phrase was always positioned in the second part of the sentence, so as to minimize differences in prominence.

The judgment task consisted of two parts: one with the prepositional phrases presented as isolated word strings, and one in which they were embedded in a sentence. The order of the two parts and of the stimuli within each part was randomized across participants. For contextualized phrases, participants were instructed to judge the underlined phrase in the given context. When the phrases appeared without a context, participants were informed that they were free to think of usage contexts.

The participants (N = 86) were first-year students at *Anonymous-Institution*. All of them grew up in the Netherlands, with Dutch as (one of) their mother tongue(s). The judgment task was done twice within a period of two to three weeks. Participants were not informed in advance that they would perform the same task twice.

Since participants were free to use their own judgment scale, familiarity scores for each participant were converted to Z-scores. In order to investigate stability, Δ-scores were then calculated for each participant's judgments of the stimuli at time 1 and time 2. Overall, judgments were found to be fairly unstable. Furthermore, stability in judgments was not higher for contextualized stimuli than it was for isolated stimuli.

Interestingly, context affected familiarity ratings differently for different stimuli. One third of the phrases were rated as more familiar when presented in a sentence; another third received higher scores when presented as isolated word strings; and the remainder did not show a significant difference.

Contrary to our expectations, there was no difference between isolated and contextualized items in the degree of individual variation in judgment, using standard deviations as a measure of diversity in judgments.

Using a regression model with CGN-based frequency measures for contextualized and isolated stimuli, it was determined that the frequency of the exact PP is a significant predictor, with more frequent PPs receiving higher familiarity ratings.

In sum, familiarity judgments in a ME-task are influenced by phrase frequency, but others factors too. These judgments are far from stable over the (rather short) time span measured. Providing a prototypical context does not reduce instability, but in many cases it does affect judgments.

In our presentation we will go into a full consideration of the variation between items and across participants in the (in)stability of judgments, the effects of context, and the relationship with corpus frequencies. We will present different explanations for our findings and discuss how they contribute to the debate on judgment data as reliable, valuable linguistic evidence in theories on entrenchment of multiword units.

References

Arnon, Inbal, and Neal Snider. 2010. More than Words: Frequency Effects for Multiword Phrases. *Journal of Memory and Language* 62: 67-82.

Bard, Ellen Gurman, Dan Robertson, and Antonalle Sorace. 1996. Magnitude Estimation and Linguistic Acceptability. *Language* 72: 32-68.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1996. Corpus based investigations of language use. *Annual Review of Applied Linguistics* 16: 115-135.

Camblin, Christine, Peter Gordon, and Tamara Swaab. 2007. The interplay of discourse congruence and lexical association during sentence processing: evidence from ERPs and eye tracking. *Journal of Memory and Language* 56 (1): 103-128.

Labov, William. N.d.. *Some observations on the foundations of linguistics.* (Unpublished Manuscript) Retrieved from http://www.ling.upenn.edu/~wlabov/Papers/Foundations.html

McGee, Iain. 2009. Adjective-noun collocations in elicited and corpus data: similarities, differences and the whys and wherefores. *Corpus Linguistics and Linguistic Theory* 5 (1): 79-103.

Schütze, Carson, and Jon Sprouse. To appear. Judgment Data. In: Sharma and Podesva (eds.), *Research Methods in Linguistics*. Cambridge: Cambridge University Press.

Tremblay, Antoine, and Harald Baayen. 2009. Holistic processing of regular four-word sequences. In: Wood (ed.), *Perspectives on Formulaic Language in Acquisition and Production*. London and New York: Continuum.

A Quantitative and Typological Approach to Correlating Linguistic Complexity

Yoon Mi Oh François Pellegrino Egidio Marsico Christophe Coupé
Laboratoire Dynamique du Langage, Université de Lyon and CNRS, France
{yoon-mi.oh, francois.pellegrino}@univ-lyon2.fr,
{egidio.marsico, christophe.coupe}@cnrs.fr

1 Hypothesis and objectives

The equal complexity hypothesis states that "all human languages are equally complex" (Bane, 2008). Menzerath's law is well-known for explaining the phenomenon of self-regulation in phonology: "the more sounds in a syllable the smaller their relative length" (Altmann, 1980). Altmann, who made the mathematical formula of this law (Forns and Ferrer-i-Cancho, 2009), assumed that it can be applied to morphology as well - "the longer the word the shorter its morphemes" (Altmann, 1980) - and proved that the clause length depends on sentence length (Teupenhayn and Altmann, 1984).

Some previous works on morphological complexity (Bane, 2008; Juola, 1998) asserted that morphology is a good starting point for complexity computation for its clearness, compared to other more ambiguous domains such as semantics. The best-known method of calculating morphological complexity is to take the numbers of linguistic constituents into account (Bane, 2008; Moscoso del Prado, 2011), with different mathematical formula to be applied to these figures. The following two paradigms are commonly employed: i) information theory (Fenk et al., 2006; Moscoso del Prado et al., 2004; Pellegrino et al., 2011) ii) Kolmogorov complexity (Bane, 2008; Juola, 1998).

The main goal of our work is to explore interactions between phonological and morphological modules by means of crossing parameters of these two linguistic levels. This paper provides preliminary results obtained from a corpus-based cross-language study.

2 Methodology and preliminary results

Our 14-language corpus is based on the Multext multilingual corpus (Campione and Véronis, 1998). For each language, 15 short texts which consist of 3-5 sentences translated from British English are recorded by 5 male and 5 female native speakers. The data of 6 languages (English (eng), German (deu), Italian (ita), Mandarin Chinese (cmn), Spanish (spa) and Vietnamese (vie)) are taken from the Multext corpus, and the data of the other 8 languages (Basque (eus), Catalan (cat), French (fra), Hungarian (hun), Japanese (jpn), Korean (kor), Turkish (tur) and Wolof (wol)) have been collected by the authors.

Two types of parameters are taken into account in this study. First, at the phonological level, and following Pellegrino et al. (2011), a set of phonological factors is employed. For each language, the *syllabic rate*, (the number of syllables pronounced per second), is computed. Additionally, using Vietnamese as an external reference, a *syllabic information density* (resp. *word information rate*) is defined for each target language as the average ratio between the total number of syllables (resp. words) in a text in Vietnamese and the number of syllables (resp. words) of this text translated in the target language.

Our method of measuring the information density computes the average amount of information carried by syllables and words at the text level. Thus, it differs from studies related to the principle of uniform information density (Frank and Jaeger, 2008), since the latter focus on the variation of information transmitted during communication. Figure 1 illustrates the negative correlation (R^2= 0.65) between these phonological factors, i.e. a trade-off between syllabic rate and information density (Pellegrino et al., 2011).

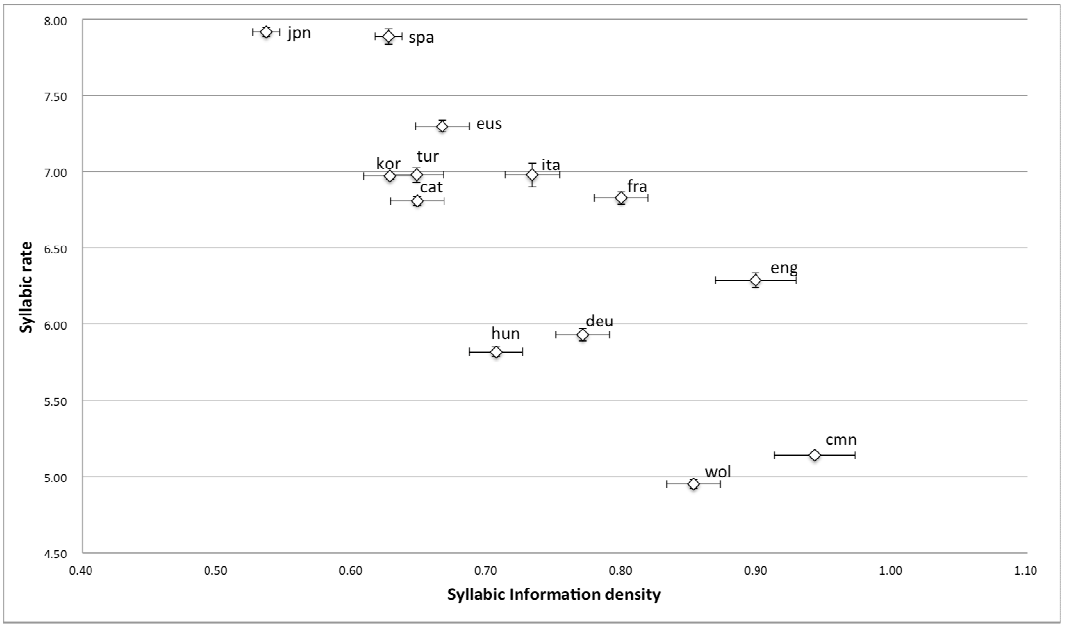

Figure 1: Syllabic rate and syllabic information density (Error bars indicate standard error)

Second, at the morphological level, the languages of our corpus can be classified into three categories, as shown in Table 1 (Greenberg, 1960).

Category	Languages
Agglutinative languages	Basque, Hungarian, Japanese, Korean, Turkish
Fusional languages	Catalan, English, French, German, Italian, Spanish, Wolof
Isolating languages	Mandarin Chinese, Vietnamese

Table 1: Morphological classification

In order to investigate the relations between phonological and morphological modules, we compare the average number of syllables per word and the information density calculated at the word level and at the syllable level, respectively in Figures 2 and 3.

Figure 2 exhibits a strong positive correlation (R^2= 0.84) between the average number of syllables per word and the information density at the word level, which logically means that the longer the word, the more information it contains. In general, there are more syllables per word in agglutinative languages (in black) than in fusional languages (in grey). Chinese as an isolating language is marked in white. Furthermore, Figure 3 shows that at the syllable level, fusional languages have a tendency towards higher information density compared to agglutinative languages.

Values of languages in the same morphological category are quite dispersed. In Figure 2, regarding fusional languages, a large difference exists, for example, between German with a very complex declension system and English with a limited morphological system (Moscoso del Prado, 2011). Japanese, which has a relatively simple phonological system, has the largest number of syllables per word and transmits the least amount of information per syllable (Figure 3). Compared to Japanese, Mandarin Chinese, an isolating language with a relatively complex phonological system, shows completely opposite values.

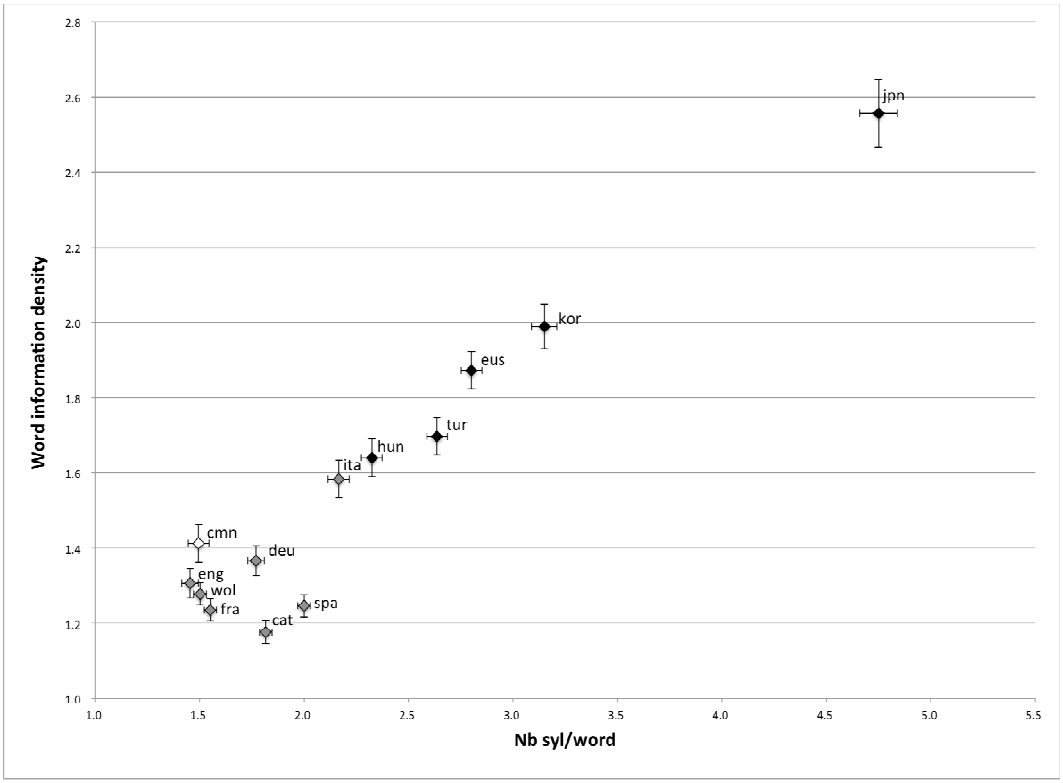

Figure 2: Word information density and mean number of syllables per word (Error bars indicate standard error)

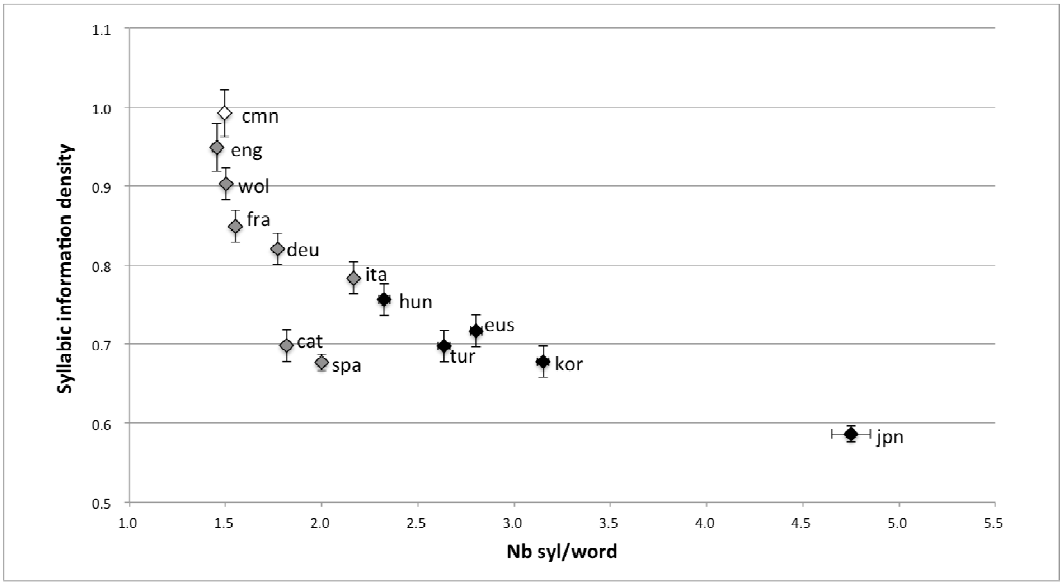

Figure 3: Syllabic information density and mean number of syllables per word (Error bars indicate standard error)

3 Discussion and further work

Fenk et al. (2006) defined word complexity as the mean number of syllables per word and syllable complexity as the mean number of phonemes per syllable, and found a negative linear correlation between these two figures. Similarly, our result shows a negative corre-

lation between word complexity and information density at the syllable level, i.e. the less complex a word, the more information per syllable.

Furthermore, according to our results, despite the dispersed values of languages in the same morphological category, some differences are observed between these categories and agglutinative languages clearly tend to have longer words than fusional languages. Fenk-Oczlon and Fenk (1985) showed that the average number of syllables per clause depends on the mean number of phonemes per syllable, but the analysis at word level had not been done before.

These preliminary results show a relation between the morphological and phonological modules. In further studies, this relation will be investigated in more details by analyzing our multilingual parallel data, and by adding more isolating languages to observe their pattern. We are currently working on unsupervised morpheme segmentation, using Morfessor (Creutz and Lagus, 2005), in order to compare our multilingual data at morpheme level. At the same time, we aim to compare the average number of words per sentence in order to correlate the linguistic complexities of three different levels.

References

Altmann, G. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2, 1–10.

Bane, M. 2008. Quantifying and measuring morphological complexity. In *Proc. of the 26th West Coast Conference on Formal Linguistics*, 69-76.

Campione, E. and Véronis, J. 1998. A multilingual prosodic database. In *Proc. of ICSLP98*, Sydney, Australia, 3163-3166.

Creutz, M., and Lagus, K. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0*. Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March. URL: http://www.cis.hut.fi/projects/morpho/

Fenk, A., Fenk-Oczlon, G. and Fenk, L. 2006. Syllable complexity as a function of word complexity. *In The VIII-th International Conference "Cognitive Modeling in Linguistics" Vol. 1*, 324-333.

Fenk-Oczlon, G., and Fenk, A. 1985. The mean length of propositions is 7 plus minus 2 syllables–but the position of languages within this range is not accidental. In *Proc. of the XXIII International Congress of Psychology: Selected/Revised Papers*, Vol. 2, 355–359.

Forns, N. and Ferrer-i-Cancho, R. 2009. The self-organization of genomes. *Complexity, 15(5)*, 34-36.

Frank, A., and Jaeger, T. F. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proc. of the Cognitive Science Society*.

Greenberg, J. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics, 26(3)*, 178-194.

Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Moscoso del Prado, F., Kostić, A., and Baayen, R.H. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94(1)*, 1-18.

Moscoso del Prado, F. 2011. The Mirage of morphological complexity, In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, 3524-3529.

Pellegrino, F., Coupé, C., and Marsico, E. 2011. A cross-language perspective on speech information rate. *Language, 87(3),* 539–558.

Teupenhayn, R., and Altmann, G. 1984. Clause length and Menzerath's law. *Glottometrika 6,* 127-138.

Two methods for automatic identification of cognates

Taraka Rama[*1], Prasant Kolachina[†2], and Sudheer Kolachina[‡3]

[1]University of Gothenburg
[2]International Institute of Information Technology, Hyderabad
[3]Massachusetts Institute of Technology

1 Introduction

Cognate identification is a important task in historical linguistics for the purpose of establishing genealogical relationships between languages (Campbell, 2004). Cognates are identified through regular sound correspondences between words, from supposedly related languages, having a similar surface and semantic forms. However, not all cognate pairs are equally similar. In other words, cognacy judgment is not a binary decision but a finely graded one. Consider a cognate pair, German and English, *hund* ∼ *hound* which reveals itself to be cognate through visual inspection. The cognacy similarity score for such a pair should be very high. Now, consider a cognate pair, Sanskrit to English, *chakra* ∼ *wheel* whose similarity is not revealed through visual inspection. Such a cognate pair should have lower similarity score since the regular sound processes have affected the original proto-word to show divergent surface yet, related forms in the distantly related languages. In the rest of the paper, we propose two methods in section 2 for the purpose of cognate identification. We describe the multi-lingual dataset in section 3 and validate our methods in section 4.

2 Methods

Automatic detection of sound correspondences is the crucial step in cognate identification. Sound correspondences can be extracted using the alignments obtained from Levenshtein distance (LD; Levenshtein 1965). LD is defined as the minimum number of insertion, deletion and substitution operations required to transform a string into another with all the operation costs set to 1. However, the alignments generated through LD might not be linguistically meaningful. For instance, the alignments between the words for ashes: Catalan 'sendra' and Italian 'tenere' would be 's' : 't', 'e' : 'e', 'n' : 'n', 'd': 'e', and 'a' and 'e'. The phoneme segment pair 'd': 'e' is linguistically implausible, since a consonant cannot align with a vowel.

Wieling et al. (2009) alleviate this problem, for the classification of Bulgarian dialectal data, by introducing an additional constraint (VC-constraint) that vowels cannot align with consonant and vice-versa. Their approach is summarized as follows:

1. Employ the VC-constraint LD to align a word pair and extract all possible phoneme segment pairs for a pair of languages.

[*]taraka.rama.kasicheyanula@gu.se
[†]prasant@research.iiit.ac.in
[‡]sudheer@mit.edu

2. The similarity of a segment pair is computed using Pair-wise Mutual Information (PMI; Church and Hanks 1990) which is defined as $\log p(x,y) - \log p(x) - \log p(y)$.

3. The segment pair similarity is converted into a distance score, in the range of $[0,1]$, through the formula

$$\frac{max_{pmi} - pmi}{max_{pmi} - min_{pmi}} \quad (1)$$

4. The pair-wise item LD is computed using the segment pair distances obtained from step 3.

Steps $1-5$ are repeated until there is no change in the segment pairs between two successive iterations. The final iteration of the above algorithm yields a list of segment pair distances.

However, the VC-constrained LD operates at a single segment level i.e. the method always operates on a single segment pair. This method, when extended to multiple length segments allows alignment between segments of length greater than 1. Bergsma and Kondrak (2007) employ the idea of multiple length segments to train a linear classifier for the automatic identification of cognates from bi-text data. They align a word pair using the basic LD and extract adjacent segment pairs. The maximum length of a segment pair is limited to 3 in their experiments. This approach is expected to identify word pairs which need not be genetically related but are borrowings. The same authors also identify their multiple segment approach similar to that of "phrases" in Statistical Machine Translation (SMT).

Pursuing the idea of "phrases", the class of generative alignment models commonly referred to in literature on SMT as IBM models (Brown et al., 1993) can be used to generate alignments across multiple length segments. These models are used to align words between translations across a language pair, and are naturally designed to generate alignments between multiple length segments across two languages, unlike the traditional LD method. The IBM models utilize information such as frequency counts and co-occurrence counts across the word lists to generate alignments, using minimal linguistic information. We extend the same approach to automatically align multiple length segments in a word pair across two languages.

For any given pair of languages, the word pairs for identical concepts are extracted to create a bilingual word list for the language pair. Each of the word pairs are aligned using the IBM models to extract multiple length segment pairs. We compute a PMI-based segment distance score for each of the multiple segment pairs using the normalization formula given in equation 1. In our experiments, we limit the maximum length of a segment in the pair to 2. The alignments are obtained using the implementation of publicly available IBM models available in Moses (Koehn et al., 2007). The toolkit additionally provides multiple heuristic algorithms to extract high quality alignments generated from the IBM models. We use all of these heuristics to extract segment pairs prior to the computation of the PMI score for each segment pair.

We observed that the original IPA transcribed data has fine distinctions such as vowel length and primary stress. We ignored the vowel length distinction and stress pattern. Further, the IPA symbols are mapped to a reduced sound class alphabet consisting of 21 symbols; 15 consonant classes and 6 vowel classes (given in List 2012) to encounter symbol sparsity. In all our experiments, k was set to 1.

The contributions of this paper is threefold:

1. We apply the linguistically motivated Levenshtein distance to the task of cognate identification on three different datasets given in List (2012).

2. We apply a popular SMT technique to align phoneme segments between semantically equivalent word pairs and use the segment pair distance to compute the LD between a word pair.

3. We introduce a new evaluation measure to quantify the performance of the two measures.

3 Dataset and Evaluation Measures

Language Group	Number of languages	Number of items
Indo-European (IE)	20	207
Germanic (GER)	7	110
Uralic (URL)	21	110

Table 1: Number of languages and items in the three language groups.

The dataset, in table 1 also contains the cognacy judgments for a item between a pair of languages. In a language pair, the pair-wise item distances are compared to the gold standard cognate judgments using point-biserial correlation (a special case of Pearson's *r*). In each iteration, we compute the average cognate identification accuracy by taking the average of the correlation for all language pairs. The improvement of the average correlation between two successive iterations is measured through a paired *t-test* with significance level set at 0.05.

4 Results

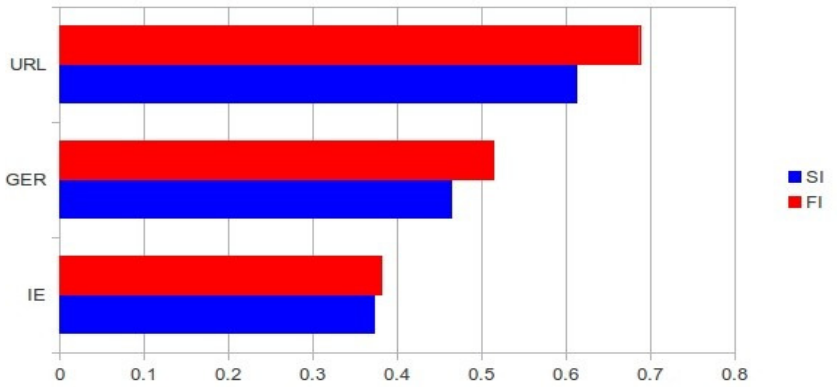

Figure 1: Results of VC-constrained LD

Figure 1 shows the improvement in the average correlation between the starting iteration (SI) and the final iteration (FI) for each language group. The starting iteration makes use of the basic VC-constrained levenshtein distance. The subsequent iterations make use of the PMI-based segment pair distances to compute the LD for a word pair. Table 2 shows the number of iterations the algorithm required to converge as well as the number of statistically significant iterations.

Language group	Number of iterations	Significant iterations
IE	2	2
GER	6	2
URL	4	4

Table 2: Number of iterations and significant iterations for each language group. The significant iterations are always less than or equal to the number of iterations to converge.

Figure 2 shows the results of SMT derived segment pair distances in computing the LD for a word pair. The result for Uralic language group is comparable to the result of the VC-constrained LD. The results for Indo-European and Germanic datasets are lower than the results for VC-constrained LD. It has to be noted that the algorithms are not directly comparable. VC-constrained LD merges the segment pairs generated from all language pairs and then computes the PMI-based distance score for a segment pair. Whereas, the SMT-based alignments are generated independently for each language pair and the PMI-based segment distances are also computed independently for each language pair.

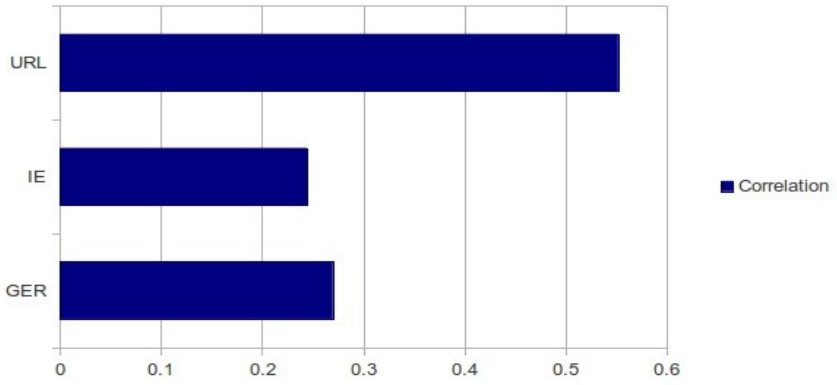

Figure 2: Results of SMT derived segment pair distances. Each bar in the figure shows the average agreement of the method between the pair-wise distances and the gold standard.

5 Conclusion

In this abstract, we described and applied two statistically driven algorithms for the task of cognate identification to three multi-lingual datasets. The initial results suggest that both the approaches are worth pursuing and can be applied to the four other language groups' datasets listed in List (2012). As a future work, we propose that the VC-constrained LD be used for computing the segment pair distances for each language pair. Also, the SMT based segment distances should be computed for the overall language pairs for a direct comparison between the two methods.

References

Shane Bergsma and Grzegorz Kondrak. Alignment-based discriminative string similarity. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 656, 2007.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Mathematics of statistical machine translation: Parameter estimation. pages 19(2):263–311, 1993.

L. Campbell. *Historical Linguistics: An Introduction*. MIT Press, 2004.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. ISSN 0891-2017.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster*

and Demonstration Sessions, pages 177–180. Association for Computational Linguistics, 2007.

VI Levenshtein. Binary codes capable of correcting spurious insertions and reversals. *Cybernetics and Control Theory*, 10:707–710, 1965.

Johann-Mattis List. Lexstat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France, April 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-0216`.

Martijn Wieling, Jelena Prokić, and John Nerbonne. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 26–34. Association for Computational Linguistics, 2009.

Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs

Sylvia Springorum Sabine Schulte im Walde Antje Roßdeutscher
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{sylvia.springorum,schulte,antje}@ims.uni-stuttgart.de

Compositionality of German particle verbs German particle verbs (PVs) are highly productive combinations of a base verb and a prefix particle. Concerning their semantics, there is an ongoing discussion whether the meaning of German particle verbs is in general compositional or not. For example, Kratzer (2003) claimed that German PVs are idiosyncratic; this stands in opposition to the semantic analyses by Lechler and Roßdeutscher (2009), Kliche (2011), among others. who demonstrated that each particle has several different readings which however form regular patterns depending on the contexts. Our position is in-between: we agree that not every PV composition is transparent, but with a fine-grained sub-lexical analysis and taking analogy and meaning shift mechanisms into account, the majority of combinations can be explained by patterns.

Our research focuses on how speakers of German combine particle senses with base verb senses. Questions which come along with this focus are: (i) how applicable, (ii) how available and (iii) how common or prototypical is a semantic pattern of a meaning composition?

Goal of this study This study presents preliminary insights into an ongoing experiment for German PVs, where the participants generate sentences with attested PVs and also with yet not attested formations which we call systematic Neologisms of German Particle Verbs (neoPV). A PV is a neoPV if it is not listed in the Duden dictionary [1] and if it is not attested in the German web corpus *SdeWac* (Faaß *et al.*, 2010). The main assumption for the experiment is that if PVs are compositional and productive, neoPVs should have at least one understandable meaning. If neoPVs are given an interpretation by way of generating sentences with them, the idea of a rule based interpretation is hard to deny. In the following, we first describe the experiment to collect the neoPV data, and then perform a quantitative and qualitative description of the preliminary results, over all data and focusing on specific subsets.

Experiment The experiment is running with Amazon Mechanical Turk (AMT). The subjects are presented with a PV and two tasks: first, they are asked to provide a rating of 0-3 whether the PV is known or unknown or how familiar they are with it. Then, they have to generate at least one sentence using the PV, such that the sentences illustrate the verb meaning. After the generation, the subjects have the opportunity to mark a checkbox, if they feel it was difficult to generate a sentence for the particular PV.

The data comprise a total of 125 PVs: Five different particles (*ab, an, auf, aus, nach*) were combined with verbs from five different semantic verb classes: (1) DE-ADJECTIVAL e.g. *kürzen* 'shorten', (2) ACHIEVMENT/ACCOMPLISHMENT e.g. *finden* 'find', (3) PHYSICAL PROCESS e.g. *stricken* 'knit', (4) MENTAL PROCESS e.g. *denken* 'think', and (5) STATE e.g. *lieben* 'love'. The chosen base verbs (BVs) were balanced for corpus frequencies in the SdeWaC.

[1] The PVs were looked up in the online version of the dictionary: www.Duden.de

Results Table 1 shows for each BV class (column 1) the amount of so far generated sentences (column 2), the familiarity of a PV (column 3), the difficulity, which refers to the previously mentioned checkbox option (column 4), the percentage of these 'difficult' cases of the unknwon PVs and the amount of neoPVs in the presented data (last column).
So far, we collected 1,470 sentences, out of which 863 contain a neoPV unknown to the subject. In 230 of these unknown cases (26.65%), the subject claimed that it was difficult to think of a sentence. Comparing across the semantic BV classes, the unknown PHYSICAL PROCESS verbs seem to be easier to handle than the other BVs (only 20.71% were difficult). This finding fits with our expectation, that it should be easier to apply a particle meaning to verbs with a homogeneous event structure than to verbs either coming with a result or a state by themselves, because particles often contribute result and state to BVs.

BV Class	**Sentences**	**Unknown**	**Diffcult**	**% of Diff.**	**neoPV**
All	1470	863	230	26.65	81
DE-ADJECTIVAL	303	199	57	28.64	19
ACHIEV./ACCOMP.	310	134	37	27.61	12
PHYSICAL PROCESS	305	140	29	20.71	15
MENTAL PROCESS	301	165	48	29.09	16
STATE	251	225	59	26.22	22

Table 1: Quantification of current results.

Results for the BV *stricken* Taking a closer look at the PHYSICAL PROCESS verb *stricken* (Table 2), we have collected 49 sentences so far, with 26 PV ratings as 'unknown' and 5 ratings as 'difficult'. The attested verbs with this BV are *aufstricken* and *anstricken*, and the neoPVs are *abstricken*, *ausstricken* and *nachstricken*. Surprisingly, the attested verb *anstricken* was judged as unknown in 5 of the 9 sentences and the also attested *aufstricken* was even judged as unknown in 9 of 11 sentences. On the contrary, we also find the reverse case, where for the neoPV *abstricken* only 3 of the 8 sentences were marked as 'unknown'.

Verbs	**Sentences**	**Unknown**	**Diffcult**	**neoPV**
All	49	26	5	3
ab-	8	3	1	+
an-	9	5	2	-
auf-	11	9	1	-
aus-	10	8	1	+
nach-	11	3	-	+

Table 2: Quantification of results for the BV *stricken*.

Results for the BV *abstricken* Table 3 shows the distribution of *ab* readings in sentences with *abstricken* in relation to whether the PV was known or unknown. For the known verbs, in one case *ab* was given the COPY reading as in *Ein Bild abmalen* 'to copy a picture', in the other case it is not clear which reading was used. The COPY reading also occured for one unknown case. We also found sentences, where *ab* has the END OF SUPPORT reading, implying an end of a contact relation as in *Knopf von einer Hose abreißen* 'to rip off the button of a trouser', a QUANTIFICATION reading as in *die Aufgaben abarbeiten* 'to complete a task step by step' and a TERMINATION reading as in *das Baby abstillen* 'to weab the baby'. The TERMINATION reading was used 3 times, and example 1 is one of the sentences. The adjective *letzten* 'last' in this case shows that there must be a semantics which terminates the *stricken* event.

(1) *Die letzten Maschen müssen abgestrickt werden. Das ist nicht schwer.*

Reading	Known	Unknown	Diffcult
END OF SUPPORT		1	1
COPY	1	1	
QUANTIFICATION		1	
TERMINATION	3		
Undef	1		

Table 3: Distribution of particle readings for *abstricken*.

'The last stitchs have to be cast off. This is not difficult'

Results for PVs with particle *ab* Table 4 shows the distribution of the readings of the particle *ab* over all 5 PHYSICAL PROCESS BVs. The most common readings are the END OF SUPPORT and the QUANTIFICATION reading occuring in around 10 sentences each. So it seems that these two readings tend to be more prototypical than the others.

Another very interesting point is the metaphorical use of a neoPV which occured several times with different particles and BVs. This means in general, that people are not only able to compose the meaning of a PV but also to embed the resulting concept in another domain and in an understandable way. The PVs in the Sentences 2 is not literal. In 2 we have the abstract object *Arbeitstelle* 'job' which was mentally attached to some future life plans, but since the job interview did not go well, the job has to be mentally detached. This was expressed by the PV *abnageln* '[ab] + to nail', an abstract END OF SUPPORT *ab* and an abstract interpretation of *nageln* 'to nail'. We see the modal context together with the dativ as an evidence, that the construction is analogous to the existing metaphorical reading of *sich etwas abschminken können/müssen* 'to get something out the head/literally: to be able to remove make up' in example 3.

(2) *Das Vorstellungsgespraech lieft gar nicht gut, die neue Arbeitsstelle kann ich mir wohl abnageln.*
'The job interview didn't go well, I have to get the new job out of my head.'

(3) *Wenn du weiter so verschwederisch lebst, kannst du dir die Reise abschminken.*
'If you keep on living lavishly then you can get the travel out of your head.'

Reading	*abnageln*	*abstricken*	*abrühren*	*abschaukeln*	*abschlafen*	Sum
END OF SUP.	5	1	1	1		8
QUANT.	3	1		1	5	10
COVER	1					1
TERMINATION		3		1		4
COPY		2				2
USE UP				1	1	2
MIX			4			4
METAPHOR	1		1	2		4
Total	15	8	9	7	10	

Table 4: Distribution of readings for *ab* + PHYSICAL PROCESS BVS.

Even though the experiment is not finished yet we showed that (i) there are several interpretations of the PVs and the particles and some of them seem to be more difficult (ii) that not all readings have to be available to everyone, cf. example with *anstricken*, which is attested, but was rated in more than half of the sentences as unknown and (iii) we have PV readings which were used more often than others (cf. the QUANTIFICATION reading). We also found metaphorical and therefore non-prototypical PV usages, like in the case of *abnageln*.

References

Faaß G., Heid, U., and Schmid, H. (2010). Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Kliche, F. (2011). Semantic Variants of German Particle Verbs with *ab*. *Leuvense Bijdragen*, (97).

Kratzer, A. (2003). *The Event Argument and the Semantics of Verbs*. http://semanticsarchive.net/Archive/GU1NWM4Z/.

Lechler, A. and Roßdeutscher, A. (2009). Analysing German Verb-Particle Construction with *auf* in a DRT-based Framework. Technical Report 4, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Quantitative contribution to the study of the syntax of spoken vs. written language

Juliette Thuilier
Université Paris-Sorbonne & Alpage (INRIA – Paris Diderot)
juliette.thuilier@paris-sorbonne.fr

1 Introduction

In French, attributive adjectives (A) can appear both before or after the noun (N):

(1) a. une agréable soirée (anteposed)
a nice evening
b. une soirée agréable (postposed)
a evening nice
"a nice evening"

Except the presence of a post-adjectival dependent that imposes the postposition of the AP (2), this phenomenon is led by various factors interacting in a complex way and favoring one position over the other.

(2) a. une musique agréable à écouter
a music nice to hear
b. *une agréable à écouter musique
a nice to hear music
"a music nice to hear"

The aim of this paper is twofold: to model this alternation phenomenon and compare the difference between the syntax of spoken French (SF) and written French (WF) on the basis of this alternation phenomenon. The methodology is inspired by the work by Bresnan et al. (2007) and Bresnan and Ford (2010) on dative alternation in English. Using statistical modeling on data extracted from written and spoken corpora, we test syntactic factors found in the literature (Abeillé and Godard, 1999; Wilmet, 1981; Forsgren, 1978; Blinkenberg, 1933 a. o.).

2 Methodology

We assume that, with statistical tools (logistic regression – Agresti, 2007 – and mixed-effect models – Gelman and Hill, 2006), we are able to free ourselves from variations due to the sampling of the corpora. Moreover, one advantage of the mixed-effect logistic regression is that it is predictive, in the sense that one can build a model on a set of data and use this model to predict the choice between anteposition and postposition on unseen data. This way, we can evaluate how well the model generalizes from the training set. Lastly, we make use of the possibility of testing the significance of interaction between different factors in order to evaluate which syntactic factors have a different behavior according to the medium used (spoken vs. written).

2.1 The database

To build our database, we first extracted the attributive As that appeared in both positions in the syntactically annotated newspaper corpus *French Treebank* (FTB, Abeillé and Clément

2004), leaving aside As with post-adjectival dependents. We then extracted the same As from the spoken corpus C-ORAL-ROM (CORAL, Cresti and Moneglia 2005). Only 130 of them were found in the spoken corpus. Besides the variable capturing the medium used (SF vs. WF), these data were annotated for 10 variables concerning the syntactic environment of each A in context: (1) the A is coordinated, (2) the A is modified by an adverbial element; the NP contains (3) an other A in postposition, (4) a relative clause, (5) a PP; the determiner of the NP is (6) demonstrative, (7) possessive, (8) a definite article; a measure of collocation for (9) the ordered sequence A+N and (10) the ordered sequence N+A (collocations estimated with χ^2, Manning and Schütze, 1999). These variables are presented in Table 1[1].

VARIABLES	TYPE	DESCRIPTION
coordination	*bool*	the adjective is coordinated or not
modifier	*bool*	the adjective is pre-modified or not
demonstrativeDeterminer	*bool*	the NP is introduced by a demonstrative determiner or not
possessiveDeterminer	*bool*	the NP is introduced by a possessive determiner or not
definiteArticle	*bool*	the NP is introduced by a definite article or not
writtenCorpus	*bool*	the adjective is extracted from a written corpus or not
PP	*bool*	there is a PP in the NP or not
relativeClause	*bool*	there is a relative clause in the NP or not
otherPostposedA	*bool*	there is a postposed adjective in the NP or not
collocationAN	*real*	score for A+N bigram (log scale)
collocationNA	*real*	score for N+A bigram (log scale)

Table 1: Annotated variables in the database

3 Observations

The database contains 6612 occurrences of attributive As (4986 in FTB, 1626 in CORAL) representing 170 lemmas, with 68.9% of anteposition (67.1% in FTB, 74.3% in CORAL). There is variation according to the lemmas: for instance, the A *unique* 'unique' is anteposed in 20% of the cases, whereas *sérieux* 'serious' appears in this position in 51.4% and *petit* 'small' in 98.6%. Moreover, there is less alternation in spoken data than in written ones: the 170 lemmas appear in both positions in FTB, while only 56 (43.1% of the 130 lemmas) are really alternating in CORAL. This seems to reveal that in spoken French, the As tend to have a more fixed behavior than in the written variant. One can hypothesize that the more the speech is spontaneous, the more the A occurs in its preferred position, that is the more frequent position.

4 Multi-factorial statistical modeling

We used mixed-effects logistic regression to estimate the probability that the anteposition will be chosen as a function of 11 predictive variables (the 10 syntactic variables and the medium: written or spoken). The construction of the model consists in estimating the coefficients that are associated with each variable. Besides the predictive variables, also called fixed effects, mixed-effects models are able to take into account the variation in the data by means of random-effects. In our case, the adjectival lemmas are the random effects in order to model the adjectival idiosyncrasies. We built a model with 11 fixed-effects and 1 random-effect. We tested all the interactions between the medium and the 10 syntactic variables interactions. We removed predictors and interactions that were non-significant at the 0.05 level step by step, but keeping in the model non-significant fixed-effects for predictors that participated in significant interactions. The model is presented in Table 2. All

[1]We also differentiated two lemmas in context for 9 As: *ancien* 'ancient/former', *pur* 'pure', *seul* 'alone/single', *simple* 'simple/modest', *sacré* 'sacred/brilliant', *commun* 'ordinary/shared', *pauvre* 'poor/unfortunate', *propre* 'own/clean', *cher* 'expensive/dear'.

the fixed-effects are significant or participate in a significant interaction and thus participate in predicting the position of the As. The condition number of the predictors used in this model is $\kappa = 8.15$, which indicates that our data display low collinearity. This model has a mean accuracy of 0.882 (10-fold cross-validation) and the mean concordance probability is C = 0.947 (10-fold cross-validation). These numbers indicate that the model's predictions are very accurate.

5 Results

Each coefficient associated with fixed-effects can be interpreted as the preference for a position: a positive coefficient indicates a preference for anteposition and a negative one for postposition. Thus the model shows that the nature of the determiner influences the position: demonstrative, possessive and definite determiners favor the anteposition. Moreover, APs containing coordinated As or adverbial modifiers tend to be postposed, which confirms that speakers tend to put "heavy" APs after the N. The occurrence of a relative clause, a PP or another A after the N also favors the anteposition. Finally, the N the A is combined with affects the choice: the more the A and the N tend to be a collocation in a given order, as in *à juste$_A$ titre$_N$* 'understandably', the more the sequence tend to occur in the given order.

There is also a significant effect of the medium: SF favors postposition compared to WF. Moreover there are significant interactions between the media and three variables: the demonstrative determiner, the possessive one and the occurrence of a modifier. The interaction effects are represented in the plots (Figure 1). First, demonstratives slightly favor anteposition in SF, whereas in WF, they have a stronger effect. Second, the general tendencies observed for possessive determiners and modifiers are strengthened in SF: NPs introduced by possessives strongly favor anteposition; and the occurrence of a pre-adjectival modifier triggers the postposition of the AP in most of the cases.

RANDOM EFFECTS

GROUPS	NAME	VARIANCE	STD.DEV.
adjectival-lemma	(Intercept)	2.4124	1.5532

Number of obs: 6612, groups: adjectival-lemma, 170

FIXED EFFECTS

	ESTIMATE	STD. ERROR	Z VALUE	Pr(>|z|)
(Intercept)	-0.70679	0.18881	-3.743	0.000182
demonstrativeDeterminer=true	0.19508	0.49344	0.395	0.692588
possessiveDeterminer=true	2.07981	0.49308	4.218	2.46e-05
definiteArticle=true	0.36049	0.10687	3.373	0.000743
writtenCorpus=true	0.36915	0.13451	2.744	0.006063
coordination=true	-1.23054	0.26661	-4.616	3.92e-06
otherPostposedA=true	0.58555	0.15354	3.814	0.000137
PP=true	0.84350	0.10448	8.073	6.84e-16
relativeClause=true	0.70982	0.21166	3.354	0.000798
modifier=true	-2.73507	0.35934	-7.611	2.71e-14
collocationAN	0.37713	0.01849	20.396	< 2e-16
collocationNA	-0.44103	0.02000	-22.055	< 2e-16
demonstrativeDeterminer:writtenCorpus	1.29997	0.55833	2.328	0.019894
possessiveDeterminer:writtenCorpus	-1.11834	0.52371	-2.135	0.032727
modifier:writtenCorpus	0.99029	0.39445	2.511	0.012055

Table 2: Model parameters

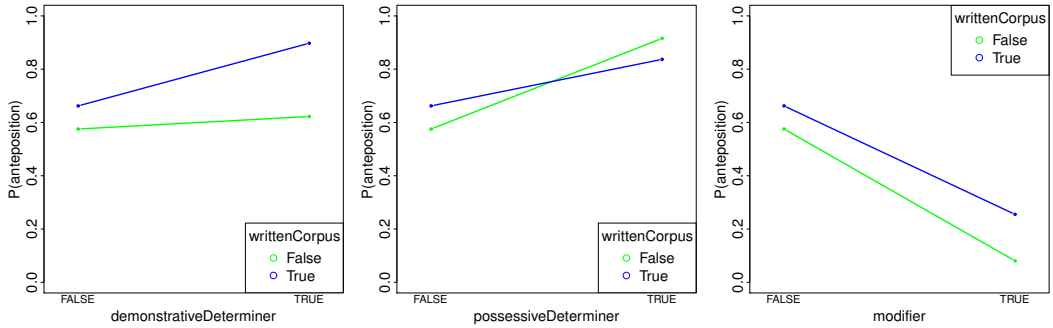

Figure 1: Partial effects of the 3 interactions

References

Abeillé and Godard. 1999. La position de l'adjectif epíthète en français : le poids des mots. *Recherches Linguistiques de Vincennes* 28: 9–32.

Agresti. 2007. *An Introduction to Categorical Data Analysis.* Wiley.

Behaghel. 1909. Von deutsher wortstellung. *Insogermanische Forschungen* 25.

Blinkenberg . 1933. *L'ordre des mots en français moderne. Deuxième partie.* Copenhague : Levin & Munksgaard.

Bresnan, Cueni, Nikitina, and Baayen. 2007. Predicting the dative alternation. In: Boume, Kraemer, and Zwarts (Eds.), *Cognitive Foundations of Interpretation*, Amsterdam : Royal Netherlands Academy of Science.

Bresnan and Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86 (1).

Forsgren. 1978. *La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique.* Stockholm: Almqvist & Wilksell.

Gelman and Hill. 2006. *Data Analysis Using Regression and Multilevel/ Hierarchical Models.* Cambridge: Cambridge University Press.

Hawkins. 1994. *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

Manning and Schütze. 1999. *Foundations of Statistical Natural Language Processing.* , Cambridge, MA: The MIT Press.

Wilmet. 1981. La place de l'épithète qualificative en français contemporain : étude grammaticale et stylistique. *Revue de linguistique romane* 45: 17–73.

Lectal conditioning of lexical collocations

Jose Tummers[1,2] Dirk Speelman[2] Dirk Geeraerts[2]
[1]Leuven University College [2]University of Leuven
jose.tummers@khleuven.be,
{dirk.speelman,dirk.geeraerts}@arts.kuleuven.be

1 Problem statement

The last decade, empirical linguistics focusing on genuine data has largely benefited from theoretical developments in Construction Grammar and from methodological and technical innovations in usage-based linguistics. In both frameworks, there is an obvious interest for lexical selectivity and idiomatic language use as part of the interplay between lexicon and grammar in probabilistic language models (Gries 2008). Lexical preference patterns are modeled along the paradigmatic axis, called collostructions (Stefanowitsch & Gries 2003, 2008), as well as the syntagmatic axis, called collocations (Sinclair 1991; Speelman et al. 2009; Wulff 2008), proving that the instantiation of constructions and constructional slots is at least partially conditioned by lexical selection restrictions.

Less attention has been paid to the lectal dimension of language use, referring to language external sources of variation. However, in a usage-based language model, the properties of the actual usage settings should be taken into account since they influence the language use (Geeraerts 2005). In this respect, Stefanowitsch & Gries (2008) explored the relation between register and collostructions.

In this contribution, we will focus on the lectal conditioning of lexical collocations. First, we will analyze how register and national variety modify the distributional properties of AN collocations in Dutch. Next, we will analyze how those lectal variables alter the impact of lexical collocations on the alternation between two inflectional variants of the adjective in Dutch definite NPs with a singular neuter head noun. In this NP construction, the adjective displays an alternation between the standard inflected form (1) and its marked uninflected counterpart (2):

(1) het vriendelijk-e kind
the friendly-INFL child
(2) het vriendelijk-ø kind
the friendly-ZERO child

Within the intricate network of variables governing this alternation, the lexical collocation strength of the AN pair exerts a major impact on the inflectional realization of the adjective, the use of the uninflected alternative being favored in AN collocations (Tummers 2005). Furthermore, the lectal variables hypothesized to modify the impact of lexical collocations on the adjectival inflection both have a significant effect on the choice of the inflectional alternative, the use of the uninflected adjective being favored by Belgian Dutch as well as informal registers in Belgian Dutch and (highly) formal registers in Netherlandic Dutch.

The following research questions will be addressed to disentangle the relation between lexical collocation strength on the one hand and the lectal variables on the other hand:

1. To what extent is the distribution of AN collocations in Dutch modified by register and national variety?
2. To what extent is the impact of AN collocations on the selection of the adjectival alternative in Dutch altered by register and national variety?

The answers to those questions will shed light on the relation between collocation strength on the one hand and the lectal variables on the other. Is there a consecutive

relationship between both, do they both act independently or do they act in mutual interaction?

2 Results and discussion

A database of 4,964 definite NPs with a singular neuter head noun (3,810 inflected and 1,154 uninflected adjectives) was extracted from the Corpus of Spoken Dutch (Oostdijk 2000). That repository of spoken Dutch contains data from Belgian and Netherlandic Dutch, the two national varieties, and various registers ranging from highly informal (colloquial speech) to highly formal (prepared speeches in parliament). The lexical collocation strength between A and N lemmas was computed using the log likelihood ratio, G^2 (Dunning 1993).

In answer to research question 1, figure 1 visualizes the G^2-distributions in the four different registers grouped by national variety, showing differences induced by both register and national variety. Moreover, the distribution shows a strong positive skew, yielding a lot of outliers which are not all included in the boxplots (range(G^2) = [0.00;1782.99]).

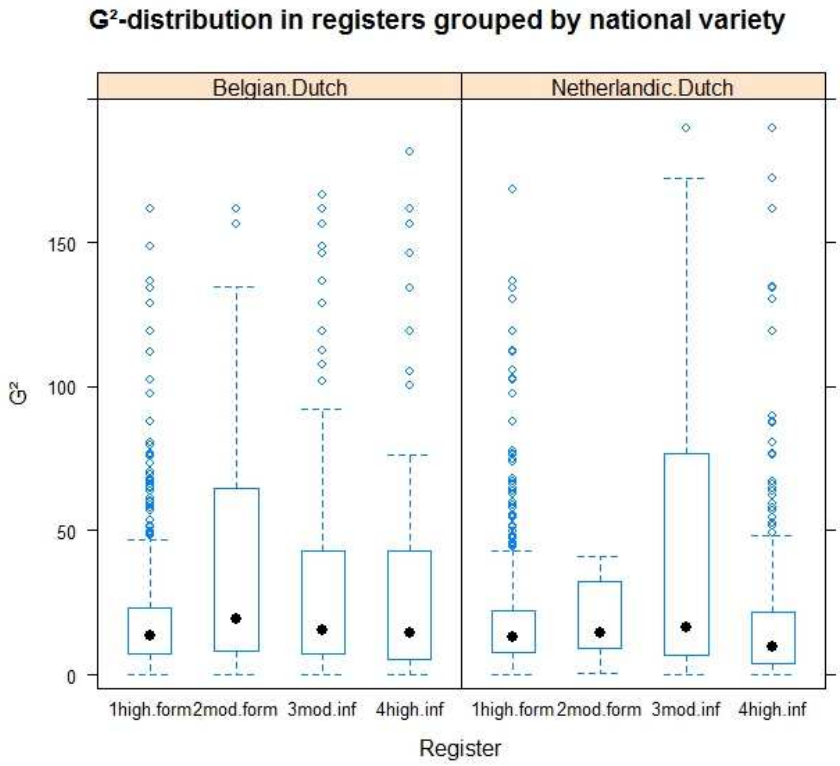

Figure 1: G^2-distribution over AN pairs in registers grouped by national variety

To model the impact of both lectal variables on the lexical collocation strength, viz. G^2, a gamma GLM has been fitted, G^2 displaying a Chi^2-distribution which in turn is a special case of the gamma distribution (Forbes et al. 2011). Table 1 presents the regression coefficients, both lectal variables (`nat.var`, `register`) being dummy coded.

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.041447	0.003003	13.804	< 2e-16 ***
nat.var=bel	-0.007020	0.003681	-1.907	0.056541 .
register=mod.form	0.009085	0.032179	0.282	0.777704
register=mod.inf	-0.032765	0.003203	-10.231	< 2e-16 ***
register=high.inf	-0.017535	0.003708	-4.729	2.32e-06 ***
nat.var=bel:register=mod.form	-0.028925	0.032297	-0.896	0.370511
nat.var=bel:register=mod.inf	0.016507	0.004426	3.730	0.000194 ***

nat.var=bel:register=high.inf	-0.001795	0.004793	-0.375	0.708046

Table 1: Gamma GLM modeling impact of national variety and register on G^2

Although no significant main effect of the national variety (`nat.var=bel`) is found, there is a significant interaction between register and national variety indicating a different stylistic conditioning of AN collocation patterns in both national varieties of Dutch.

To deal with research question 2, a logistic regression analysis has been performed (`rms` library in R, Harrell 2001) with $\ln(^{P(A.uninflected)}/_{1-P(A.uninflected)})$ as response variable and G^2 (`llr`), national variety (`nat.var`, dummy coding) and register (`register`, dummy coding) as explanatory variables (model statistics: likelihood ratio $Chi^2 = 648.11$, df = 15, $p < 0.0001$, C = 0.732). The regression coefficients (table 2) show an adjustment of the impact of the lexical collocation strength on the inflectional alternation by both lectal variables and their interaction**.**

Variable	**Coef**	**S.E.**	**Wald Z**	**Pr(>|Z|)**
Intercept	-2.3088	0.1113	-20.75	<0.0001
llr	0.0144	0.0020	7.25	<0.0001
nat.var=bel	0.5516	0.1352	4.08	<0.0001
register=mod.form	2.1823	1.0109	2.16	0.0309
register=mod.inf	1.3902	0.1699	8.18	<0.0001
register=high.inf	0.2149	0.1680	1.28	0.2008
llr:nat.var=bel	-0.0075	0.0023	-3.29	0.0010
llr:register=mod.form	-0.0642	0.0508	-1.26	0.2069
llr:register=mod.inf	-0.0101	0.0021	-4.74	<0.0001
llr:register=high.inf	-0.0033	0.0026	-1.26	0.2059
nat.var=bel:register=mod.form	-1.4334	1.0215	-1.40	0.1605
nat.var=bel:register=mod.inf	-0.0682	0.2170	-0.31	0.7532
nat.var=bel:register=high.inf	1.2169	0.2277	5.35	<0.0001
llr:nat.var=bel:register=mod.form	0.0627	0.0509	1.23	0.2178
llr:nat.var=bel:register=mod.inf	0.0054	0.0026	2.11	0.0350
llr:nat.var=bel:register=high.inf	-0.0001	0.0032	-0.02	0.9811

Table 2: Logistic regression modeling the impact of G^2, national variety and register on inflectional alternation attributive adjective

First, the impact of the collocation strength on the selection of the uninflected adjective is significantly lower in Belgian than in Netherlandic Dutch (reference value). Next, the effect of the collocation strength on the selection of the uninflected adjective in the moderately informal register (`llr:register=mod.inf`) is significantly lower than for the most formal register (reference value) and the other registers. Finally, the propensity of AN collocations to select the uninflected adjective in the moderately informal register, as compared to the most formal register, is significantly higher in Belgian than in Netherlandic Dutch, as can be inferred from the significant triple interaction (`llr:nat.var=bel:register=mod.inf`).

3 Conclusion

In sum, lexical collocation strength and lectal sensitivity operate in mutual interaction. First, the lexical collocation strength in AN pairs is subject to lectal adjustments. Second, the selection criteria of the adjectival alternatives use lexical collocation strength in a different way depending on the lectal settings, as national variety, register and their interaction significantly constrain the effect of lexical collocation strength on the inflectional variation. Hence, we argue that a comprehensive usage-based language model needs to include a lectal dimension. In this respect, we refer to Cognitive Linguistics, where the recognition of the importance of lectal constraints on language use resulted in Cognitive sociolinguistics (Geeraerts et al. 2010).

References

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1), 61-74.

Forbes, C., M. Evans, N. Hastings, and B. Peacock. 2011. *Statistical Distributions*. Hoboken: John Wiley & Sons.

Geeraerts, D. 2005. Lectal variation and empirical data in Cognitive Linguistics. In: F. J. Ruiz de Mendoza Ibáñez, and M. S. Peña Cervel (eds.), *Cognitive Linguistics and Interdisciplinary Dynamics*, 163-190. Berlin: Mouton de Gruyter.

Geeraerts, D., G. Kristiansen, and Y. Peirsman (eds). 2010. *Advances in cognitive sociolinguistics*. Berlin: Walter de Gruyter.

Gries, S. Th. 2008. Phraseology and linguistic theory: a brief survey. In: S. Granger, and F. Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 3-25. Amsterdam: John Benjamins.

Harrell, F. E. 2001. *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*. New York: Springer.

Oostdijk, N. 2000. Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 5 (3), 280-284.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Speelman, D., J. Tummers, and D. Geeraerts. 2009. Lexical patterning in a construction grammar. The effect of lexical co-occurrence patterns on the inflectional variation in Dutch attributive adjectives. *Constructions and Frames*, 1 (1), 87-118.

Stefanowitsch, A., and S. Th.Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209-243.

Stefanowitsch, A., and S. Th. Gries. 2008. Channel and constructional meaning: A collostructional case study. In: R. Dirven, and G. Kristiansen (eds.), *Cognitive Sociolinguistics: Language variation, Cultural Models, Social Systems*, 129-152. Berlin: Mouton de Gruyter.

Tummers, J. 2005. *Het naakte adjectief. Kwantitatief-empirisch onderzoek naar de adjectivische buigingsalternantie bij neutra*. PhD dissertation, KULeuven, Faculty of Arts.

Wulff, S. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. London/New York: Continuum Press.

Translation-driven mapping of semantic fields: the case of Dutch and French inceptive verbs

Lore Vandevoorde Gert De Sutter Koen Plevoets
Ghent University
{Lore.Vandevoorde, Gert.DeSutter, Koen.Plevoets}@UGent.be

The (semi-)automatic retrieval of semantically similar words has become of increasing importance to lexical semantics and lexical variation studies (e.g., Peirsman et al., 2010), which has led to the advent of vector-based approaches like Latent Semantic Analysis (Landauer and Dumais, 1997), first and second order bag-of-words models (Manning and Schütze, 1999) and the behavioral profiles method (Divjak and Gries, 2006, 2009). These models are generally characterized as distributional, which means that they capture word meaning in relation to their context in large corpora. The present corpus-based study will put forward the use of translational data and translation corpora as another reliable way of retrieving semantic relationships and mapping them in semantic fields, complementary to the Semantic Vector Spaces. In computational linguistics, the use of more than one language to identify lexical relationships has already proved to be a successful way of resolving problems of ambiguity (Dagan et al., 1991) and as a cross-lingual solution for word sense disambiguation (Lefever, 2012). Drawing on Dyvik's (2004, p. 311) assumption that "semantically closely related words ought to have strongly overlapping sets of translations", overlapping sets of translations should commensurably reveal the semantic relations between translations, between translations and their source language items and, more importantly, between the source language items themselves. The present corpus-based study will map out the semantic field of the Dutch inceptive verb BEGINNEN and its most salient French translation COMMENCER by creating semantic fields through what we will call "back-and-forth translation". The method is carried out as follows: first, all translations of a given (set of) lexeme(s) in a large corpus are checked manually. Then, inversely, all translations of these translations back into the initial source language (we will call this "back-translations") are looked up. These so-called back-translations enable us to access the structure of the semantic field of the initial (set of) lexeme(s) via the first-order translations, thus guaranteeing a verification of the initial set of lexemes as well as a broadening of the semantic field without any (word class) restriction imposed by the initial selection of lexemes.

We used translational data extracted from the Dutch Parallel Corpus, a ten-million-word parallel and comparable corpus, balanced with respect to five text types and four translation directions (Macken, et al., 2011). In order to generate the semantic field of BEGINNEN, a concise set of near-synonyms was selected, consisting of *beginnen, aanvangen, een aanvang nemen, starten, van start gaan* and *aanvatten*. We based our selection on lexicographic data and inter-annotator substitution testing. The French translations of this set of onomasiological variants of BEGINNEN (n=528) were manually checked, returning a total of 17 different translations. Then, the 17 translations were inversely queried from the corpus as source-language lexemes. The Dutch translations of this set (n=1563) yielded 47 translations back into Dutch. The French variants of COMMENCER (*entamer*, *démarrer* and *débuter*) were submitted to the same procedure: manual checking (n=253) returned 8 different translations that were subsequently translated back into French, returning 31 back-translations (n=1393).

The resulting frequency tables were analyzed with the technique of correspondence analysis (Greenacre, 2007; Lebart et al., 1998). Correspondence analysis arrives at a lower-dimensional representation of the row and column categories, analogous to a Semantic Vector Space. More specifically, the multidimensional data structure of the 6 variants for BEGINNEN with 17 French translations were approximated in 2 dimensions, thus mapping out the semantic field of BEGINNEN. The 47 back-translations were subsequently projected onto this space as so-called "supplementary points". The rationale of this approach is that the projected back-translations do not reveal onomasiological but semasiological clusters. The same procedure was followed for the variants of COMMENCER.

The Dutch back-translations of BEGINNEN and the French back-translations of COMMENCER were plotted into two separate graphs, depicting their respective semantic fields. We observe that both graphs show lexemes clustering together. Figure 1 shows that most lexemes are in the plot's origin, e.g. *beginnen* [to begin], *meteen* [right away], *ten eerste* [firstly], *aanvang* [onset]. This cluster can consequently be interpreted as the prototypical center, consisting of lexemes with the basic meaning of the inceptive category, viz. "start of a general process". A second cluster appears slightly to the right of the central cluster, consisting of lexemes like *starten* [to start], *lanceren* [to launch], *op gang brengen* [to bring about], *starter* [starter], *actief* [active]. They generally refer to the "starting up of a business, a company, a medical treatment or an interpersonal relation". More to the right, we find *opzetten* [to set up], *invoeren* [to establish], *instellen* [to set up] and *in werking treden* [become effective], mostly referring to a "rule or legislation becoming effective". The outlying cluster (bottom left of the origin) consists of *aanvangen* [to start], *een aanvang nemen* [to commence], *sluiten* [to close] and *ingaan* [to take effect], commonly appearing in texts provided by governmental instances and usually referring to a "lease or hire agreement taking effect". When looking at Figure 2 for COMMENCER, we find a fairly similar graph to the one of BEGINNEN with most lexemes around the origin (e.g. *se lancer* [to dive into], *commencer* [to begin], *départ* [start] and *d'abord* [first(ly)]) and two separate clusters: one to the right of the central cluster with lexemes like *démarrer* [to start], *lancer* [to launch] and *mettre sur pied* [to set up], usually referring to the "beginning of a project, an initiative or a business" and an outlying cluster bottom left of the origin with lexemes commonly referring to the "initiation of a legal situation" (e.g. *prendre effet* [become effective], *prendre cours* [to take effect], *aborder* [to bring up]).

As appears from our results, translational data are an interesting source for the bottom-up identification of a semantic field's structure and for the differentiation of prototypical meanings from peripheral ones. Although we are not able yet to compare our translational approach with the distributional approaches mentioned above, the creation of semantic fields on the basis of translational data appears to have several advantages. Firstly, it does not require complex annotating techniques making it far less time consuming than some other (distributional) methods. Secondly, our translation-driven approach yields a semantic field with different word classes, which broadens the structure of the generated semantic fields, a strategy not often adopted by distributional models. Finally, our method provides an opportunity for a straightforward, cross-linguistic comparison of semantic fields.

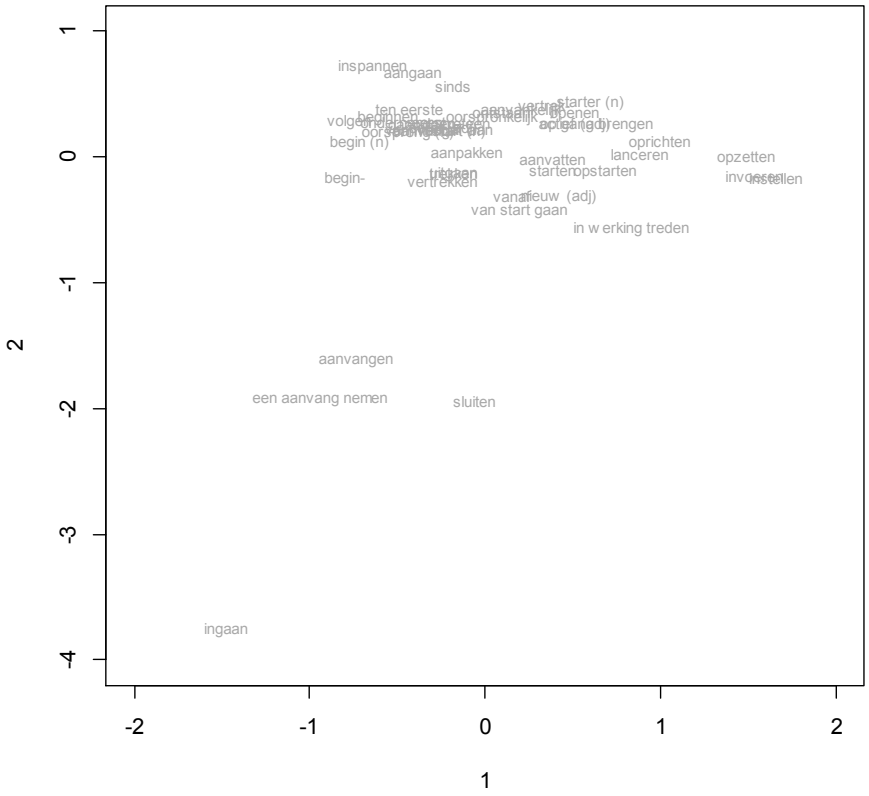

Figure 1: Semantic field of BEGINNEN

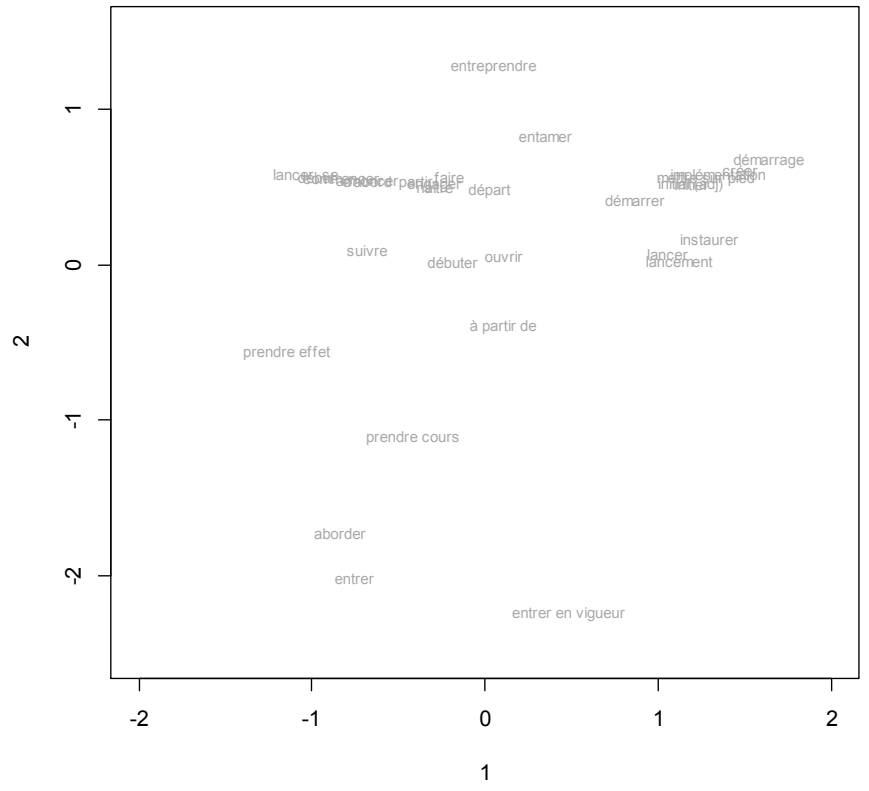

Figure 2: Semantic field of COMMENCER

References

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In: *Proceedings of the 29th annual meeting of the Association for Computational Linguistics*, 130--137.

Divjak, Dagmar, and Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering Behavioral Profiles. *Corpus Linguistics and Linguistic Theory*, 2 (1): 23--60.

Divjak, Dagmar, and Stefan Th. Gries. 2009. Corpus-based cognitive semantics: a contrastive study of phasal verbs in English and Russian. In: Lewandowska-Tomasczyk and Dziwirek (eds.), *Studies in Cognitive Corpus Linguistics*, 273--296. Frankfurt am Main: Peter Lang.

Dyvik, Helge. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. In: Aijmer and Altenberg (eds.), *Advances in Corpus Linguistics*, 311--326. Amsterdam and New York: Rodopi.

Greenacre, Michael. 2007. *Correspondence analysis in practice*. Boca Raton: Chapman and Hall/CRC.

Landauer, Thomas, and Susan Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104 (2): 211--240.

Lebart, Ludovic, André Salem and Lisette Berry. 1998. *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

Lefever, Els. 2012. *ParaSense: parallel corpora for word sense disambiguation.* Ghent University: Ghent.

Macken, Lieve, Orphée De Clercq and Hans Paulussen. 2011. Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. *Meta*, 56 (2): 374--390.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Peirsman, Yves, Dirk Geeraerts and Dirk Speelman. 2010. The Automatic Identification of Lexical Variation between Language Varieties. *Journal of Natural Language Engineering*, 16 (4): 469--491.

Evaluating cluster quality in Semantic Vector Space

Thomas Wielfaert Kris Heylen Jakub Kozakoszczak
Leonid Soshinskiy Dirk Speelman
University of Leuven
{thomas.wielfaert,kris.heylen,dirk.speelman}@arts.kuleuven.be
{leonid.soshinskiy,jakub.kozakoszczak}@student.kuleuven.be

1 Visual analysis of distributional models

In recent years, distributional models of semantics have become the mainstay of large-scale modelling of lexical semantics in Computational Linguistics (see Turney and Pantel 2010 for an overview). These vector-based approaches also hold a large potential for research in Linguistics proper: They allow linguists to base their analysis on large amounts of usage data, thus vastly extending their empirical basis, and they make it possible to detect potentially interesting patterns of how lexical meaning is contextually realised.

So far, there have been relatively few applications of distributional vector models in theoretical linguistics, mainly because of the technical complexity and the lack of a linguist-friendly interface to explore the output, so that they remain largely black boxes. Heylen et al. (2012) made a first attempt to open up the Semantic Vector Spaces for linguistic investigation through interactive visualisations of the semantic similarities between usage instances (word tokens) that are identified by a distributional model. A lexicologist can peruse a 2D represenation together with the concordances and a colour coding of explanatory variables (e.g. region or register). Wielfaert et al. (2013) extends this approach by visualising multiple models in the same interface and adding meta-information about the extent to which specific context features influence the model's output. However, with a large number of different distributional models, it is not feasible for a linguist to compare all their visualisations and assess how well or which type of semantics the models capture. Therefore, this paper introduces two quantitative measures to evaluate the quality of distributional models directly and systematically against an expert's analysis of semantic structure. We evaluate this measure both on Dutch and English data.

2 Quantitative evaluation

One of the strengths of Semantic Vector Spaces is their parameter-richness, which allows to define distributional contexts in many different ways. One can for instance use a window-defined bag-of-words approach or contexts filtered by syntactic dependencies. One can vary the size of the context window, include or exclude function words, filter by part-of-speech, assign weights to context features by collocational strength etc. Each of these parameter settings gives a lexicologist a different perspective on the data and can capture different types of contextually determined lexical semantics. However, at the same time, this parameter-richness is also the largest weakness of Semantic Vector Spaces because the number of possible solutions grows exponentially with the number of parameters that is varied. As a consequence, a lexicologist cannot arrive at an overall assessment of how all these different parameters settings affect the type of semantics captured by distributional models. Although a 2D representation makes it possible to visually compare one specific model's output with a human expert's analysis, as the number of solutions grows, it becomes indispensible to have a measure that can reliably quantify how well many different distributional models corroborate or contradict the researcher's hypothesis. Our aim is therefore to develop a

measure that enables a systematic, large-scale comparison of model outputs against expert analyses.

In Computational Linguistics, token-level distributional models are typically used in Word Sense Disambiguation (WSD) tasks and their evalution is based on a human "gold standard" in the form of manually disambiguated concordances. For the evaluation of our measure we make use of similar sense-classified data sets. For English, we use the test set from the SemEval 2010 Word Sense Induction & Disambiguation task. However, these data sets typically distinguish dictionary-style, lexicographic senses that do not cover all the semantic distintinctions that theoretically inspired lexicologists are interested in. Therefore we also created a finer grained, lexicologically annotated evaluation dataset of a Dutch polysemous noun (monitor).

In computational WSD, identifying semantic structure is seen as a clustering problem where tokens have to be assigned to the 'correct' word sense. The output of a distributional model (a semantic similarity matrix) is therefore submitted to a clustering algorithm, and, following traditional practice in Information Retrieval, the cluster solutions are in their turn evaluated in terms of purity, normalised mutual information, Rand index and F measure (Manning et al., 2008). As linguists however, we are not interested in an evaluation that depends on a specific cluster algorithm; rather, we want to evaluate directly how well a lexicologist's analysis of semantic structure is present in the distributional models' output. We have experimented with two such direct quality measures. The first one, 'cluster quality' is taken from Speelman and Geeraerts (2008) and is very similar to the McClain-Rao clustering index (McClain and Rao, 1975). The basic idea is that for each token we calculate the ratio between the within-cluster and between-cluster distances to other tokens and then aggregate over all tokens. For the distance measure we either use 1 minus the cosine similarities that are outputed by the distributional model, or the Euclidean distances between coordinates after dimension reduction of the cosine similarity matrix with nonmetric Multidimensional Scaling (isoMDS), which is the technique used in the 2D visualisations described above. The lower this ratio of within-cluster and between-cluster distances, the better the 'cluster quality'.

Because 'cluster quality' relies heavily on distances, extreme outliers have the potential to bias the result. Therefore, we implemented a second measure we call 'k-nearest neighbour quality'. Here, the idea is that in a good model, tokens should be mainly surrounded by tokens that belong to the same sense cluster. If we take the k-nearest tokens and divide the number of tokens belonging to the same cluster by k, we get the percentage of neighbouring tokens with the same sense. Again, we aggregate over all tokens to get the 'k-nearest neighbour quality'. With this measure, a good cluster solution is represented by a number approaching 1 (100% or perfect quality).

Both quality measures were applied to the output of a range of differently parametrized distributional models for the English and Dutch disambiguated data sets. The quality rankings by the measures were then compared to the quality assessment by a human expert that scrutenized the visualisation of the different models. Both quality measures result in similar model rankings that, in their turn, by-and-large correspond to the linguistic assessments. However, the measures do react slightly differently to specific parameter settings.

References

Heylen, Kris, Dirk Speelman and Dirk Geeraerts. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In: *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Language Patters and Uncovering Language History from Multilingual Resources*, 16–24.

Manning, Christopher D., Prabhakar Raghavan and Schütze, Hinrich. 2008. *Introduction to Information Retrieval.* New York: Cambridge University Press.

McClain, John O. and Vithala R. Rao. 1975. Clustisz: A program to test for the quality of clustering of a set of objects. In: *Journal of Marketing Resaerch*, Vol 12 (4): 456–460.

Speelman, Dirk, and Dirk Geeraerts. 2008. The Role of Concept Characteristics in Lexical Dialectometry. *International Journal of Humanities and Arts Computing* 2 (1-2): 221–242.

Turney, Peter D. and Patrick Pantel. 2010. Looking at word meaning. From Frequency to Meaning: Vector Space Models of Semantics. In: *Journal of Artificial Intelligence*, Vol 37: 141–188.

Wielfaert, Thomas, Kris Heylen and Dirk Speelman. 2013. Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique. In: *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, 154–166, Sables d'Olonne.

A cognitively grounded measure of pronunciation distance

Martijn Wieling[1,3] Jelke Bloem[2] John Nerbonne[3] R. Harald Baayen[1,4]
[1]University of Tübingen, [2]University of Amsterdam, [3]University of Groningen, [4]University of Alberta
wieling@gmail.com, j.bloem@uva.nl, j.nerbonne@rug.nl, harald.baayen@uni-tuebingen.de

1 Introduction

Obtaining a suitable difference measure (i.e. distance) between two pronunciations is important, not only for dialectologists who are interested in finding the relationship between different dialects (e.g., Heeringa, 2004), but also for language researchers investigating the relationship between the world's languages (e.g., Bakker et al., 2009), or those investigating second language acquisition (e.g., Flege et al., 2006). Obtaining distances between word pronunciations enables quantitative analyses in which the effect of various factors can be investigated.

A commonly used automatic measure of pronunciation distance is the Levenshtein distance (Levenshtein, 1965) which calculates the minimal number of insertions and deletions to transform one phonetically transcribed string into the other. While improvements have been proposed to make the method phonetically sensitive (Wieling et al., 2012), it suffers from two important drawbacks. The first is that there is no cognitive basis for using the Levenshtein distance as a pronunciation distance measure. The second is that the Levenshtein distance does not allow asymmetric distances (generally characterizing perceptual pronunciation distances; Gooskens and Heeringa, 2004).

Here we propose a new method, Naive Discriminative Learning (NDL; Baayen et al., 2011), which does not suffer from these drawbacks. The idea behind this approach (grounded in human learning theory; Rescorla and Wagner, 1972) is that we model how well a listener understands meaning when listening to a speaker with a certain accent. In this model, the past experience (i.e. exposure to speech) of a listener shapes how well a word's phonetic cues activate meaning by means of association strengths. When a cue is present together with a certain word (representing meaning), their association strength increases, whereas it decreases when the cue is present but the word (meaning) is not.

After determining the association strengths between all cues and meanings in the network of an adult listener (Danks, 2003; using the R-package 'ndl'), the activation of a meaning for a specific set of cues is calculated by summing the corresponding association strengths. For example, when the listener model is based on native American English (AE) speech, we can compare the activation of a certain meaning for a set of cues on the basis of a native as opposed to a non-native AE pronunciation. Presumably the non-native cues (through lower association strengths) will give rise to a lower activation of the meaning compared to fully native cues. If we then calculate the difference between these activations, we obtain a measure of distance between the two pronunciations of a word.

2 Material

The Speech Accent archive (Weinberger and Kunath, 2011) is digitally available at http://accent.gmu.edu and contains a large sample of speech samples in English from people with various language backgrounds. Each speaker reads the same paragraph of text in English (containing 69 words, of which 55 are unique).

All speech samples are transcribed according to the International Phonetic Alphabet, and the associated audio files are available. In 2010, we extracted all available 989 transcribed samples and their audio from the Speech Accent Archive. In this study, we use a subset consisting of all 115 native U.S.-born English speakers (used as the native reference pronunciations) and 286 mostly non-native speakers for whom we obtained foreignness ratings.

3 Methods

To obtain the NDL-based network of association strengths representing a native AE listener, we randomly selected 58 native AE speakers whose pronunciations were converted to cues (i.e. trigrams of sound segments, including markers representing word boundaries) for the corresponding meanings. As the association strength between a cue and a meaning will obviously depend on the relative frequency with which they co-occur, we extracted word frequency information from the Google N-Gram Corpus (Brants and Franz, 2009). We then constructed the model, yielding a network of association strengths (representing a native AE listener).

Using this network, we first determined the activation of each meaning when supplying the phonetic trigram cues of the remaining 57 native AE speakers. These activations were averaged (across speakers) in order to estimate how well an average native AE speaker is understood by our simulated native AE listener for each meaning separately. In similar fashion, we calculated how well each of the 286 speakers is understood by our simulated native AE listener (for each meaning). To determine the NDL-based pronunciation distance between each individual speaker and the average native AE speaker, we simply calculated the difference between their activations averaged across all meanings.

4 Results

To determine how well these NDL-based pronunciation distances matched perceptual distances we developed a questionnaire in which participants listened to 50 different speech samples and rated their native-likeness (on a scale from 1 to 7). As the questionnaire was advertised in a post on Language Log by Mark Liberman, more than 1100 native AE speakers participated, resulting in at least 50 ratings per speech sample (Cronbach's alpha: 0.85). The Pearson correlation between the perceptual native-likeness ratings and log-transformed NDL-based pronunciation distances was $r = -0.82$ ($p < 0.001$). These results are comparable to those using the Levenshtein distance (Wieling et al., submitted), which is also illustrated by the high correlation between the two types of computational distances: $r = 0.89$ ($p < 0.001$).

5 Discussion

The high correlation between the perceptual native-likeness ratings and NDL-based pronunciation distances indicates that our new measure indeed captures pronunciation distances. While the Levenshtein distance offers comparable performance and is computationally efficient (it takes about 10 seconds, compared to 50 seconds for the complete NDL procedure), it has no cognitive basis supporting a link with perceptual pronunciation distances, and it does not allow asymmetric distances (such as those reported by Gooskens and Heeringa, 2004). NDL does not suffer from these drawbacks, and as it also allows for the inclusion of non-segmental cues (such as intonation markers), it is a promising alternative to the Levenshtein distance (which does not allow the inclusion of non-segmental information).

References

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118, 438–482.

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., Egorov, D., Mailhammer, R., Grant, A., and Holman, E. W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1), 169–181.

Brants, T. and Franz, A. (2009). Web 1T 5-gram, 10 European Languages. Version 1. Linguistic Data Consortium, Philadelphia.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, 47, 109–121.

Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H. and Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 34, 153-175.

Gooskens, C. and W. Heeringa (2004). Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. *Language Variation and Change*, 16(3), 189-207.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163, 845–848. In Russian.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*, New York: Appleton-Century-Crofts, pp. 64–99.

Weinberger, S. H. and S. A. Kunath (2011). The Speech Accent Archive: towards a typology of English accents. *Language and Computers*, 73, 265-281.

Wieling, M., Bloem, J., Mignella, K, Timmermeister, M. and Nerbonne, J. (submitted). Automatically measuring the strength of foreign accents in English.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2), 307-314.

Distance-to-V, length and verb disposition effects on PP placement in Belgian Dutch. A corpus-based multifactorial investigation

Annelore Willems Gert De Sutter
University of Ghent
{annelore.willems, gert.desutter} @ugent.be

The present paper reports on a corpus-based multifactorial investigation of PP placement in Dutch subordinate clauses. Language users fundamentally have the choice to put PPs either before V-final (midfield position) or after V-final (postfield position). Consider example 1 and 2:

(1) dat de trainer$_{[subject]}$ door een laptop naast het veld$_{[PP]}$ vervangen wordt$_{[V\text{-}final]}$
that the coach$_{[subject]}$ by a laptop next to the field$_{[PP]}$ replaced is$_{[V\text{-}final]}$

(2) dat de trainer$_{[subject]}$ vervangen wordt$_{[V\text{-}final]}$ door een laptop naast het veld$_{[PP]}$
That the coach$_{[subject]}$ replaced is$_{[V\text{-}final]}$ by a laptop next to the field$_{[PP]}$
'That the coach is replaced by a laptop next to the field'

On the basis of the corpus materials in the journalistic component of the Dutch Parallel Corpus (Macken et al. 2011), which yielded 1900 manually verified attestations, we performed a mixed effects model with PP position (midfield vs. postfield) as binary response variable and *distance-to-V*, *distance between V and the end of the clause* and *length of the PP* as fixed predictor variables and *verb lexemes* as a random predictor variable. The fixed predictor variables were respectively operationalized as the number of words[1] between subject and verb, the number of words between the verb and the end of the clause and the number of words of the PP. First, a mixed effects model was fitted with only main effects. This model (table 1) shows a highly significant result for the *length of the PP* and the *distance between V and the end of the clause*, and no significant result for the *distance-to-V*. Based on this model, we can conclude that longer PPs tend to occur more in the postfield position (cf. the principle of end-weight; Wasow 2002) and that PPs occur less often in the postfield if there are other elements after V-final (thus countering the end-weight effect). To interpret the effect of the random factor (individual verbs), we performed a distinctive collexeme analysis (Gries & Stefanowitsch 2004). This analysis shows that complex verbs stimulate postfield position.

	Variance	Std. Dev.
Verbs	0.86	0.93
	Odds ratio	p-value
Length PP	2.94	<2e-16 ***
Distance-to-V		0.494
Distance between V and end of sentence	0.25	<2e-16 ***

Table 1

In a second model, we add the interaction effect between the two significant main effects (table 2). Although the interaction is significant, the main effect *distance between V and the end of the clause* disappears. Model diagnostics show that this is due to multicollinearity (vif > 8).

[1] We also measured the length of the three fixed factors in terms of syllables. High correlations were attested for all three factors between the operationalization in terms of words and syllables (resp. 0.93, 0.97 and 0.94).

	Variance	Std. Dev.
Verbs	0.88	0.94
	Odds ratio	p-value
Length PP	6.47	4.55e-13 ***
Distance-to-V		0.48
Distance between V and end of sentence		0.24
Interaction length PP and Distance between V and end of sentence	0.52	0.0001 ***

Table 2

In order to avoid this problem, we decided to perform a third model with one significant main effect, *length of the PP*, and the interaction effect, which does not suffer from multicollinearity. The interaction effect entails that the linear trend to place long PPs in the postfield is suppressed somewhat for long PPs in postfields that are not empty (see Figure 1). The predictive power of this third model is good (c=86).

	Variance	Std. Dev.
Verbs	0.87	0.93
	Odds ratio	p-value
Length PP	8.47	<2e-16 ***
Interaction length PP and Distance between V and end of sentence	0.41	<2e-16 ***

Table 3

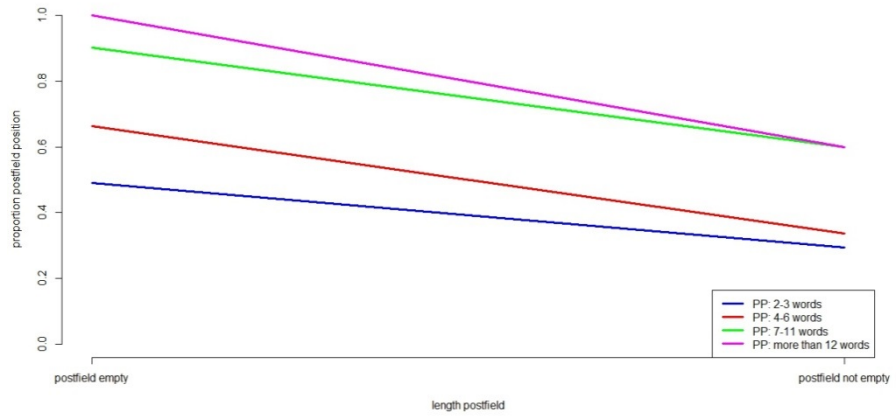

Figure 1: plot of the interaction effect

On the basis of these results, we are able to confirm the multifactorial nature of PP placement in Dutch, as is already often shown for other types of syntactic variation (e.g., Grondelaers 2000, Gries 2003, De Sutter 2005). More interestingly, our results can be used to refine two common assumptions in traditional Dutch syntactic theory and psycholinguistic theories of sentence comprehension:

(1) the structural position before V-final (midfield) in Dutch subordinate clauses (example 1) is not the standard slot for PPs. In Dutch syntax, the midfield position is generally considered to be the standard position for PPs, while the postfield position is often seen as an obvious alternative to release an overstrained midfield. If the brace construction becomes too heavy, language users tend to narrow down the amount of information between the two poles of the brace construction by shifting the PP to the end of the sentence. This view is elaborated in the *Algemene Nederlandse Spraakkunst* (1997), and by many generative grammarians, who explain the variation by an optional movement transformation to the right (e.g. Koster 1974, Jansen 1979). Our results, however, show that postfield position is more often preferred than midfield position (58% vs. 42%), even in circumstances where the midfield is not overladen, thereby refuting the overstrained midfield assumption.

(2) the distance between subject and V is not to be reduced as much as possible, as our data show that subject and verb in subordinate clauses are mostly not adjacent. As a consequence, psycholinguistic theories as Gibson's Dependency Locality Theory (2000)

needs to be nuanced, at least for Dutch. Gibson proposes that structures with shorter dependencies are preferred and easier to process (Gibson 2000, Temperley 2006). However, our data show that Dutch language users do not strive at maximally reducing the distance between subject and verb, as could be expected on the basis of Gibson's theory.

ANS: Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., van den Toorn, M.C. (red.). 1997. *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk. Groningen: Martinus Nijhoff / Deurne: Wolters Plantyn.

De Sutter, G. 2005. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. Unpublished PhD University of Leuven.

Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In: Miyashita, Y., Marantz, A., & O'Neil, W. (Eds*.), Image, language, brain*, 95-126. Cambridge, MA: MIT Press.

Gries, S. Th. 2003. *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London & New York: Continuum Press.

Gries, S. Th., Stefanowitsch, A. 2004. Extending Collostructional Analysis: a corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, 9: 97-129.

Grondelaers, S. 2000. *De distributie van niet-anaforisch* er *buiten de eerste zinsplaats. Sociolexicologische, functionele en psycholinguïstische aspecten van* er*'s status als presentatief signaal*. Unpublished PhD University of Leuven.

Jansen, F. 1979. On tracing conditioning factors of movement rules: extraposition of PP in spoken Dutch. In: M. van de Velde en W. Vandeweghe (Hrsg.), *Sprachstruktur, Individuum und Gesellschaft*, 83-89. Berlin: Niemeyer.

Koster, J. 1974. Het werkwoord als spiegelcentrum. *Spektator*, 3: 601-618.

Macken, L., Declercq, O., Paulussen, H. 2011. Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. *META*, 56(2): 374-390.

Rietveld, T., van Hout, R. 2005. *Statistics in Language Research: Analysis of Variance*. Berlin: Mouton de Gruyter.

Temperley, D. 2006. Minimization of dependency length in written English, *Cognition*, doi:10.1016/j.cognition.2006.09.011

Wasow, T. 2002. *Postverbal Behavior*. CSLI Publications.