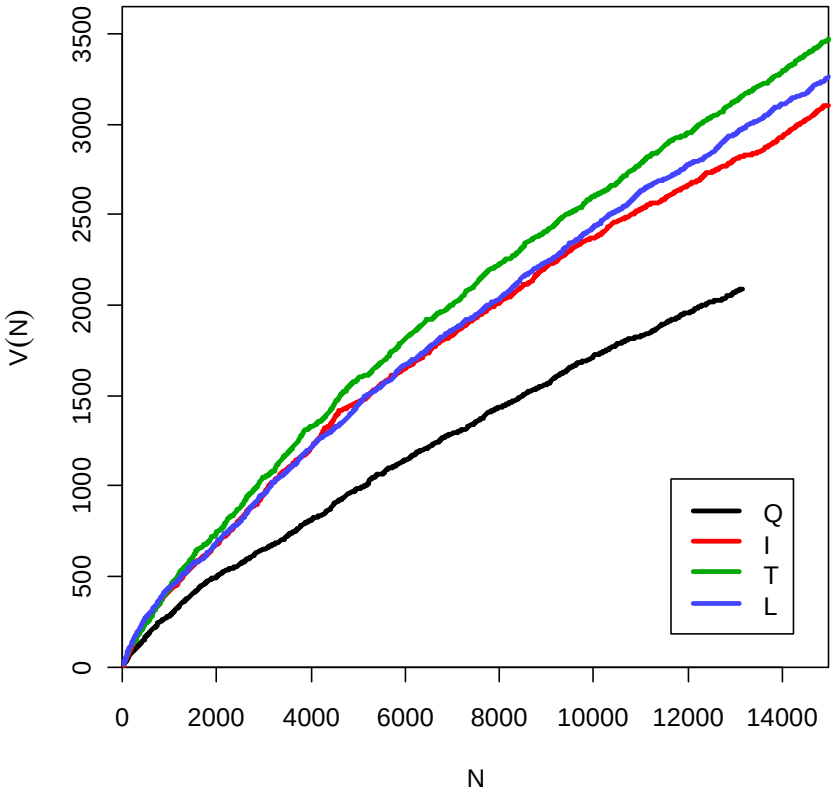
Proceedings of Quantitative Investigations in Theoretical Linguistics 4

QITL-4



29-31 March 2011
Humboldt-Universität zu Berlin

Edited by Amir Zeldes and Anke Lüdeling

Proceedings of Quantitative Investigations in Theoretical Linguistics 4
29-31 March 2011
Humboldt-Universität zu Berlin
http://korpling.german.hu-berlin.de/qitl4/

Invited Speakers

Heike Behrens, University of Basel
Hermann Moisl, University of Newcastle upon Tyne
Ingo Plag, University of Siegen

Program Committee

Antti Arppe, University of Helsinki
Harald Baayen, University of Alberta
Marco Baroni, University of Trento/CIMeC
Douglas Biber, Northern Arizona University
Peter Bosch, University of Osnabrück
Jeremy Boyd, University of Illinois
Michael Cysouw, Ludwig Maximilian University of Munich
Mark Davies, Brigham Young University
Stefan Engelberg, Institute for the German Language (IDS), Mannheim
Stefan Evert, University of Osnabrück
Livio Gaeta, University of Naples Federico II
Stefan Th. Gries, University of California, Santa Barbara
Jennifer Hay, University of Canterbury
Graham Katz, Georgetown University
Tibor Kiss, Ruhr University Bochum
Tony Kroch, University of Pennsylvania
Jonas Kuhn, Institute for Natural Language Processing, University of Stuttgart
Roland Meyer, University of Regensburg
Adam Przepiórkowski, Institute of Computer Science, Polish Academy of Sciences
Katharina Spalek, Humboldt University of Berlin
Ruprecht von Waldenfels, University of Bern
Stefanie Wulff, University of North Texas

Organizing Committee

Anke Lüdeling, Humboldt University of Berlin
Amir Zeldes, Humboldt University of Berlin

Proceedings of Quantitative Investigations in Theoretical Linguistics 4

29-31 March 2011
Humboldt-Universität zu Berlin

Edited by Amir Zeldes and Anke Lüdeling

Contents

Foreword

It gives us great pleasure to be able to host the fourth Quantitative Investigations in Theoretical Linguistics (QITL-4) in Berlin, under the auspices of the Institute for German Language and Linguistics at Humboldt University. The previous QITL conferences in Osnabrück and Helsinki remain among the most enjoyable and enriching conferences on empirical linguistics in our memory, and we are excited about the contributions the conference has attracted, covering some of the most important, original and interesting work that has arisen in the interim. Because of our wish to stay true to the one-session-only concept of the previous QITL conferences, out of some 80 submissions we could only accept 28 presentations: 15 talks and, following the introduction of a poster session in Helsinki, 13 poster presentations. In addition, our three invited speakers have graciously agreed to hold plenary talks on their work in different fields of quantitative linguistic research.

In planning the conference we have attempted to represent the best, methodologically and theoretically most diverse contributions from those contemporary researchers in linguistics who not only delve into data to begin with, but also use their findings to inform theoretical models of the language faculty. Our declared goal has been to bridge the gap between the theory of grammar and language data, in a way that enriches both without compromising either one. We firmly believe and hope that QITL has and will continue to serve this purpose.

This conference could not have been organized without the help of very many people. Special thanks are due to the members of the program committee in dealing with the sizable load of reviewing needed to shape the final program. We are also indebted to our colleagues and students for helping out at the conference, as well as to other members of the institute and the university who helped in organizing conference materials, printing this volume, providing technical support for the conference, refreshments and other forms of logistic support. Finally, we would like to thank all participants in the conference for contributing to the discussion on quantitative research in theoretical linguistics.

Berlin, March 2011,

Anke Lüdeling and Amir Zeldes

Invited Talks

Heike Behrens
University of Basel

What counts? The graded nature of rule abstraction in child language

Quantitative corpus linguistics plays a decisive role in usage-based, constructivist approaches to language acquisition. First, evidence is needed that the input available to the child provides the necessary information for abstracting the target language system(s). Second, the distributional properties of the language input and its uptake by the learner help to identify the learning mechanisms, especially with regard to refining the relative contribution of functional factors and perceptual factors like salience and frequency. Recent research has shown that the input available to the child highlights a subset of the linguistic system in which the basic morphological and syntactic properties are available in a rather consistent fashion. The findings give rise to a third perspective in quantitative acquisition research: what does the child generalize over? If we assume that children do not have access to adult-like categories, and do not command the full range of options of a particular structure, we have to define how to conceive of the intermediate levels of acquisition. In this presentation I will use extensive corpus data on German child language to show how German noun inflection (plural classes, declension classes) is acquired through a gradual generalization over the phonotactic properties of the noun root. This results in acquisition sequences that are largely error-free although or because the underlying generalizations are rather local.

Hermann Moisl
University of Newcastle upon Tyne

Corpus-based generation of linguistic hypotheses using quantitative methods

In linguistics as in other sciences, the de facto standard methodology is based on Karl Popper's concept of the falsifiable hypothesis, whereby a hypothesis is proposed in answer to a research question about some domain of interest and then tested by observation of the domain. Because of their centrality, it is natural to ask how hypotheses are generated.

The consensus in philosophy of science is that hypothesis generation is non-algorithmic, that is, not reducible to a formula, but is rather driven by human intellectual creativity in response to a research question using a combination of deductive inference from existing axioms and theorems, and inductive inference of generalizations from observation of the domain. Hypothesis generation by deductive inference has long been dominant in generative linguistics and, to a lesser extent, in other subdisciplines like historical linguistics and sociolinguistics. The advent in recent decades of large amounts of digital electronic text amenable to computational analysis has, however, made hypothesis generation by inductive inference viable. This paper shows how mathematical and statistical techniques such as cluster analysis and singular value composition can be used for that purpose.

Ingo Plag
University of Siegen

Compound Stress Assignment Emerges from the Lexicon

In both psycholinguistic and theoretical-linguistic circles there is a debate about the nature and role of symbolic rules, associative networks, and analogical or exemplar-based models in the organization of language in general, and of morphology and the lexicon in particular. Compounding has featured prominently in this debate and the present paper complements this line research by focusing on an aspect of semi-regularity in compounds that had received little attention until a few years back: stress assignment in English noun-noun compounds.

English noun-noun compounds are traditionally assumed to be subject to the Compound Stress Rule, which states that the left-hand constituent is more prominent than the right-hand constituent. However, not all English noun-noun compounds abide to the Compound Stress Rule (cf. for example, *ópera glasses*, *wátch-maker* and *Óxford Street* versus *steel brídge*, *morning páper*, and *Madison Ávenue*). Rightward stress is far from exceptional, and the nature of the observable variability is still under debate.

In this paper I will argue that compound stress assignment emerges from the lexicon. We carried out a number of experimental and corpus-based quantitative and computational studies of a large number of English compounds that show that deterministic, rule-based approaches are not very successful in predicting compound stress. In contrast, it is possible to use the individual properties of stored compounds and different types of lexical relatedness between these compounds to successfully model compound stress assignment. This modeling can be done with the help of different kinds of multiple regression or by using analogical computational algorithms. Depending on the respective data set, these models vary in their predictive accuracy, but generally outperform deterministic models.

Our results are in line with the bulk of more recent psycholinguistic research on lexical processing and are impossible to account for under a rule-based theory of compound stress.

Presentations

Alena Anishchanka, Natalia Levshina, Dirk Speelman, Dirk Geeraerts
University of Leuven, QLVL (Quantitative Lexicology and Variational Linguistics)
Alena.Anishchanka@arts.kuleuven.be, Natalia.Levshina@arts.kuleuven.be,
Dirk.Speelman@arts.kuleuven.be, dirk.geeraerts@arts.kuleuven.be

Shades of salience: Multivariate analysis of prototypicality effects in color terms

Research question

The proposed study aims to develop a rigorous quantitative method to measure the prototypicality effects in the semantics of lexical units, using color terminology in American advertising as the source of data. We approach this goal from two perspectives. Firstly, we analyze the psycholinguistic techniques developed in basic color categorization studies for measuring salience (basicness) of color terms and compare them to corpus-based measurements of color term salience. Secondly, we analyze a number of color terms using a bottom-up quantitative procedure for measuring their relative salience based on formal linguistic properties. Following the usage-based approach to semantics, we specifically avoid the *a priori* distinction between more prototypical (basic) and less prototypical (non-basic) color terms. Using the convergent evidence from psycholinguistic and corpus-based measurements, we suggest that a number of techniques developed in corpus-driven studies could be added to the arsenal of measurements applied in the basic color term (BCT) tradition. This would allow extending the analyses and generalizations to include a wider range of color terms and categories.

Background

Over the last half century, anthropological, psycholinguistic and linguistic studies of color categorization have developed a number of quantitative experimental procedures for measuring the psychological and linguistic salience (basicness) of color terms (see, for instance, Berlin & Kay 1969, Rosch Heider 1972, Rosch & Mervis 1975, MacLaury 1997, Corbett & Davies 1997). Many of these techniques such as reaction time and consistency in (color) naming experiments, frequency and ranking in elicitation tasks, frequency, derivational potential and length in text corpora have been successfully applied to other semantic domains in search of prototype representations. However, in the color domain itself the discussed measurements have been mostly limited in their application to basic color term analyses.

Alternatively, the recent corpus-based studies of color terms demonstrate that their basicness or salience is a matter of degree and cannot be reduced to the dichotomy of basic vs. non-basic color terms (Kerttula 2002) and that it might be affected by contextual factors (Steinvall 2002). Furthermore, the comparison of different behavioral and linguistic tests (Corbett & Davies 1997) demonstrates that not all of them are equally successful in distinguishing between basic and non-basic color terms. These findings give evidence to an internal hierarchy and graded membership in color categories in line with the multivariate model of semantics developed in Cognitive

Linguistics (Geeraerts et al. 1994, Gries 2003, Glynn 2007, Geeraerts 2010). They might also suggest the heterogeneous nature of salience that includes linguistic, categorical, and cultural levels (Geeraerts 2006).

Data and variables

The study is based on an extensive self-compiled corpus of color names and color samples used by US manufacturers and retailers in online marketing in four product categories (automobiles, clothing, make-up, and house paints).

For the purposes of the presented analyses, we randomly selected a sample of 16400 observations equally representing the four product categories. The data were annotated for a range of quantified formal and corpus-based linguistic parameters intended to operationalize the salience of color terms (see Table 1). For the analysis of correlation between corpus-based and experimental parameters of salience we used measurements applied in psycholinguistic color naming and elicitation experiments (Boynton & Olson 1990, Corbett & Davies 1997, Taft & Sivik 1997, Sturges & Whitfield 1997) (see Table 1).

TABLE 1. MEASUREMENTS OF SALIENCE.

Formal linguistic characteristics
length of the color term in alphanumeric characters
length of the color term in constituents
length of the color term in phonemes
length of the color term in syllables
Corpus-based characteristics
token frequency
type token frequency ratio
usage as an independent color name
attraction to the head or modifier slot in compound color terms as a simple ratio
attraction to the head or modifier slot in compound color terms as a log-likelihood ratio
the size of the color category according to the distribution of its exemplars in rgb color space
(un)restricted distribution across different product categories
derivational productivity based on type frequency of derived color terms
derivational productivity based on token frequency of derived color terms
Psycholinguistic parameters in color naming experiments
response time in milliseconds
consistency of naming
Psycholinguistic parameters in color name elicitation experiments
relative frequency
sequence rank
Evolutionary parameters
position in the evolutionary BCT hierarchy suggested by Berlin and Kay

The results of statistical analyses

In the first part of the study, we have tested the relationships between corpus-based and experimental parameters of salience using Kendall tau rank correlation coefficient and hierarchical cluster analysis based on the data available for 34 color categories.

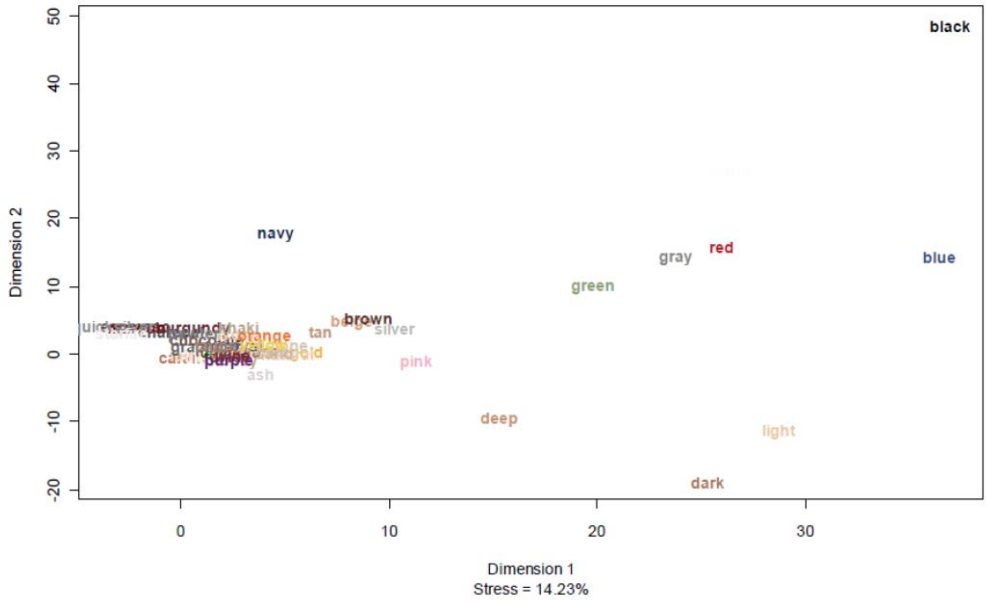
The results suggest several groups of salience parameters. The purely formal measurements of color term length show the highest correlation and form a distinct cluster. Corpus-based measurements are distributed in two clusters interrelated with other parameters. In one cluster several corpus-based characteristics including derivational productivity and usage in the head position correlate with Berlin and Kay

evolutionary sequence. In the other cluster such corpus-based measurements as token frequency and independent usage group with psycholinguistic measurements suggesting convergence of experimental and corpus evidence.

In the second part of the study, we performed a series of Kruskal's non-metric multidimensional scaling analyses exploring the relative salience of color terms used in advertising. The results confirm that basicness or salience of color terms is a continuous non-homogeneous parameter rather than a dichotomy. We can observe a continuum from the traditionally recognized BCT (*green, blue, red*) followed by the so-called secondary BCT (*orange, purple, pink*) through well entrenched non-basic monolexemic names (*wine, burgundy, charcoal)* to compounds with color modifiers (*light blue*) and finally to the most idiosyncratic compound terms (*deep sea blue, titanium silver*).

If we further zoom in on monolexemic color terms (see Figure 1), we can observe that most of the primary BCT (*black, blue, red, green*) are more distinct in their linguistic behavior compared to secondary BCT (*pink, orange, brown, purple*) and especially to very densely clustered non-basic color terms. At the same time, certain non-basic color terms (*silver, tan, beige*) come close in their linguistic characteristics to secondary BCT suggesting a continuum between basic and non-basic color terms.

FIGURE 1. CORPUS-BASED SALIENCE OF MONOLEXEMIC COLOR TERMS.



The proposed analyses make several contributions to the studies of linguistic salience and color categorization. Firstly, the comparison of corpus-based and experimental measurements of salience suggests both convergence of the two paradigms and the distinct place of corpus-based parameters. This can be seen as evidence of multidimensional nature of linguistic salience and prototypicality effects that create the basis for conceptual variation. Secondly, the chosen granularity of the analyses specifically addresses the gap between the categorization studies focusing on the most salient (basic) color categories and idiosyncratic color terms hand-picked for the studies of color terms in advertising (Stoeva-Holm 1996, Bergh 2007, Graumann 2007, Wyler 2007). Finally, considering the potentially unlimited source of data, the proposed model allows developing generalizations on a larger scale than has been suggested in the

previous basic color categorization analyses and the studies of color terms in advertising.

References

Bergh, Gunnar 2007. The semiosis of Swedish car color names: Descriptive and amplifying functions. In *Anthropology of Color: Interdisciplinary Multilevel Modeling,* MacLaury, Robert, Paramei, Galina and Dedrick, Don (eds.), 337-345. Amsterdam: John Benjamins.

Berlin, Brent and Kay, Paul 1969. *Basic Color Terms: Their Universality and Evolution.* Berkeley: University of California Press.

Boynton, Robert and Olson, Conrad 1990. Salience of chromatic basic color terms confirmed by three measures. *Vision Research* 30(9), 1311-1317.

Corbett, Greville and Davies, Ian 1997. Establishing basic color terms: measures and techniques. In *Color categories in thought and language,* Hardin, Clyde and Maffi, Luisa (eds.), 197-223. Cambridge: Cambridge University Press.

Geeraerts, Dirk 2006. Salience phenomena in the lexicon. A typology. In Geeraerts, Dirk *Words and Other Wonders. Papers on Lexical and Semantic Topics* [Cognitive Linguistics Research 33], 76-98. Berlin: Mouton de Gruyter.

Geeraerts, Dirk 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk, Grondelaers, Stef and Bakema, Peter 1994. *The Structure of Lexical Variation. Meaning, Naming, and Context* [Cognitive Linguistics Research 5]. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2007. Mapping meaning: Towards a usage-based methodology in Cognitive Linguistics. PhD dissertation, Katholieke Universiteit Leuven.

Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement* [Open Linguistic Series]: London/New York: Continuum Press.

Graumann, Andrea 2007. Color names and dynamic imagery. In *Speaking of colors and odors* [Converging Evidence in Language and Communication Research 8]*,* Plümacher, Martine and Holz, Peter (eds.), pp.129-140. Amsterdam: John Benjamins.

Kerttula, Seija 2002. *English Colour Terms: Etymology, Chronology, and Relative Basicness*. [Mémoires de la Société Néophilologique de Helsinki], Vol. LX. Helsinki: Société Néophilologique.

MacLaury, Robert 1997. Color and Cognition in Mesoamerica: Constructing Categories as Vantages. Austin: University of Texas Press.

Rosch Heider, Eleanor 1972. Universals in color naming and memory. *Journal of Experimental Psychology* 93(1), 10-20.

Rosch, Eleanor and Mervis, Carolyn 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* 7, 573-605.

Steinvall, Anders 2002. English Color Terms in Context. PhD dissertation, Umeå University: Skrifter från moderna språk 3.

Stoeva-Holm, Dessislava 1996. Farbbezeichnungen in deutschen Modetexten. Eine morphologisch-semantische Untersuchung. [Studia Germanistica Upsaliensia, 34]. PhD dissertation, Uppsala: Universitetsförlag Almqvist & Wiksell.

Sturges, Julia and Whitfield, T.W.Allan 1997. Salient Features of Munsell Colour Space as a Function of Monolexemic Naming and Response Latencies. *Vision Research* 37(3), 307-313.

Taft, C. and Sivik, L. 1997. Salient color terms in four languages. *Scandinavian Journal of Psychology*, 38, 29-34.

Wyler, Siegfried 2007. Color terms between elegance and beauty: The verbalization of color with textiles and cosmetics. In *Speaking of colors and odors* [Converging Evidence in Language and Communication Research 8], Plümacher, Martine and Holz, Peter (eds.), pp.113-128. Amsterdam: John Benjamins.

R. Harald Baayen, Antti Arppe
University of Alberta, University of Helsinki
baayen@ualberta.ca, antti.arppe@helsinki.fi

Statistical classification and principles of human learning

In the application of any statistical analysis method to the modeling of linguistic phenomena, a recurring question is how to understand the statistical results from a cognitive perspective. Although quantitative models may provide detailed and useful insights into which factors enhance the probability of particular linguistic phenomena, they tend leave unanswered how actual speakers come to learn and use their language in the way they do.

The present study addresses this question by introducing a new, parameter-free model for linguistic choice behavior based on naive discriminative learning that is driven fully and only by the distributional properties of its input. The learning principles on which this model is based, the so-called Rescorla-Wagner equations (see appendix), were first proposed by Wagner and Rescorla in 1972 (Wagner & Rescorla 1972), and have proved to be amazingly fruitful in psychology as a model for human and animal learning (Miller, Barnet, & Grahame 1995). A technical innovation due to Danks (2003) makes it possible to estimate the weights of the Rescorla-Wagner equations when learning has reached a state of equilibrium. Baayen, Milin, Filipović Durdević, Hendrix, and Marelli (submitted) incorporated the equilibrium equations of Danks (2003) into a general discriminative learning model that is naive in the sense of naive Bayes classifiers. These authors show that naive discriminative learning provides accurate predictions of response latencies in the visual lexical decision task. The model reproduces a wide range of effects in the morphological processing literature with a minimum of representational assumptions, using a learning engine that, in its simplest form, has no free parameters.

In this paper, we pit this parameter-free statistical engine derived from human learning principles against several well-established statistical classifiers: random forests (Breiman 2001; Strobl, Malley, & Tutz 2009), support vector machines (Vapnik 1995), memory-based learning (Daelemans & Bosch 2005) and polytomous logistic regression (according to the one-vs.-rest heuristic, see e.g. Arppe (2008)).

As our linguistic example case, we have selected the near-synonymous set of the four most frequent Finnish verbs denoting THINK, namely *ajatella*, *miettiä*, *pohtia*, *harkita* ‘think, reflect, ponder, consider’, which have been comprehensively studied by Arppe (2008) using newspaper and Internet newsgroup discussion corpora. Altogether 3,404 occurrences of these four THINK verbs and their sentential contexts were analyzed in terms of their morphological and lexical as well as syntactic structure (following Functional Dependency Grammar, (Tapanainen & Järvinen 1997)), supplemented with semantic and structural subclassifications. Of some 6000 contextual features, 46 were selected for the present study, as these 46 emerged from previous analyses as the most predictive ones when taken together. This subset of predictors included the most common morphological properties and general semantic characteristics of the verb-chain in which the think verb occurred, and detailed information on the syntactic structure (functional roles and various subclassifications) linked with the think verbs in their sentential context. Arppe (2008) observed that using polytomous logistic regression (with any of several common heuristics) as a classifier seems to reach a ceiling at a Recall rate of roughly two-thirds of the sentences in the research corpus. The

results could not be substantially improved with the addition of further granularity in semantic and structural subclassification of the syntactic roles, and effectively similar results were obtained when partially varying (even randomly) the selection of contextual features.

TABLE 1. CLASSIFICATION DIAGNOSTICS FOR FIVE MODELS FITTED TO THE FINNISH DATA SET (N = 3404).

	$\lambda_{prediction}$	$\tau_{classification}$	Recall (proportion correct)
polytomous logistic regression	0.368	0.488	0.645
support vector machine	0.334	0.461	0.626
memory-based learning	0.286	0.422	0.599
random forests	0.326	0.455	0.621
naive discriminative learning	0.349	0.473	0.634

The classification results for the four statistical models and the naive discriminative learning model are summarized in Table 1. The measure for proportionate reduction of prediction error, $\lambda_{prediction}$, tells us how much better the models perform by using the selected set of predictors compared to what would be achieved by systematically selecting the most frequent verb in the data, while the measure for proportionate reduction of classification error, $\tau_{classification}$, informs us how much better the models reproduce, in the long run, the actually occurring proportions of verbs evident in the data, in comparison to the baseline case of homogeneous proportionate distribution (Menard 1995). As the results for random forests may change slightly across different runs, also a mean Recall =0.622 was estimated for a series of 50 random forests (range: 0.617-0.626). From Table 1 we learn that polytomous regression performs best, followed by naive discriminative learning. A proportions test comparing the top two models suggests they have equivalent recall. What we can conclude at this point is that discriminative learning performs as well as other established classifiers, at least on this data set. Interestingly, naive discriminative reading achieves this level of accuracy without a single free parameter, and therefore provides the theoretically most parsimonious fit of all models surveyed here.

At a high level of abstraction, each of the five models surveyed above provides a good characterization of a Finnish native speaker's knowledge of the optimal choice of a think verb given morphological, syntactic and other contextual information. Although roughly equivalent in terms of predictive accuracy, it is only memory-based learning and naive discriminative learning that have some cognitive plausibility — we believe it is unlikely that the brain would actually be searching for support vectors, that it would be estimating beta weights, or that it would be constructing forests of conditional inference trees. Memory-based learning is an attractive paradigm for probabilistic inference in language processing, that is in many ways compatible with usage-based and exemplar-based approaches. A potential disadvantage of memory-based learning is that it requires vast amounts of memory for the exemplars, combined with on-line computations on nearest neighbor sets. By contrast, discriminative learning assumes that the adult competence is the result of a long process of discriminative learning. This model is extremely sparse in the number of representations and connections required: for the present data set, all we need is 4 representations , one for each of the think verbs, 46 representations for the binary predictor values, and 4 ∗46 = 148 connection weights. The support for a particular verb given the input is calculated straightforwardly by summation of the weights on the connections linking the input predictors to the verb.

Exemplar knowledge is not stored explicitly in the form of 3404 exemplar vectors, but implicitly in just 148 connection weights. We hypothesize that naive discriminative learning implements the simplest possible mathematical characterization of probabilistic linguistic competence, compatible with the insight that grammar is usage-based, but without assuming that usage is calculated over an entire exemplar space. Instead, we assume that usage is acquired piecemeal in a much simpler weight space.

References

Arppe, A. 2008. *Univariate, Bivariate and Multivariate Methods in Corpus-based Lexicography. A Study of Synonymy*. Helsinki: Department of General Linguistics, University of Helsinki.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P. and Marelli, M. submitted. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning.

Breiman, L. 2001. *Random forests. Machine Learning* 45, 5-32.

Daelemans, W. and Bosch, A. Van den. 2005. Memory-based Language Processing. Cambridge: Cambridge University Press.

Danks, D. 2003. Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology* 47(2), 109-121.

Menard, S. 1995. Applied logistic regression analysis. Thousand Oaks: Sage Publications.

Miller, R. R., Barnet, R. C. and Grahame, N. J. 1995. Assessment of the Rescorla-Wagner model. *Psychological Bulletin* 117(3), 363-386.

Strobl, C., Malley, J. and Tutz, G. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods* 14(4), 26.

Tapanainen, P. and Järvinen, T. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, April 1997, 64-71. Washington, D.C.: Association of Computational Linguistics.

Vapnik, V. 1995. The nature of statistical learning theory. Berlin: Springer-Verlag.

Wagner, A. and Rescorla, R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F. (eds.), *Classical Conditioning II*, 64-99. Appleton-Century-Crofts.

Appendix

Let PRESENT(X, t) denote the presence of a cue (predictor value) or outcome (one of the four Finnish THINK verbs) X at time t, and ABSENT(X, t) denote its absence at time t. The Rescorla-Wagner equations specify the association strength V_i^{t+1} of cue Ci with outcome O at time $t + 1$ using a recurrence equation, as follows:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{1}$$

The change in association strength ΔV_it defined as

$$\Delta V_i^t = \begin{cases} 0 & \text{if ABSENT}(Ci,t) \\ \alpha_i\beta_1(\lambda - \Sigma_{\text{PRESENT}(Cj,t)} Vj) & \text{if PRESENT(Cj,t) \& PRESENT(O, t)} \\ \alpha_i\beta_2(0 - \Sigma_{\text{PRESENT}(Cj,t)} Vj) & \text{if PRESENT(Cj,t) \& ABSENT(O, t)} \end{cases} \tag{2}$$

The equilibrium equations of Danks (2003),

$$\Pr(O|Ci) - \Pr(Cj|Ci)Vj = 0 \tag{3}$$

make it possible to estimate the weights for an 'adult' system by solving the above set of equations using the co-occurrence vector of a specific outcome (verb) given the different predictor values and the co-occurrence matrix of predictor values.

Markus Bader, Jana Häussler
University of Konstanz, University of Potsdam
markus.bader@uni-konstanz.de, jana.haeussler@uni-potsdam.de

From graded ratings to binary decisions: A case study on argument alternations in German

The growing use of experimental methods in linguistics has led to a widespread recognition that grammaticality is a gradient phenomenon. With regard to language performance, this raises several important questions: (i) How does graded grammaticality relate to language comprehension? In particular, is graded grammaticality a consequence of the fact that the perception of grammaticality is mediated by the language comprehension mechanisms? (ii) How does graded grammaticality relate to language production? In particular, do production frequencies determine degree of grammaticality? (iii) What are the processes leading to traditional binary grammaticality judgments?

We present a formal model addressing all three questions. The major claims of our model are as follows: (1) Graded grammaticality is a property of the grammar itself (e.g., Pater 2009) and not an epiphenomenon caused by performance mechanisms (e.g., Hawkins 2006). (2) During language production, the probability of a sentence is determined inter alia by its continuous grammaticality score. That is, in contrast to many usage-based models (see Manning 2003), we claim that graded grammaticality determines usage and not the other way round. (3) During language comprehension, sentences are automatically assigned continuous grammaticality values. In contrast to models denying a direct relationship between graded and binary judgments (Featherston 2005), we hypothesize that, when required by the task, continuous values are mapped directly onto binary grammaticality judgments in the way of Signal Detection Theory (Green & Swets 1966).

The evidence for this model comes from a series of experiments and an associated corpus study. The experiments investigated identical sentence materials with different judgment procedures: magnitude estimation (ME) and binary grammaticality judgments (BGJ). The sentence materials covers 120 ditransitive verbs in six syntactic contexts according to the factors *Voice* (active, regular passive and *bekommen* ('get') passive) and *Number of arguments* (2 or 3 arguments). In the following, we concentrate on the *bekommen* passive which is known to be subject to lexical variation (see (1)).

(1) *Ich glaube, dass Peter ein Buch geschenkt bekam / ??gestohlen bekam.*
I believe that Peter a book presented got stolen got.
'I believe that Peter was presented/stolen a book.'

The corpus study is based on the *deWaC* corpus described in Baroni et al. (2009), a huge corpus of German built by web crawling. With about 1,278,177,539 tokens, the deWaC corpus is of a size sufficient for present purposes. At the same time, this corpus

avoids the disadvantages brought about by querying the web directly by means of one of the popular search engines.

The experiments and the corpus study confirm this lexical variation for the 120 verbs in the *bekommen* passive:

- ME (z-transformed acceptability scores): range = -1.63 – 2.15
- BGJ (percentages of grammatical judgments): range = 19% - 100%
- Corpus frequencies: range = 0-2500

The direct mapping of graded grammaticality scores to binary grammaticality judgments is confirmed by a logistic regression analysis predicting BGJ results from ME scores which revealed a reasonably good fit (C =.79). The claim that gradient grammaticality is not caused by performance mechanisms is based on the finding of gradient judgments for sentences which are closely matched with respect to syntactic complexity, length and construction frequencies.

Several pieces of evidence support the claim that grammaticality is among the factors determining frequency and not the other way round. First, high grammaticality was found to be independent of frequency whereas low grammaticality was consistently associated with low frequency. Second, a poisson regression analysis predicting syntactic frequencies from verb frequencies and ME scores showed that both predictor variables lead to significant and substantial reductions in deviance. Third, we run a preference experiment along the lines of Bresnan & Ford (2010). Participants read a context question establishing a topic and then had to choose between two syntactic realizations of the answer (active voice versus *bekommen* passive). When the underlying dative object was established as topic (see (2)), the preferred answer was the *bekommen* passive variant which allows the topic to be realized as a subject.

(2) Was gibt es Neues von eurem Opa? ('What's about your grandpa?')

a. Topic = subject (*bekommen* passive)
Unser Opa hat von Robert einen Rasenmäher überreicht bekommen.
our grandpa has by R. a lawnmower handed-over got
'Our grandpa was handed over a lawnmower by Robert.'

b. Topic = dative object (active)
Unserem Opa hat Robert einen Rasenmäher überreicht.
our grandpa has R. a lawnmower handed-over
'To our grandpa, Robert handed over a lawnmower.'

Importantly, the choice of *bekommen* passive answers was modulated by the verb-specific grammaticality scores estimated from the prior judgment experiments. For verbs that had received degraded grammaticality scores in the *bekommen* passive, the *bekommen* passive answer was chosen less often. Frequency information also had an influence, although a smaller one. For verbs that were judged as fully compatible with the *bekommen* passive, the probability of choosing the *bekommen* passive variant was higher for verbs with higher corpus frequency for the *bekommen* passive.

References

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano and Zanchetta, Eros 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal* 23, 209-226.

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

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

Green, David M. and Swets, John A. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hawkins, John 2006. Gradedness as relative efficiency in the processing of syntax and semantics. In *Gradience in grammar: Generative perspectives*, Fanselow, Gisbert, Féry, Caroline, Vogel, Ralf and Schlesewsky, Matthias (eds.), 207-226. New York: Oxford University Press.

Manning, Christopher D. 2003. Probabilistic syntax. In *Probabilistic linguistics*, Bod, Rens, Hay, Jennifer and Jannedy, Stefanie (eds.), 289-341. Cambridge, MA: MIT Press.

Pater, Joe 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33, 999-1035.

Danielle Barth
University of Oregon
dmathieu@uoregon.edu

Modeling reduction of *is, am* and *are* in grammaticalized constructions

Introduction

Grammaticalization is a type of language change characterized by the creation of a grammatical element from a lexical element or another grammatical element. A classic example is the development of the English future auxiliary *will* from the lexical verb meaning 'want'. In the grammaticalization literature, phonetic reduction is often listed as a potential, although certainly not necessary, consequence to grammaticalization. For example, the auxiliary *will* can reduce to *'ll* as in (1), but the lexical verb cannot as in (2).

(1) *I'll see you later,*
(2) **I'll it to be so.*

However, phonetic reduction can be due to many factors, including frequency of use (Bybee 2007), low semantic weight (Heine 1993, Lehmann 1995), lexical class (van Bergem 1995) and context of use, etc. While changes in all of these parameters are associated with grammaticalization, they do not always go together. This paper examines a case where the less grammaticalized source construction is more frequent than the more grammaticalized target construction.

The Copular, Progressive and Passive Constructions

This paper examines the particular case of *be*. The forms *is, am* and *are* can reduce to *'s, 'm* and *'re* in both the lexical source construction of the copula as the main verb (3) and two grammaticalized constructions of progressive (4) and passive (5) as the auxiliary.

(3) *She's a welder.*

(4) *She's working.*
(5) *She's seen.*

The copular source construction is more frequent than the grammaticalized progressive and passive constructions as seen in a search in the Corpus of Contemporary American English (COCA) (Davies 2008-). All three variants could be characterized as having very little semantic weight because they are all grammatical elements. So this leads to the research question: Does one of these three construction types exhibit reduction more often, and if so which? And to follow up: what factors could influence the reduction of *is, am* and *are*? Following grammaticalization theory would predict that the passive and progressive constructions should display more reduction than the source copula construction. If frequency influences reduction then the copula construction should show the most reduction and the passive construction should show the least.

Modeling Reduction

The spoken section of the COCA, which contains transcripts of American TV and Radio programs with spontaneous speech, was used to build a sample of 3080 entries coded for presence/absence of reduction based on the orthographic transcription. A range of independent variables were coded by hand. Reduction was modeled using mixed-effects logistic regression. Random factors included the speakers and programs in which the target of interest occurred. Fixed factors included construction type, transitional frequency of collocates, log frequency of collocates and various phonetic and syntactic variables.

A series of models were built exploring the effects of these factors. Model comparison was used to select the simplest model that was as predictive as the most complex model. Models were compared with log likelihood tests and bootstrapping was done on logistic regression models with the random factors removed.

This paper presents the final four models: one for the combined results of *is, am* and *are* data and one for each individual verb-auxiliary set. The final models indicated that construction type played a role in influencing reduction, but in a more complicated way than expected. The progressive construction had the most reduction and was significantly different than either of the other two construction types. The copular and passive constructions were not significantly different from each other in any model except for the model of *am* only data.

Other significant factors included the type of subject NP, the preceding and following phonemes, the length of the preceding NP, preceding utterance and preceding word length, the occurrence of a reduced variant of *be* in the preceding utterance, log frequency of collocates and transitional probability of following collocates. The full model has a C index of concordance of .91, above the .8 threshold that Baayen (2008) argues is needed for adequate predictive capability. Models for each verb-auxiliary set also have C's above .8.

These results indicate than neither frequency nor grammaticalization alone provides adequate explanation in this case. The tendencies of grammaticalization favor the progressive construction in comparison to the copular construction and the tendencies of frequency favor the progressive construction in comparison to the passive construction.

Further analysis was done, coding separately from the corpus instances of the words *going* or *gone* following the target variant of *be*. This filtered out most cases of

the future construction, which appeared to be a factor influencing the results and was also the most frequent word following a variant of *be* in the progressive construction. In this subset of the data, the progressive still had significantly more reduction than the passive construction. The progressive also still showed more reduction than the copular in the full and *is* models.

Together these results show that, at the very least, reduction of *is*, *am*, and *are* varies by construction type and this pattern is robust. Frequency plays some role, because the least frequent construction type displays the least reduction. However, frequency is not enough to account for the highly frequent copular construction having less reduction than the progressive construction, even when instances of the future construction are coded separately. Grammaticalization also plays a role here because the grammaticalized progressive displays more reduction than its source copular construction. In the case of the future construction, it displays the most reduction and is more frequent than its source spatial construction, which is a classic example of grammaticalization. Neither grammaticalization nor high frequency guarantees that reduction will take place. Reduction is simply a possible outcome. What this study shows is that both are needed to account for the reduction of *is*, *am*, and *are*.

References

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

van Bergem, D. 1995. *Acoustic and lexical vowel reduction*. Amsterdam: IFOTT.

Bybee, J. L. 2007. *Frequency of use and the organization of language*. Oxford & New York: Oxford University Press.

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. Available online at http://www.americancorpus.org .

Heine, B. 1993. *Auxiliaries*. Oxford: Oxford University Press.

Lehmann, C. 1995. *Thoughts on Grammaticalization*. München, Newcastle: Lincom Europa.

Daniel Berndt[1], Gemma Boleda[2], Berit Gehrke[3], Louise McNally[4]
Codeus[1], Universitat Politècnica de Catalunya[2], Universitat Pompeu Fabra[3,4]
daniel@codeus.net, gboleda@lsi.upc.edu, berit.gehrke@upf.edu,
louise.mcnally@upf.edu

Semantic factors in the choice between ethnic adjectives and PP counterparts: Quantitative evidence

Introduction

Many languages offer competing strategies for expressing the participant roles associated with event nominalizations, such as the use of a denominal adjective ((1a)) vs. a PP ((1b)):

(1) a. **French** agreement to participate in the negotiations
b. agreement **by France** to participate in the negotiations

Since Kayne (1981) theoretical work on nominalizations has focused mainly on whether such PPs and adjectives are true arguments of the nominalization or simply modifiers

which happen to provide participant role information, with no conclusive results (see e.g. Grimshaw 1990, Alexiadou 2001, Van de Velde 2004, McNally & Boleda 2004); little attention has been devoted to the factors determining when one or the other option is used (but see Bartning 1986). The goal of this research is to address this latter question, in the hope that a better understanding of these factors will also offer new insight into the argument vs. adjunct debate.

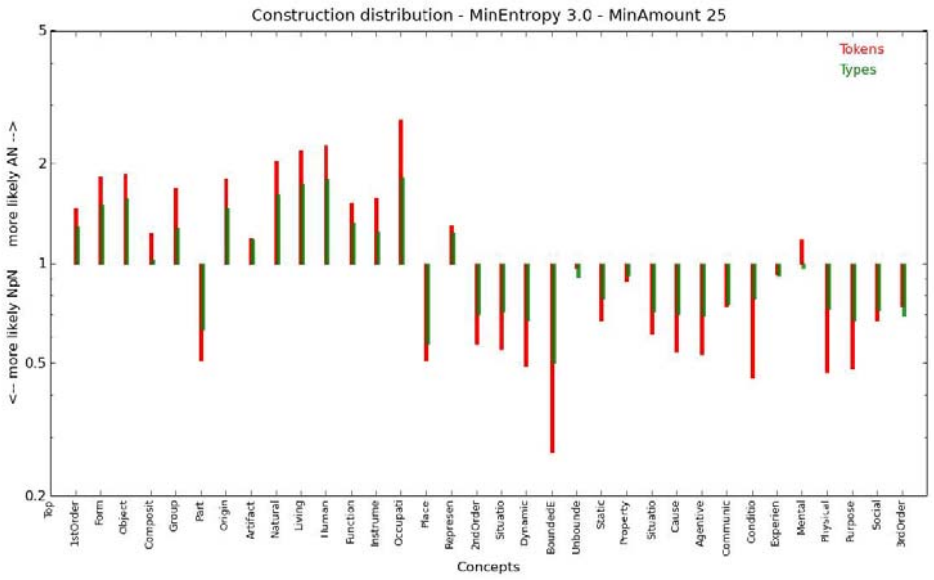
Methodology

We hypothesize that the choice between (1a)-(1b) depends on various factors, including whether the noun is deverbal, the argument structure of the underlying verb, prior or subsequent mention of the participant in question (in (1), France), and what we call concept stability – the degree to which the full noun phrase describes a well-established class of (abstract or concrete) entities. We tested for these factors in a study on the British National Corpus. To reduce unintended sources of variation, we limited our study to so-called ethnic adjectives (Alexiadou and Stavrou to appear) and their nominal counterparts, such as *French-France*. We examined Adjective+N and N+P+Nation_noun examples from 49 different nations whose adjective (*French*) and proper noun (*France*) forms occur 1,000-30,000 times in the BNC, filtering the examples whose head noun was too infrequent (≤ 24 occurrences) or too nation-specific (e.g., *reunification*). To determine the semantic class of the head nouns, we used the WordNet-based Top Concept Ontology (Álvez et al. 2008). For the analysis of nominalizations, we considered only data involving a manually-selected list of 45 nouns.

Results

Unlike nouns denoting physical objects, abstract nouns, including nominalizations, prefer the prepositional construction (see Figure 1; note that nouns in the Top Ontology categories *Part* and *Place*, e.g. *border*, *area*, are an exception).

FIGURE 1. DISTRIBUTION OF HEAD NOUNS ACROSS TOP CONCEPT ONTOLOGY CONCEPTS IN THE ADJECTIVE (AN) AND PREPOSITIONAL (NPN) CONSTRUCTIONS. THE Y-AXIS DEPICTS THE RATIO (AN_CONCEPT/AN)/(NPN_CONCEPT/NPN).



The adjective construction occurs with a much smaller range of nouns than does the PP construction, an effect that is more pronounced with infrequent nations and when only nominalizations are considered (see Figure 2 and Table 1). These results suggest that use of the adjective construction positively correlates with concept stability: Adjective+nominalization combinations are arguably less likely to form stable concepts than adjective+concrete noun combinations (cp. e.g. *a French wine* vs. *a French agreement*).

FIGURE 2. VARIATION OF HEAD NOUNS IN THE ADJECTIVE (AN) AND PREPOSITIONAL (NPN) CONSTRUCTIONS. PERCENTAGES (Y-AXIS) OVER THE LEMMATA THAT OCCUR IN THE AN AND NPN CONSTRUCTIONS POOLED TOGETHER. GRAPH B EXCLUDES THE 20 MOST FREQUENT NATIONS, WHILE GRAPH C DEPICTS ONLY THESE (AS THERE ARE TOO FEW DATA WITH NOMINALIZATIONS FOR INFREQUENT NATIONS).

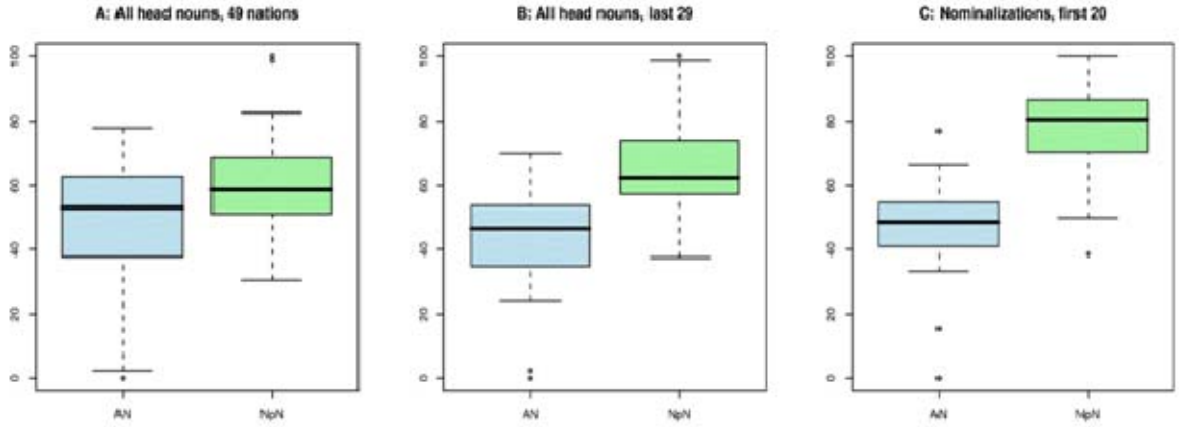


TABLE 1. RESULTS OF PAIRED T-TESTS ON DATA IN FIGURE 2.

data	t	df	p-value
A	-2.49	48	0.016
B	-3.77	28	0.0008
C	-4.36	19	0.0003

Conclusion

Though the other factors are pending analysis, these initial results indicate an asymmetry in the distribution of the constructions in (1); and the strong association between the use of the adjective construction and concept stability specifically lends support to an analysis of ethnic adjectives as classifying modifiers rather than as argument-saturating expressions as posited in related work. To this respect, note that if ethnic adjectives were able to saturate arguments, we would expect them to appear more frequently with, and with a wider variety of, nominalizations.

References

Alexiadou, Artemis 2001. *Functional Structure in Nominals: Nominalization and Ergativity*. Amsterdam: John Benjamins.

Alexiadou, Artemis and Stavrou, Melita to appear. Ethnic adjectives as pseudo-adjectives: A case study in syntax-morphology interaction and the structure of DP. *Studia Linguistica.*

Álvez, Javier, Atserias, Jordi, Carrera, Jordi, Climent, Salvador, Laparra, Egoitz, Oliver, Antoni and Rigau, German 2008. Complete and Consistent Annotation of WordNet using the Top Concept

Ontology. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, LREC 2008.
Bartning, Inge 1986. Le parallélisme entre les syntagmes Nom + *adjectif ethnique* et les syntagmes prépositionnels correspondant en *Nom + de (+ Dét) + Nom géographique*. *Revue Romane* 21, 3-46.
Grimshaw, Jane 1990. *Argument Structure*. Cambridge, MA: MIT Press.
Kayne, Richard 1981. On certain differences between French and English. *Linguistic Inquiry* 12, 349-371.
McNally, Louise and Gemma Boleda 2004. Relational adjectives as properties of kinds. In *Empirical Issues in Syntax and Semantics 5*, Bonami, Olivier and Cabredo Hofherr, Patricia (eds.), 179-196.
Van der Velde, Danièle 2004. A linguistic answer to a philosophical question: Are events generic and repeatable, or irreducibly particular? Ms., U. Lille 3.

Felix Bildhauer, Philippa Cook
Freie Universität Berlin
felix.bildhauer@fu-berlin.de, philippa.cook@fu-berlin.de

Multiple fronting vs. VP fronting in German

Multiple fronting in German

German is classed as a V2 language, that is, normally exactly one constituent occupies the position before the finite verb (the *prefield*) in declarative main clauses. When a sentence contains a non-finite verb (e. g. a participle, an infinitive), the constituent in the prefield may be a verb phrase, consisting of the non-finite verb plus any number of its complements and adjuncts, as in (1).

(1) (DeReKo/Cosmas: NUN93/OKT.01456)
[Ganz auf den Straßenausbau verzichten] **will** die CSU keinesfalls
Completely on the road.development renounce wants the CSU not
'The CSU doesn't want to abandon the development of the road completely.'

In what have been claimed to constitute rare, exceptional cases, however, more than one constituent appears to precede the finite verb, as illustrated in (2):

(2) (DeReKo/Cosmas: R99/JAN.01605)
[Dem Saft] [eine kräftigere Farbe] **geben** Blutorangen.
to.the juice a more.vivid colour give blood.oranges
'What gives the juice a more vivid colour is blood oranges.'

There has been ongoing debate in the theoretical literature concerning the status of examples seemingly violating this V2 constraint, mostly based on constructed examples and considering the data without any context. On the other hand, Müller (2003) discusses a collection of naturally occurring examples, classified by syntactic category and grammatical function of the *prefield* elements. At this point, the question arises as to which constraints govern multiple fronting in German and in particular, which relation multiple fronting bears to the (putatively) more canonical vp-fronting construction exemplified in (1). Interestingly, we find alternation between these two constructions even when the non-finite verb and much of the remaining material in the *prefield* are kept constant (compare (1) to (3)).

(3) (DeReKo/Cosmas: BRZ06/DEZ.08579)
[Ganz] [auf das Skalpell] **können** die Ärzte allerdings nicht verzichten
Completely on the scalpel can the doctors however not renounce
'However, doctors aren't able yet to work completely without a scalpel.'

Since discourse related factors have been argued to play a role in constraining/licensing multiple fronting, the present study explores in detail the role of information structure in tipping the balance towards one or the other of the constructions exemplified in (1) and (3). To our knowledge, no such analysis has ever been undertaken, almost certainly because the data needed for such a comparison are relatively scarce.

Pilot study

Method

Using the w-archive at IDS/Mannheim (2.5 billion word tokens), we conducted a corpus study that compares multiple fronting (MF) and vp-fronting (VPF) with respect to selected features of information structure. Since these corpora are not parsed, there is no straightforward way of extracting the structures we are interested in. As a workaround, a specific instantiation of the multiple fronting / vp-fronting pattern was chosen that can be represented as: ADV PP (X) Vfin…(Y), with a finite verb occuring in either the X or Y position (=VPF vs. MF, respectively). For the pilot study, the pattern was limited to specific lexical material for the adverb, the non-finite verb, and the preposition heading the PP, resulting in two sets: *sich negativ/positiv auf ...auswirken* 'to have a negative/positive effect on sth.' and *ganz auf ...verzichten* 'to do completely without sth.', as shown schematically in Figure 1. All sentence token satisfying these criteria were extracted from the corpus. For each sentence token, the following factors were annotated a.) focus; b.) givenness of the PP; c.) givenness of the subject; d.) length of the PP; e.) length of the subject-NP.

FIGURE 1. PATTERNS EXAMINED IN PILOT STUDY.

	SET A	SET B
MF	[*Ganz*] [*auf* NP] V_{fin} …*verzichten*$_{nfin}$	[*Positiv/Negativ*] [*auf* NP] V_{fin} …*auswirken*$_{nfin}$
VPF	[*Ganz auf* NP *verzichten*$_{nfin}$] V_{fin} …	[*Positiv/Negativ auf* NP *auswirken*$_{nfin}$] V_{fin} …

Results

Remarkably, the "canonical" construction (VPF) turned out to be less frequent than the "non-canonical" one (MF) (Figure (2)).

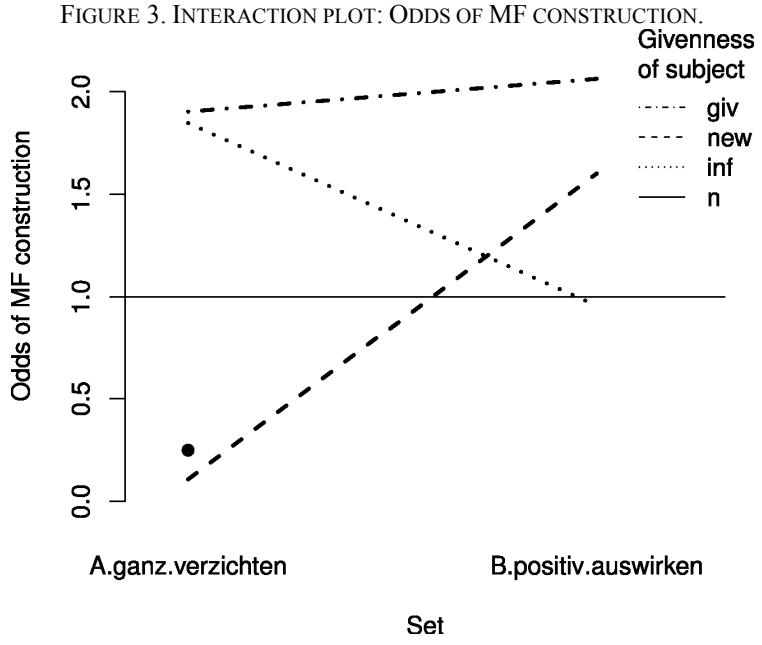
FIGURE 2. EXTRACTED SENTENCE TOKENS BY SET AND CONSTRUCTION.

	SET A	SET B	TOTAL
MF	81	74	155
VPF	50	26	76
TOTAL	131	100	231

A logistic regression was performed on the annotated data (with the construction type as the dependent variable), followed by a stepwise model selection process (optimizing AIC). Length (both of the PP and subject-NP) as well as givenness of the PP could be eliminated as factors from the optimal model. Figure 3 plots the odds of the

MF construction against the givenness of the subject-NP (*given*, *new*, *inferrable*, *none*) for the two sets.

FIGURE 3. INTERACTION PLOT: ODDS OF MF CONSTRUCTION.



Firstly, subject-NPs with givenness-status *none* (negative quantified NPs like *niemand* 'nobody') almost exclusively occur in the VPF construction; however, the effect is only marginally significant (p = .07) due to the low overall number of items in this category (n=5). Second, *new* subject-NPs strongly disfavour the MF construction (p = .04), but only in the A-set; the effect is not significant anymore in the B-set.

FIGURE 4. RESULTS OF LOGISTIC REGRESSION.

	COEFFICIENT	STD. ERROR	Z	P	
(Intercept)	0.64	0.22	3.02	0.003	**
set=B	0.08	0.91	0.09	0.93	
givenness-subj=inf	-0.03	0.55	-0.05	0.96	
givenness-subj=new	-2.87	1.41	-2.04	0.04	*
givenness-subj=none	-2.03	1.14	-1.78	0.07	.
focus=subj	0.84	0.86	0.98	0.33	
focus=other	-16.21	1029.12	-0.02	0.99	
set=B × givenness-subj=inf	-0.74	1.15	-0.64	0.52	
set=B × givenness-subjv=new	2.61	1.49	1.76	0.08	.
set=B × givenness-subj=none	NA	NA	NA	NA	

Discussion

The results of our pilot study suggest that, contrary to what was expected, none of the information structural factors that were considered explain the observed MF-VPF alternation across the two sets of data. Instead, we find item-specific effects (that is, specific to the lexical material that was held constant in each set). These findings, if they carry over to a larger, less constrained data set, are interesting from a theoretical point of view because they would seem to stress the role of item specific knowledge as opposed to higher-level generalizations in this particular domain of grammar.

An extended version of the study will include more data, both in absolute terms and in terms of variation. The general pattern will be the same as the one examined in the pilot study, i. e.:

(4) ADV XP (X) Vfin…(Y)

However, NPs as well as any PP are allowed in the slot labeled 'XP' in (4), and the non-finite verb may vary freely. Moreover, the initial adverb may vary across a set of semantically similar adverbs (thus we have synonym sets centered around *negativ*/*positiv* and *ganz*. In addition, one more such synonym set will be considered, centered around *weiterhin* 'still'). Figure 5 illustrates this for the synsets *ganz* and *weiterhin*. As can be seen, extending the study in this way yields another 840 sentences tokens that will allow us to compare the two constructions on a more solid empirical basis.

FIGURE 5. DATA TO BE INCLUDED IN EXTENDED STUDY.

SYNSET	GANZ		WEITERHIN	
ADVERBS	*ganz, komplett, gänzlich, vollständig, völlig*		*weiterhin, nach wie vor, weiters, fortwährend, immer noch*	
TOTAL	651		189	
CONSTRUCTION	MF	VPF	MF	VPF
TOKENS	276	375	94	95

References

Müller, Stefan. 2003. Mehrfache Vorfeldbesetzung. *Deutsche Sprache* 31(1), 29-62, available at: http://hpsg.fu-berlin.de/~stefan/Pub/mehr-vf-ds.html [28 February 2011].

Peter Bosch, Sascha Alexeyenko, Kirsten Brukamp, Maria Cieschinger, Xiaoye Deng, Peter König
University of Osnabrück, Institute of Cognitive Science

Definite reference is not always based on salience

The currently probably most common view of definite reference is that speakers using definite referential expression (DREs) of the form "the [NP]", mean to refer to exactly one specific entity that satisfies the descriptive content of the NP, and would not use a DRE unless this uniqueness condition were fulfilled.

It has been observed, however, that DREs are frequently *also used in situations* where the uniqueness condition is not fulfilled, and apparently with no harm to communicative success. This has led to the hypothesis (von Heusinger 1995, Lewis 1979, Roberts 2003, and many others) that, perhaps, the uniqueness condition should not be formulated absolutely, in the sense that there is exactly one entity in the discourse context that satisfies the descriptive content of the NP, but rather in the weaker sense that there is exactly one such entity that is clearly more salient in the discourse situation than any potential competitors that would also satisfy the descriptive content of the NP.
We tested a specific interpretation of this hypothesis in a visual world study, i.e., we tracked subjects' eye movements with respect to a visual scene, while they listened to

stories containing DREs that could be interpreted as referential with respect to either one or more objects in the display.

Our results show that unambiguous DREs which referred to objects in the display that had already unambiguously been referred to earlier in the story (anaphorically definite referential expressions) or unambiguous DREs which literally fulfilled the uniqueness condition with respect to the scene in the display, without their referent having been mentioned before (firstmention uniquely referential definite expressions), both led to a significant increase of focussing frequency for the intended referents, with a peak at about 1500 ms after onset of the DRE.

When the DREs were not anaphorically definite and there were several referents in the display satisfying the descriptive content of the NP, the DREs obviously did not satisfy the uniqueness condition and were strictly speaking ambiguous. Ambiguous DREs still led to a reliable increase of focussing frequency, however, when one of their referents was more salient that its competitors, either because it was visually isolated from its competitors, or because it was visually close to a previously mentioned referent. In these saliency conditions the relative focussing frequency for the referents was clearly smaller than in the anaphoric and uniqueness cases, and there was a clear delay in the rise of focussing frequency, peaking only at about 2200 ms after onset of the DRE. Focussing frequencies in the saliency conditions were still much closer though to the anaphoric and uniqueness cases than in the condition of fully ambiguous DREs, where no referent led to a significant rise in relative focussing frequency.

We conclude that in situations where the uniqueness requirement for DREs is neither literally satisfied by the visual scene nor supported by unambiguous anaphoric reference, the interpretation of DREs does not proceed by the same cognitive processes that support the regular interpretation of DREs. Visual salience of referents alone still leads to a reliable interpretation of DREs, but is supported by interpretation processes that take more time and yield a lower focussing frequency.

References

von Heusinger, Klaus 1995. Formal Aspects of a Pragmatic Theory of Definiteness. In P. Bærentzen (ed.) 1995. *Aspekte der Sprachbeschreibung. Akten des 29. Linguistischen Kolloquiums in Aarhus 1994*, 81-84. [Linguistische Arbeiten 342]. Tübingen: Niemeyer.

Lewis, David .1979. Scorekeeping in a Language Game. In Bäuerle, Rainer, Egli, Urs, von Stechow, Arnim (eds.), *Semantics from Different Points of View*, 172-187. Berlin; Heidelberg; New York: Springer.

Roberts, Craige 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy* 26, 287-350.

Oliver Boxell
The University of Essex
osboxe@essex.ac.uk

Discourse-linking and long-distance syntactic dependency formation

The background and aims of the study

Lexically-specified or D(iscourse)-linked *wh*-phrases (e.g. *which horse*) possess properties which are distinct from their non D-linked counterparts (e.g. *who*). In particular, it seems that *wh*-dependencies which form with a D-linked antecedent are

less sensitive to constraints like islands (Chomsky 1962; Ross 1967) and Superiority (Chomsky 1973). For example, (1a) shows a non D-linked *wh*-island violation which is ungrammatical, but notice that the grammaticality of the same violation is significantly ameliorated when the antecedent is D-linked in (1b).

(1) (a) *Who did Jim wonder whether Mary loved_?
(b) Which man did Jim wonder whether Mary loved_?

Pesetsky (1987) accounted for this by positing that D-linked dependencies can form via unselective binding rather than the conventional *wh*-movement, on the basis that constraints such as islands and Superiority are thought of as constraints on movement specifically. The present study aimed to test this formal account of D-linking effects by examining its psychological reality during processing.

Locality constraints motivate the idea that *wh*-movement unfolds successive-cyclically (= using multiple within-clause steps), leaving a representation of the antecedent at each clause-boundary as it does so. Binding, however, is thought to co-index an antecedent with its underlying position in a single operation. It has previously been possible to identify the real-time online instantiation of intermediate antecedent representations during the processing of long-distance *wh*-dependencies (Gibson & Warren 2004; Marinis et al. 2005). The present study appropriates the method used in these studies while adding a +/-D-linked antecedent condition. If Pesetsky (1987) is correct, evidence of the intermediate reactivation of antecedents during long-distance *wh*-dependency formation should be restricted to those with non D-linked antecedents. This is because, according to the theory, only non D-linked dependencies should form in multiple steps if D-linked ones use a (single-step) binding operation.

The experimental method and materials

The main task used a self-paced reading method (Just et al.1982) in which participants read through sentences segment-by-segment while reaction-time software recorded the reading time for each one. Longer reading times were associated with greater processing cost. The materials included two critical manipulations. The first was whether the *wh*-dependency being set-up had a D-linked antecedent or not (+/-D-linking) and the second was whether or not the dependency included an intervening (CP) clause-boundary which could provide the syntactic space for an intermediate representation of the antecedent to be instantiated (+/-Intermediate representation). Since Gibson & Warren 2004 and Marinis et al. 2005 among others used materials on which these are based, and found that the effects being searched for in the critical region are specific to those including *wh*-dependency formation, additional no-*wh*-movement control conditions were foregone. The critical conditions are shown in (2)-(5):

(2) **-D-Linking; + Intermediate representation:** The manager wondered **who** the secretary claimed [**CP *who*** that the new salesman had pleased ***who*** in the meeting].
(3) **-D-Linking; -Intermediate representation:** The manager wondered **who** the secretary's claim about the new salesman had pleased ***who*** in the meeting.
(4) **+D-Linking; + Intermediate representation:** The manager wondered **which gentleman** the secretary claimed [**CP *which gentleman*** that the new salesman had pleased ***which gentleman*** in the meeting].

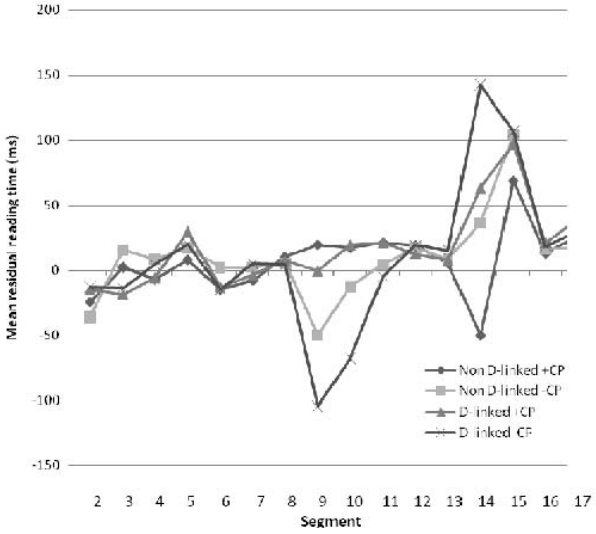
(5) **+D-Linking; -Intermediate representation:** The manager wondered **which gentleman** the secretary's claim about the new salesman had pleased ***which gentleman*** in the meeting.

The rationale was that an intermediate representation of an antecedent would facilitate subsequent antecedent reactivation at the underlying verb position since it would make that argument more locally activated in the parse. Thus, if Pesetsky's (1987) theory is to be consistent with these data, such facilitation should be restricted to dependencies with non D-linked antecedents.

Results and discussion

Figure 1 displays the average residual reading times for each segment. Reading times at the critical subcategoriser (*pleased*), which was Segment 14, were statistically significantly quicker in conditions with an intervening CP relative to conditions without one. Planned comparisons revealed this was true for both non D-linked and D-linked antecedents. Thus, these data suggest that both non D-linked and D-linked antecedents were reactivated at intermediate clause-boundaries within their dependencies. It is argued that this is consistent with multiple-step movement-type dependency formation, but inconsistent with single-step binding-type dependency formation. As such, these data are not consistent with Pesetsky's hypothesis.

FIGURE 1. MEAN RESIDUAL READING TIMES PER SEGMENT (MS).



These data may also provide reason for positing a new processing-based theory for the amelioration effects of D-linking on certain constraints, in which the widely-attested early-discourse structure instantiation it provides (e.g. Radó 1998; Frazier & Clifton 2002; Diaconescu & Goodluck 2004) "stabilises" potential constraint violations as and when they arise in a time-locked position in the parse. This would account for the data seen in Segment 9 of Figure 1, whereby the presence of a D-linked antecedent significantly reduces processing time when encountering what is argued to be a *potential* island boundary.

Finally, these data provide the first evidence of successive-cyclicity occurring online during long-distance dependency formation involving complement clauses, and thus lend support to the formalisms and theories which propose such multiple-step

dependencies in these environments. Additionally, a dissociation is revealed between the kind of antecedent reactivation found at underlying subcategorisers versus that which is found at intermediate positions. While the lexical-specificity of D-linked antecedents significantly increases the amount of time taken for reconstructing them at the verb (as can be seen in Segment 14 of Figure 1), this is not the case for intermediate reactivations, where a D-linked antecedent is reactivated significantly faster than a non D-linked one (as is demonstrated by the +CP conditions of Segment 9 in Figure 1). Thus, it is argued that while antecedent reconstruction at the verb involves full lexical information access, at intermediate positions it is likely to be a purely structural phenomenon.

Concluding remarks

The data collected in the present study provide some quantified psycholinguistic evidence which indicates that intermediate representations of antecedents are present in long-distance *wh*-dependencies, both when they are non D-linked and crucially when they are D-linked. As such, the ameliorative properties of D-linking over certain "movement" constraints like islands do not seem to be explicable, as Pesetsky (1987) suggests, with an unselective binding account. This is because the presence of intermediate representations is indicative of successive-cyclic movement-type dependency formation rather than single-step binding-type dependency formation.

References

Chomsky, N. 1962. The logical basis of linguistic theory. In *Preprints of the papers for the ninth International Congress of Linguists,* Halle, M. (ed.) 509-74.

Chomsky, N. 1973. Conditions on transformations. In *A Festschrift for Morris Halle*, Anderson, S. R. and Kiparsky, P. (eds.) New York: Holt, Rinehart and Winston.

Diaconescu, R. and Goodluck, H. 2004. The pronoun attraction effect for D(iscourse)-linked phrases: Evidence from speakers of a null subject language. *Journal of Psycholinguistic Research* 33, 303-320.

Frazier, L. and Clifton, C. 2002. Processing D-linked Phrases. *Journal of Psycholinguistic Research* 31, 633-659.

Gibson, E. and Warren, T. 2004. Reading time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7, 55-78.

Just, M., Carpenter, P. and Woolley, J. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 111, 228-238.

Marinis, T., Roberts, L., Felser, C. and Clahsen, H. 2005. Gaps in second language sentence processing. *Studies in Second Language Acquisition* 27, 53-78.

Pesetsky, D. 1987. Wh-in-Situ: Movement and Unselective Binding. In *The Representation of (In)definiteness*, Reuland E. J. and ter Meulen A. G. B. (eds.) Cambridge, MA: MIT Press.

Rádo, J. 1998. Discourse-linking and topicality: Parsing *wh*-questions in English and Hungarian. Poster presented at the 11th Annual CUNY Conference on Human Sentence Processing. Rutgers: New Brunswick.

Ross, J. R. 1967. Constraints on Variables in Syntax. PhD dissertation, MIT.

Lisa Brunetti, Stefan Bott
Laboratoire Dynamique du Langage / Université Lumière Lyon2, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
lisa.brunetti@univ-lyon2.fr, sbott@lsi.upc.edu

Subject inversion in Romance: a corpus-based study

Introduction

We present a corpus based study that investigates which factors trigger a postverbal position of the subject in Spanish utterances, as in (1). This study is the first part of a larger study which includes other Romance languages (Catalan and Italian).

Spanish
(1) *y voilà, salió la ranita pequeña a través de la ventana*
and there-he-is came-out the frog small across of the window
'And there came the small frog through the window'

The literature on Romance Subject Inversion (RSI) associates this phenomenon either with semantic, syntactic, or pragmatic properties. According to several authors, RSI depends on the semantic and thematic properties of the verb. In particular, RSI has been associated with intransitive presentational verbs or verbs of existence or appearance (cf. Hatcher 1956, Lambrecht 1994). This fact would explain the preference for inverted subjects with unidentifiable reference. Other scholars have emphasized the relation between inversion and unaccusative verbs. While most part of the literature has focused on the syntactic aspects of such a relation, a few works have focused on the relation between inversion and the fact that these verbs select a non-agent subject (cf. Lambrecht 1995, 2000, Kennedy 1999). From a pragmatic point of view, an inverted subject is considered as either focused or part of the sentence focus (for Spanish, see Contreras 1976, Zubizarreta 1998, 1999). A different proposal comes from Marandin 2003, who argues that the crucial pragmatic factor triggering RSI is that the predicate is given. Finally, from the point of view of clause type, inversion in Spanish obligatorily occurs in interrogative clauses, and frequently – but not obligatorily – in relative clauses (Torrego 1984).

Although RSI has been the object of much interest in the literature, research based on an exhaustive quantitative analysis of naturally occurring data is, to the best of our knowledge, still scarce. On the other hand, the complexity of this phenomenon and the fact that it is often claimed to be strongly related to discourse factors demand an account that is not solely based on constructed sentences and introspective judgments. The present work wants to be a first step towards the filling of this empirical gap. Our aim is to quantitatively determine the burden of different factors in predicting RSI and more precisely, to understand how far RSI can be attributed to purely syntactic/lexico-semantic rather than pragmatic features.

Data

A Spanish oral corpus of 25000 words was used, transcribed from the recordings of free narrations of textless stories for children (Meyer 1969). This corpus is part of a larger multilingual corpus which also includes Catalan, Italian, German, and English narrations with the same characteristics. A total of 1221 pre-and postverbal subjects

were found. These occurrences were annotated for 34 different features. There were 13 features we could find less than 10 times in the corpus, so we excluded them and we used the remaining 21 features for the study.

As for the lexical-semantic properties of the subject, we annotated indefinite, generic, and quantified subjects. We also annotated whether subjects lacked agentive properties. From a syntactic point of view, we signaled whether the subject was sentential or modified by a restrictive relative clause. Verbs were divided by 11 classes according to their argument properties: transitive, intransitive, unaccusative, (object experiencer) psychological verbs, and different types of verbs with reflexive morphology, divided into: pure reflexives, lexicalized reflexives, reciprocals, de-causatives, auto-causatives, antipassives, psychological reflexives (Creissels 2006). As for the class of unaccusatives, its existence is not uncontroversial in the literature. For this reason, we also classified verbs according to those semantic properties that are generally, but not always, associated with the unaccusative class, namely: verbs of movement toward a point, of existence, absence, commencement, continuation, appearance, occurrence, and stance (Hatcher 1956). A third set of features concerns clause types: relative, temporal, concessive, final, conditional, causal, and declarative subordinate clauses, wh-and y/n-direct and indirect questions, exclamatives. We finally added two pragmatic features: *givP*, for a predicates that are discourse given, and *sbj_new*, for subjects that are discourse new.

Analysis

A standard chi-square test was applied for the correlation of each feature with the occurrence of RSI. The results show that the correlation is significant for 18 features (see Table 1). The three top-scoring features are related to argument structure, hence to lexico-semantic factors. The crucial factor seems to be the subject lacking volition/control on the event (see also the high score of appearance, occurrence, and decausative-reflexive verbs, which all select a non-volitional subject). The fourth top-scoring feature is syntactic: SI in Spanish is highly predictable with a relative clause. A comparison with Catalan and Italian data shows that the same holds for Catalan but not for Italian, where this correlation is not significant.

The two pragmatic features – predicate givenness and subject newness in discourse – show both a significant correlation with inversion, but the latter appears to be more significant than the former. If we look at these results in the Italian corpus, we see an even greater distance in terms of predictability between the two features: although both are significantly correlated with inversion, predicate-givenness has a higher p-value than subject-newness. From these results we cannot conclude with Marandin 2003 that the pragmatic trigger of SI is predicate-givenness. The traditional assumption that inversion is related to the information status of the subject seems to be confirmed by our data.

The percentage of inversion per speaker was also calculated. The variation widely ranges from 3% to over 37%, by which we conclude that stylistic choices are crucial for SI selection. A significant correlation is also observed with the story *Frog goes to dinner*, whose plot favors a narration with frequent topic shifts. This fact supports the idea that the organization of discourse strongly influences the subject position. However, we must say that no such correlation was found in Catalan and Italian data.

In addition to pure descriptive statistics we carried out a small experiment with C4.5 decision tree classifiers (in the J48 implementation of Weka, Witten & Frank 2005). We used a tenfold cross-classification to remedy the sparseness of data. Although the amount of data was not sufficient to train a strongly reliable classifier, its overall accuracy and the precision of predicting +RSI are fairly acceptable (83,8% and 73,6%). However, the recall of +RSI prediction is poor (34,9%).

The decision tree algorithm allowed us for an analysis of the misclassified examples. Error analysis showed that 36,5% of the false negative cases (wrongly classified as -RSI) and 64% of the false positive cases would also be acceptable with a preverbal subject, which explains the low recall for +RSI: in many cases SI is simply not obligatory. Interestingly, inverted subjects are more predictable then preverbal ones when the cues for one particular construction are fewer. In other words, inversion appears to be the default case, while preverbal subjects are required under more specific circumstances. An observation of the contexts of false positives further reveals that many misclassifications co-occur with discourse phenomena, like topic shift or contrast. This finding confirms us how discourse plays a crucial role in inversion, and that future research will have to focus on the addition of more, and more sophisticated, pragmatic features.

TABLE 1.

	Spanish	**chi-square**	**p-value**
1	non-agent sbj (no volition)	129,785933	<0,001
2	Unaccusative V	96,391268	<0,001
3	V of Appearance	72,9405773	<0,001
4	relative clause	71,0134903	<0,001
5	V of directed movement	47,7637428	<0,001
6	indefinite sbj	36,5741623	<0,001
7	V of Occurrence	36,1585354	<0,001
8	discourse new sbj	34,1611428	<0,001
9	Transitive V	32,518726	<0,001
10	sbj todo	30,9036948	<0,001
11	Given Predicate	25,5316137	<0,001
12	Intransitive V	23,0607253	<0,001
13	Decausative reflexive V	22,4422042	<0,001
14	Copula V	9,3327433	<0,005
15	V of Stance	5,94830663	<0,05
16	« Frog goes to dinner »	5,0754379	<0,05
17	quantified subject	4,82319875	<0,05
18	Psychological reflexive V	4,80499178	<0,05

References

Contreras, Heles 1976. *A Theory of Word Order with Special Reference to Spanish*. Amsterdam: North Holland.

Creissels, Denis 2006. *Syntaxe générale, une introduction typologique*. Paris: Hermès.

Hatcher, Anne G. 1956. Theme and underlying question. Two studies of Spanish word order. *Word* 12, 14-31.

Kennedy, Becky 1999. Focus constituency. *Journal of Pragmatics* 31, 1203-1230.

Witten, Ian H. and Frank, Eibe 2005. *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition). San Francisco: Morgan Kaufmann

Lambrecht, Knud 1994. *Information Structure and Sentence Form. A Theory of Topic, Focus, and the Mental Representations of Discourse Referents*, [Cambridge Studies in Linguistics 71], Cambridge: Cambridge University Press.
Lambrecht, Knud 1995. The pragmatics of case: On the relationship between semantic, grammatical, and pragmatic roles in English and French. In Shibatani, Masayoshi and Sandra A. Thompson (eds.), *Essays in Semantics and Pragmatics. In Honor of Charles J. Fillmore*, 145-190. Amsterdam: Benjamins.
Lambrecht, Knud 2000. When subjects behave like objects: An analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language* 24, 611-682.
Marandin Jean-Marie 2003. Inversion du sujet et structure de l'information dans les langues romanes. In Godard, Danièle (ed.), *Langues romanes. Problèmes de la phrase simple*. Paris: Editions du CNRS.
Torrego, Esther 1984. On inversion in Spanish and some of its effects. *Linguistic Inquiry* 15, 102-129.
Zubizarreta, María Luísa 1998. *Prosody, Focus and Word Order*. Cambridge, Mass.: MIT Press.
Zubizarreta, María Luísa 1999. Las funciones informativas: tema y foco. In Bosque, Ignacio and Demonte, Violeta (eds.), *Gramática descriptiva de la lengua española vol. 3*, 4215-4244. Madrid: Espasa Calpe.

Felix Golcher, Marc Reznicek
Humboldt Universität zu Berlin
felix.golcher@hu-berlin.de, marc.reznicek@staff.hu-berlin.de

Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus

Introduction

The process of transfer of structures from the mother tongue (L1) has been shown to take place on different linguistic levels like lexis, syntax, discourse etc. (Ellis 2009:295). Up to now though, it has been difficult to quantify the relative strength of transfer on those different levels at the same time. This will be the main objective of this study. We therefore separately look at the surface level text, sequences of the part-of-speech (POS) tags as a rough representation of the syntactic structure and the sequence of lemmas reflecting mainly the choice of lexical items.

Furthermore we will compare the results not only for the original learner texts, but also for an annotation layer that represents the reconstructed utterances of the learners, thereby enabling us to measure effects that stem from ungrammaticality in the text compared to L1-induced phenomena that do not lead to ungrammatical structures.

We compare argumentative texts on four different topics written by advanced learners of German in the Falko Learner Corpus of German. Our results indicate that besides a known influence of text type and genre (Byrnes et al. 2004) one has to take into account the confounding variable of text topic.

Research on n-grams

Research done on POS tag sequences shows that while certain linguistic structures of learner language display systematic over- and underuse for all mother tongue (L1) groups (Zeldes et al. 2008), many seem to be L1 specific (Aarts et al. 1998, Borin et al. 2004). In those studies the length of the n-grams have been arbitrarily constrained to small numbers (mostly $n<5$). Our approach takes into account all n-grams found in a text to calculate a similarity between two texts by comparing the two sets of n-grams.

The more n-grams appear in both texts, the higher is the rate of similarity between those two texts. This might allow us to find satisfactory results even in a relatively small corpus like FALKO. The stylometric measure we use is described in Golcher 2007.

Stylometry

If we train our algorithm on sets of German texts stemming from different mother tongue speakers, we can use this "knowledge" to classify unseen texts into those L1 sets. The more the transparent L1 features show up in the target language text, the better the success of the classification for that language will be. High accuracy would therefore support the claim of very coherent L1 traces which can be interpreted as a result of transfer. Low accuracy would either refute the quality of the method or could be interpreted as a result of L1 independent factors making learner texts too similar to classify them correctly.

If the L1 has an influence on the similarity of two texts, we would expect it to show up at different strengths depending on the amount and type of information included in the text. While the raw text comprises information about lexical choice, morphology, syntax etc., in the chain of automatically annotated part-of-speech tags the lexical information has been lost and only a schematic version of the morpho-syntactic information has been left. In the chain of lemma tags on the other hand morphology has been suppressed while surface linearity and lexical choice have been conserved. Comparing these three layers, one can at least partially separate two kinds of transfer effects: purely lexical effects should not show up on the POS level, while syntactic transfer which cannot be attributed to simple linearity phenomena should be less visible. Stylometric methods have been applied to the task of identifying a writer's mother tongue before (Koppel et al. 2005a, b; Tsur & Rappoport 2007, Golcher 2007), this however has been done on the surface level only. As an interesting side effect our study partially replicates the results of the papers cited above which are all based on the ICLE corpus (Granger 2002) with data from German English learners.

Since the corpus also includes a target hypothesis (TH) for every learner utterance, which forms a reconstructed target language version of the original utterance (Lüdeling 2008), as well as POS and lemma information for this TH, we can add a comparison of the results for the TH with the original learner texts. Since the TH inherently reflects a mediation between the learner text and a normally produced native speaker text (Reznicek et al. 2010), the L1 effect should be less visible for all three levels of representation. This would then help to evaluate the quality of the methods predictive power of transfer effects.

The algorithm

The measure used for quantifying the similarity of two texts is based on the frequency distribution of all substrings in both texts. Basically, the products of all those frequency pairs are summed up logarithmically. The immense number of substrings of the texts is handled by storing them as suffix trees. Written as a formula, the similarity measure $S(T_1,T_2)$ is defined as

$$S(T_1, T_2) = \sum_{all\ S} \log\left[F_{T1}(s) \cdot F_{T2}(s) + 1\right] \qquad (1)$$

Here, the sum runs over all substrings *s*. $F_{T1}$and F_{T2} denote the frequencies of *s* within the two texts. The logarithm ensures that the much higher frequencies especially of very short substrings do not mask the potentially interesting low frequency substrings. Since the logarithm of 1 is 0, all substrings which do not occur in both texts make no contribution to *S*, which is as it should be. The addition of one is also needed in order to include substrings occurring once in both texts in *S*.

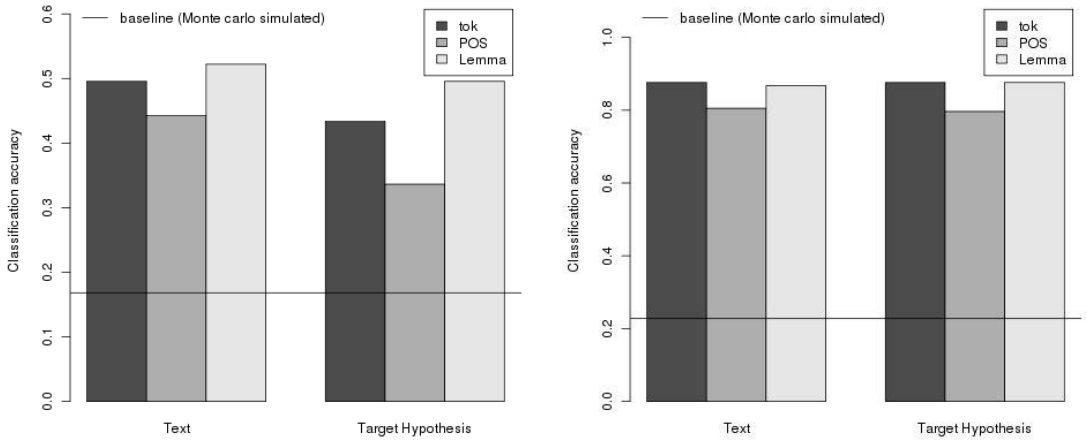
Since the substring frequencies will grow in general as the text gets longer, a normalization procedure is needed. In practice, a simple averaging strategy has turned out to be efficient (see Golcher 2007, where the described method has been shown to be highly competitive within a range of stylometric tasks).

Empirical analysis

To minimize third language interactions the data was restricted to include texts of non bilingual speakers only. We used essays from the five largest L1 groups only. This left us with 42 English, 37 Danish, 14 French, 10 Russian, and 10 Turkish native speakers. All files together consisted of just below 60,000 tokens.

For all possible text pairs, *S* was computed. This yields a symmetric 126 by 126 matrix. As hinted above, this matrix is normalized by dividing each cell by the mean of the respective row and column (Golcher 2007). This removes all text specific contributions to *S*, most notably the impact of text length. The variance due to the topic of the essays is still rather large. Thus we split all *S* values into two sets, where both contributing texts do or do not belong to the same topic. The *S* values of both sets were divided by their means. Then, each file in turn was classified into the L1 group with which it shared the highest mean *S* values, doubly normalized as described. In a second run, the texts were classified according to their topic in the same way. Here, we used the simple normalized *S* values without regard to the L1 influence. This is justified by the relatively small L1 effect compared to topic. Figure 1 shows the results.

FIGURE 1. CLASSIFICATION ACCURACIES FOR L1 (LEFT PANEL) AND TOPIC (RIGHT PANEL). FOR CLASSIFYING L1, THE TOPIC INFLUENCE WAS MITIGATED BY THE HEURISTICS DESCRIBED IN THE TEXT.



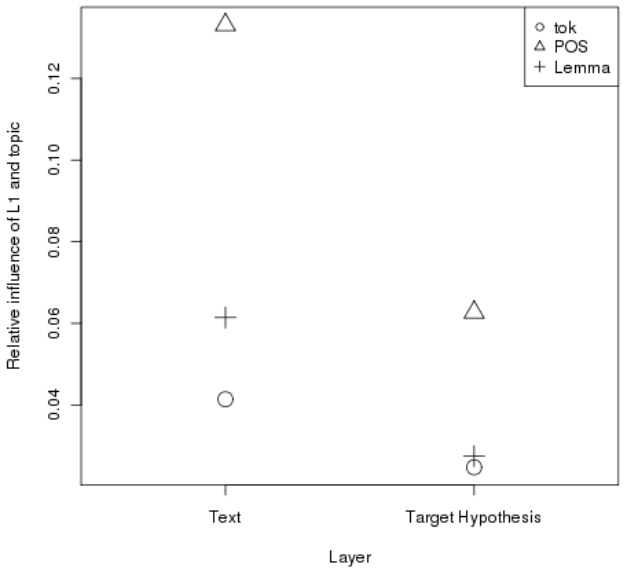
From the left panel of Figure 1 we can draw the following conclusions about the predictability of L1:

1. The overall results of correct predictions for the L1 is significantly higher than random, with a *p* value close to 0.
2. There is a difference in the performance for the different levels of representation: Lemma is best, followed by tokens and POS.
3. The same pattern is repeated for the target hypothesis, but at a lower level.

The accuracy for topic classification on the other hand is very high. Predictably, it is lowest in the POS representation, even if it is still astonishingly high. This striking persistence of the topic predictability can be seen as a hint to a spurious effect: The topical influence could be transferred from the lexical level, for example by identical material in different essays, copied from the essay title. This spuriousness cannot be expected to apply to the more interesting L1 influence in any way.

From these results we see how the classification accuracy of L1 and topic varies with regard to the two variables of text/TH and linguistic representation. The baselines for L1 and topic classifications differ though, and slightly different forms of *S* were used in both cases. This makes it hard to directly compare the impact of L1 and topic on *S*, that is on text similarity. Thus we fitted a linear mixed model (Pinheiro & Bates 2000) to the data. Assuming a linear dependency of *S* on the other variables is an approximation of course, but such a model allows us to quantify L1 and topic effects at the same time. The model has to be a mixed one in order to account for the crossed random effects of the individual essays. The results are shown in Figure 2.

FIGURE 2. COMPARING EFFECT SIZES AS GIVEN BY THE FIT TO A LINEAR MIXED MODEL. DISPLAYED IS THE ESTIMATE FOR THE L1 EFFECT DIVIDED BY THE ESTIMATE FOR THE TOPIC EFFECT.



We can summarize them as follows:

1. The effect of topic is always much more pronounced than the effect of L1. Nevertheless, sameness of L1 was always a significant factor within the model.

2. In the POS representation, L1 is relatively much more influential. Thus, even if a topic effect is still rather strong in this representation, the L1 effect resides much more in this structural representation of the text.
3. The same pattern is repeated in the target hypothesis, but the L1 effect is much weaker in comparison to the topic effect. This is in line with expectations: The target hypothesis aims at correcting errors and errors can be assumed to be L1 specific. And, since the lexical content is supposed to change rather little between the two corpus layers, the topic effect should be very similar.

Conclusion

First, the very high accuracy of the L1 classification strongly supports the claim that L1 dependent transfer effects are present in the learner texts.

The separation of the representation levels shows that transfer dominates on the lexical level. Still there are clear indications of a comparable L1 transfer on the syntactic level. Since the TH corrects mainly morphosyntactic errors, this shows that there are at least in part L1 specific errors that contribute to the coherence of the L1 groups. On the other hand the numbers make it clear that transfer is a lot stronger on levels that do not lead to ungrammatical structures in the strict sense.

Interestingly, the L1 effect on the lemma level is even higher than on the token level, which means that by subtracting morphological clues the L1 effect grows. This can be explained considering that morphology can be related to cues like tense and modality. In fact, topics seem to differ in their likelihood to trigger certain tenses and modalities therefore diminishing L1 influence.

The numbers in the last figure allow the conclusion that the influence of text topic on forms of expression is much larger than has been considered so far, even with regard to syntactic structures. Future studies should therefore control the data not only for text type, register and genre but for topic as well.

References

Aarts, Jan and Granger, Sylviane 1998. Tag sequences in learner corpora. a key to interlanguage grammar and discourse. In *Learner English on computer*. [Studies in language and linguistics],Granger, Sylviane (ed.), 132-141. London: Longman.

Borin, Lars and Prütz, Klas 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In *Corpora and language learners*.[Studies in corpus linguistics 17], Aston, Guy, Bernardini, Silvia and Stewart, Dominic (eds.), 67-87. Amsterdam, Philadelphia: John Benjamins.

Byrnes, Heidi and Sprang, Katherine A. 2004. Fostering advanced L2 literacy. A genre-based, cognitive approach. In *Advanced foreign language learning. A challenge to college programs.*[Issues in language program direction 2003], Byrnes, Heidi, Hiram, Maxim H. (eds.), Boston MA: Thomson/Heinle.

Ellis, Rod 2009. *The study of second language acquisition*. 2. ed. Oxford: Oxford Univ. Press.

Falko: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en.

Golcher, Felix 2007. A new text statistical measure and its application to stylometry. *Proceedings of Corpus Linguistics.* University of Birmingham. 2007. URL: http://amor.cms.hu-berlin.de/~golcherf/cl07.pdf.

Granger, Sylviane, Dagneaux, Estelle and Meunier, Fanny 2002. *The International Corpus of Learner English. Handbook and CD-ROM.* Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane 2008. Learner corpora. In *Corpus linguistics. An international Handbook* [Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29 1], Lüdeling, Anke and Kytö, Merja (eds.), 259-275. Berlin, New York: Mouton de Gruyter.

Koppel, Moshe, Schler, Jonathan, Zigdon, Kfir 2005a. Automatically Determining an Anonymous Author's Native Language. *Intelligence and Security Informatics.* Lecture Notes in Computer Science. Springer. 209-217.

Koppel, Moshe, Schler, Jonathan, Zigdon, Kfir 2005b. Determining an Author's Native Language by Mining a Text for Errors. In *Proceedings of KDD '05*. Chicago IL.
Lüdeling, Anke, Doolittle, Seanna, Hirschmann, Hagen, Schmidt, Karin and Walter, Maik. Das Lernerkorpus Falko 2008. *Deutsch als Fremdsprache* 45(2), 67-73.
Pinheiro, Jose, Bates, Douglas 2000. *Mixed Effects Models in S and S-Plus.* Springer, Berlin.
Reznicek, Marc, Walter, Maik, Schmidt, Karin, Lüdeling, Anke, Hirschmann, Hagen and Krummes, Cedric 2010. *Das Falko-Handbuch. Korpusaufbau und Annotationen.* Version 1.0. Berlin: Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin. URL: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko [15 October 2010].
Tsur, Oren, Rappoport, Ari 2007. Using Classifier Features for Studying the Effect of Native Language Choice of Written Second Language Words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition.* ACL. 9-16.
Zeldes, Amir, Lüdeling, Anke, Hirschmann, Hagen 2008. What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data. In *Proceedings of QITL3*, Arppe, Antti, Sinnemäki, Kaius and Nikanne, Urpo (eds.), 74-77. Helsinki. URL: http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes_et_al.pdf.

Jack Grieve
University of Leuven
jack.grieve@arts.kuleuven.be

The use of spatial autocorrelation statistics for the analysis of regional linguistic variation

Introduction

In geolinguistics, there is no standard method for testing if an individual linguistic variable, measured across a set of locations, exhibits a statistically significant regional pattern. In traditional dialectology regional patterns in the values of individual linguistic variables are identified through subjective analyses (Kurath 1949; Labov et al. 2006), whereas in modern dialectometry regional patterns in the values of individual linguistic variables are generally ignored (Séguy 1973; Goebl 2006; Nerbonne 2006), as the focus of this field is aggregated regional linguistic variation. This paper introduces two measures of spatial autocorrelation and demonstrates how these statistics can be used to identify significant patterns of regional linguistic variation in the values of individual linguistic variables.

Spatial autocorrelation

Spatial autocorrelation (Odland 1988; Lee & Kretzschmar 1993) is a measure of spatial dependency that quantifies the degree of spatial clustering or dispersion in the values of a variable measured across a set of locations. There are two basic types of spatial autocorrelation statistics: global measures identify whether the values of a variable exhibit a significant overall pattern of regional clustering, whereas local measures identify the location of significant high and low value clusters.

In order to determine if the values of a linguistic variable exhibit significant spatial clustering across a set of locations, global Moran's *I* (Moran 1948; Odland 1988) can be used to test for significant levels of positive global spatial autocorrelation. The value of Moran's *I* ranges from -1 to +1, where a significant negative value indicates that nearby locations tend to have different values (i.e. spatial dispersion), an

insignificant value indicates that nearby locations tend to have random values, and a significant positive value indicates that nearby locations tend to have similar values (i.e. spatial clustering). By testing for significant levels of positive global spatial autocorrelation, it is therefore possible to statistically identify the presence of regional clustering in the values of individual linguistic variables.

In order to determine the location of high and low value clusters in the values of a linguistic variable, local Getis-Ord *Gi** (Ord & Getis 1995) can be used to test each variable for local spatial autocorrelation. Unlike measures of global spatial autocorrelation, which return *one value for each variable* indicating the degree of regional clustering across the *entire* distribution of that variable, measures of local spatial autocorrelation return *one value for each location* for each variable indicating the degree to which that particular location is part of a high or low value cluster. The results of a local spatial autocorrelation analysis can then be mapped across the set of locations in order to identify the position of high and low value clusters.

Application

In order to demonstrate the application of these two measures of spatial autocorrelation, a regional analysis of continuous grammatical, lexical and phonetic variation will be presented based on two American datasets: the phonetic data gathered for the Atlas of North American English (Labov et al. 2006), and lexical-grammatical data extracted from a 25 million word corpus of letters to the editor (Grieve 2009). It will be argued that spatial autocorrelation statistics are particularly useful for identifying regional patterns in the type of geolinguistic data that tends to be obtained when continuously measured linguistic variables are analyzed, which often does not produce the types of clear cut regional patterns that are common in traditional categorical analyses.

The utility of the spatial autocorrelation statistics is demonstrated in the two figures reproduced below.

FIGURE 1. *DO NOT* CONTRACTION RATE.

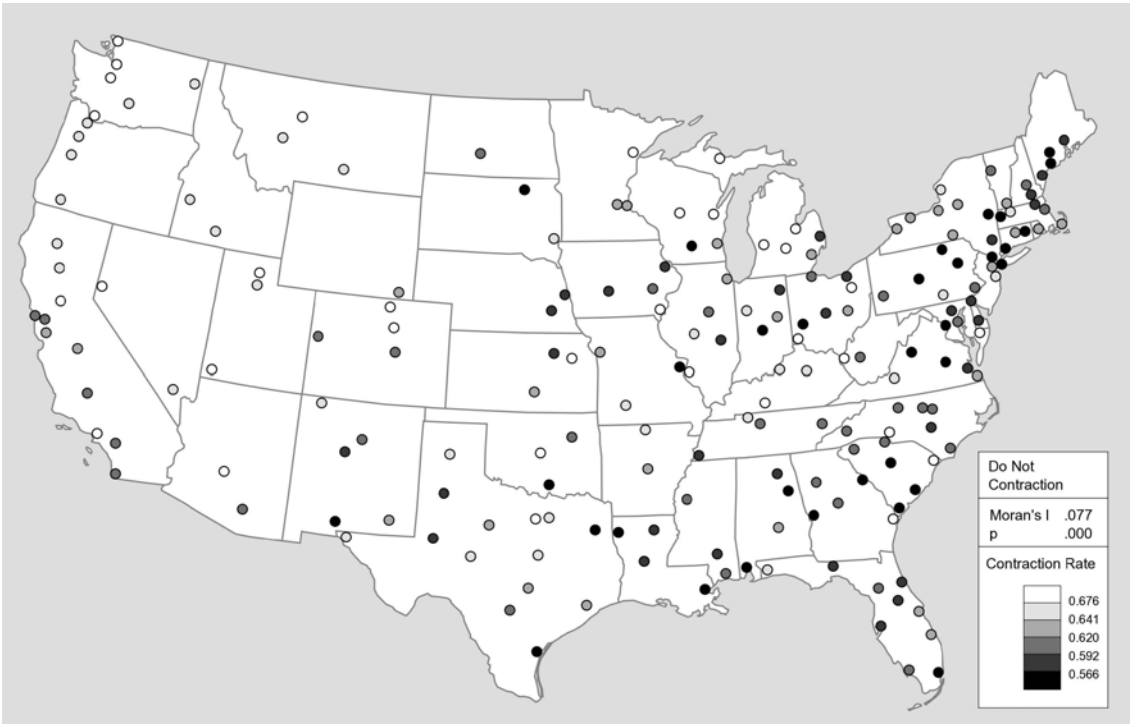
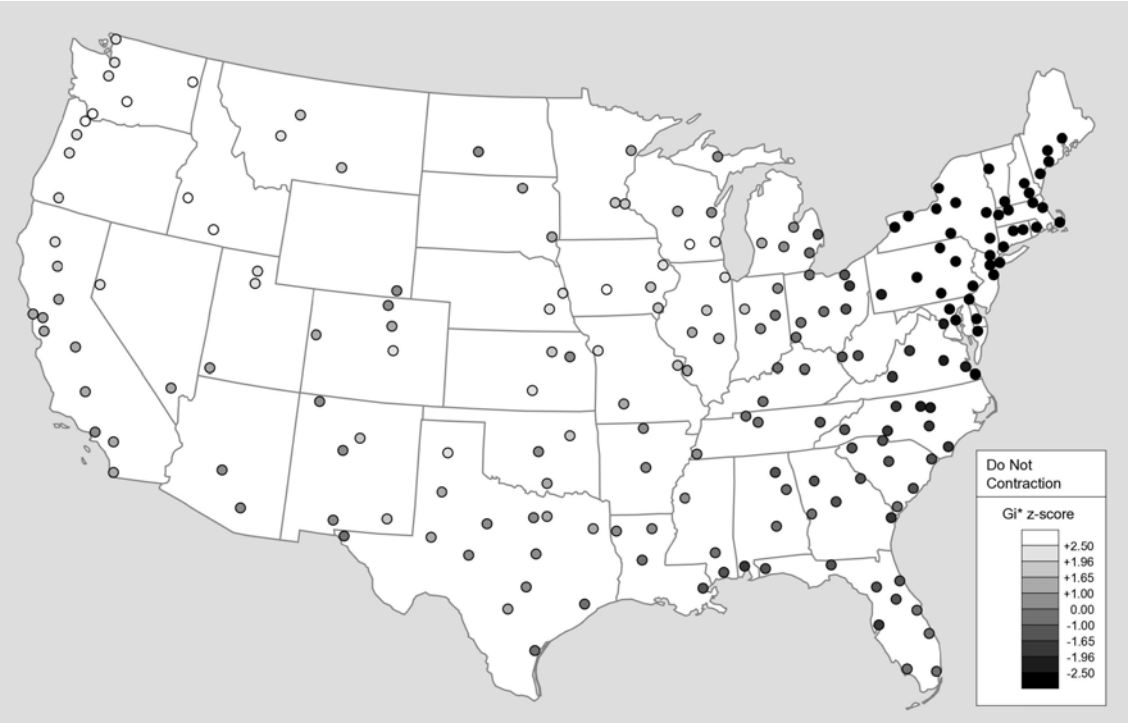


Figure 1 plots *do not* contraction rate (e.g. *doesn't* vs. *does not*) across the 206 cities represented in the letter to the editor corpus. Despite the fact that there is no clear regional pattern in the raw values of this variable, at least by traditional standards, the results of the Moran's I analysis ($I = .077$, $p < .0001$) show that the variable actually exhibits highly significant spatial clustering. The location of these clusters can then be determined by conducting an analysis of local spatial autocorrelation. The results of the Getis-Ord Gi* analysis are presented in Figure 2, which plots the Getis-Ord Gi* z-scores for *do not* contraction rate across the 206 cities, showing quite clearly that *do not* contraction is relatively more common in the Northwest and Western Midwest, and relatively less common in the Northeast.

FIGURE 2. *DO NOT* CONTRACTION GETIS-ORD GI*.



References

Goebl, H. 2006. Recent Advances is Salzburg Dialectometry. *Literary and Linguistic Computing* 21, 411-435.

Grieve J. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English.* Ph.D. Dissertation. Northern Arizona University

Kurath, H. 1949. *Word Geography of the Eastern United States.* University of Michigan Press.

Labov, W., Ash, S. and Boberg, C. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change.* New York: Mouton de Gruyter.

Lee, J. and Kretzschmar, W. 1993. Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Geographical Information Systems* 7, 541-60.

Moran, P.A.P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37, 243- 251.

Nerbonne, J. 2006. Indentifying Linguistic Structures in Aggregate Comparison. *Literary and Linguistic Computing* 21, 463-475.

Odland, J.D. 1988. *Spatial Autocorrelation*. Sage Publications.

Ord, J.K. and Getis, A. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27, 286-306.

Séguy, J. 1973. La dialectometrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane,* 37, 1-24.

Xiao He, Elsi Kaiser
University of Southern California
hex@usc.edu, emkaiser@usc.edu

Reference resolution in Chinese: New insights from psycholinguistics

Introduction

We report two self-paced reading experiments investigating Binding-Theoretic (BT) effects during real-time processing. Previous work on English produced mixed results. Badecker & Straub (2002) investigated sentences like (1) and argued that inaccessible/non-BT-licensed referents (John) are temporarily activated during reference resolution and compete with accessible/BT-licensed referents. However, Nicol & Swinney (1989) argue that inaccessible/non-BT-licensed referents do not compete (see also Sturt 2003).

(1) *John thought that Bill owed himself another chance to solve the problem.*

To enrich our understanding of factors influencing referents' ability to compete, we examined sentences where referents' accessibility depends on intervening referents' person features, namely, the Blocking Effect in Chinese: (i) If the local subject is 3rd person, the reflexive *ziji* 'self' can refer to a long-distance (LD) or a local antecedent (2b) & (2d). (ii) If the local subject is a 1st/2nd person pronoun, LD binding is blocked (2c) (Huang & Liu 2001, Pan 2000).

(2) *{Wo/Zhansan} gaosu bieren {wo/Lisi} juede ZIJI nenggou jin hao daxue.*
{I/Zhangsan} told others {I/Lisi} feel SELF could get-in a good university.

(2a) 1p-1p
Wo1 *gaosu bieren* ***wo1*** *juede ZIJI**1*** *nenggou jin hao daxue.*
I*1* tell others **I*1*** feel SELF***1*** can get-in good university
'**I*1*** told others **I*1*** feel SELF***1*** could get-into a good university.'

(2b) 1p-3p → LD antecedent available
Wo1 *gaosu bieren* ***Lisi****2 juede ZIJI1/2 …*
I1 tell others **Lisi**2 feel SELF1/2 …
'**I**1 told others **Lisi**2 feel SELF1/2 …

(2c) 3p-1p → LD antecedent unavailable, 'blocked'
Zhangsan1 *gaosu bieren* ***wo****2 juede ZIJI*1/2 …*
Zhangsan1 tell others **I**2 feel SELF*1/2 …
'**Zhangsan**1 told others **I**2 feel SELF*1/2 …

(2d) 3p-3p → LD antecedent available
Zhangsan1 *gaosu bieren* ***Lisi****2 juede ZIJI1/2 …*

Zhangsan1 tell others **Lisi**2 feel SELF1/2 …
'**Zhangsan**1 told others **Lisi**2 feel SELF1/2 …

Although these facts are standardly reported in the literature, native speaker judgments suggest that Blocking effects may be less absolute than is often assumed. Thus, in addition to (i) probing the effects that person-feature Blocking has on the availability of potential antecedents during real-time processing, we also (ii) test whether naïve Chinese speakers exhibit Blocking effects with 1p/2p interveners. We use offline comprehension question data and on-line reading times to test whether intervening 1p/2p pronouns block LD referents from competing with local antecedents.

To gain insights into the reasons underlying Blocking, (iii) we also tested whether 1p and 2p differ. Existing research disagrees regarding the underlying reasons for Blocking (e.g., perspective-taking, animacy, feature checking, Huang & Liu 2001, Pollard & Xue 1998, Cole & Wang 1996). We explore a novel prediction related to perspective-taking: non-BT work by Brunyé et al. (2009) found that in multi-sentence discourses, 2p pronouns were actually more effective at inducing perspective-taking than 1p. Specifically, comprehenders are more likely to take the perspective of 2p than 1p. If Chinese Blocking effects are attributable to perspective taking, we may find stronger Blocking with 2p than 1p.

Experiments

In *Experiment 1,* we manipulated matrix and embedded subjects (1st person pronoun/3rd person name) in a self-paced reading experiment with 4 conditions: (a) 1p-1p: both matrix and embedded subjects are 1p (2a); (b) 1p-3p: Matrix subject is 1p; embedded subject is 3p (2b); (c) 3p-1p: matrix is 3p; embedded is 1p (2c); (d) 3p-3p: both are 3p (2d). Twenty participants read sentences word-by-word; reading times (RT) were recorded. Forced-choice questions testing interpretation of *ziji* followed targets (n=32). *Experiment 2* had the same design but now 2p pronouns were used instead of 1p. Twenty-eight new participants took part in Experiment 2.

(3) Sample forced-choice question for (2d):

Shui neng jin hao daxue?
Who can get-in good university
'Who can get in a good university?'
(A). Zhangsan or (B). Lisi

Predictions

Antecedent choices: If Blocking determines final interpretation, the LD antecedent should be available in 1p-3p/2p-3p and 3p-3p, but crucially not in 3p-1p/3p-2p, due to intervening 1p/2p.

RTs: According to prior work, RT slowdowns indicate competition. 1p-3p, 2p3p and 3p-3p should exhibit slowdown from matrix-vs.-embedded-subject competition because both LD and local referents are accessible. 1p-1p/2p-2p should show no competition/slowdown because only one referent is present. If *inaccessible referents are filtered out* (Nicol & Swinney 1989), 3p-1p/3p-2p should show no competition/slowdown. If *inaccessible referents compete* (Badecker & Straub 2002) and the matrix subject is prominent enough, 3p-1p/3p-2p should exhibit

competition/slowdown. RTs were trimmed and log-transformed, and dependent variables were centered to reduce multicollinearity. Data were primarily analyzed with mixed-effects regression (Baayen et al. 2008, Jaeger 2008).

Experiment 1 Results

Antecedent choices: Antecedent choices reveal a *bias for local subjects* (3p-3p: 85.7% local subject, 1p-3p: 95.9%, 3p-1p: 73.1%, *p*'s<0.001). Strikingly, this preference is significantly weaker in 3p-1p (26.9% matrix subject choices, p's<.025), the condition where Blocking predicts the matrix subject to be unavailable.

RTs: At *ziji* and onwards, RTs in 1p-3p, 3p-3p and 3p-1p were significantly longer than in 1p-1p, suggesting that even in 3p-1p, the matrix subject was not blocked from competing. (3p were overall slower than 1p; this is expected (cf. Warren & Gibson 2005) and not crucial here.)

Response-contingent analysis: However, could 3p-1p slowdowns be driven by the subset of trials where participants (unexpectedly) chose the matrix subject (violated Blocking)? If we analyze trials where participants chose the local subject (Blocking is effective), will we still see competition from the matrix subject? When we look only at trials with local-antecedent-interpretations, slowdowns in 3p-1p are no longer significant.

In sum, (i) 1p does not consistently block access to the matrix subject, (ii) but when it does (Blocking is effective), Blocking is strong enough to reduce competition from the matrix subject.

Experiment 2 Results

Antecedent choices: Again we found a strong bias for local subjects (3p-3p: 87% local subject, 2p-3p: 100%, 3p-2p: 93.8%, *p*'s<0.001). 3p-2p shows a stronger local preference than 3p-1p in Exp1 (p<.005): An intervening 2p produces a more consistent blocking effect than 1p.

RTs: 3p-2p was not significantly slower than 2p-2p (marginally slower, p=.069), suggesting 3p is not accessible enough to cause a significant slowdown. As a whole, it appears that an intervening 2p may be a stronger blocker (and may reduce competition more) than an intervening 1p.

Conclusions

Our results suggest that person-feature blocking can reduce competition during on-line processing, but that 2p pronouns block more effectively than 1p pronouns, as illustrated by antecedent choices and reading times. On the basis of Brunyé et al.'s (2009) findings, this difference in patterning between 1st person pronoun and 2nd person pronoun may be caused by perspective taking: We suggest that identifying with the addressee (2p pronoun) leads comprehenders to more consistently interpret the reflexive as referring to the addressee (local 2p subject), resulting in a consistent blocking effect.

References

Baayen, R.H., Davidson, D.J. and Bates, D.M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 390-412.

Badecker, W. and Straub, K. 2002. The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory and Cognition* 28, 748-769.

Brunyé, T.T., Ditman, T., Mahoney, C.R., Augustyn, J.S. and Taylor, H.A. 2009. When you and I share perspectives: pronouns modulate perspective taking during narrative. *Psychological Science* 20, 27-32.
Cole, P. and Wang, C. 1996. Antecedents and blockers of long-distance reflexives: the case of Chinese *ziji*. *Linguistic Inquiry* 27(3), 357-390.
Huang, C.-T. J. and Liu, C.-S. L. 2001. Logophoricity, attitudes and ziji at the interface. In *Syntax and Semantics* [Long Distance Reflexives 33], Cole, Petter, Hermon, Gabriella and Huang, C.T. James (eds.), 141-195. New York: Academic Press.
Jaeger, T.F. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language* 59, 434-446.
Nicol, J. and Swinney, D. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research* 18, 5-19.
Pan, H. 2000. Why the blocking effect? In *Syntax and Semantics* [Long Distance Reflexives 33], Cole, Petter, Hermon, Gabriella and Huang, C.T. James (eds.), 279-316. New York: Academic Press.
Pollard, C. and Xue, P. 1998. Chinese reflexive *ziji*: syntactic reflexives vs. nonsyntactic reflexives. *Journal of East Asian Linguistics* 7, 287-318.
Sturt, P. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language* 48, 542-562 .
Warren, T. and Gibson, E. 2005. Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes* 20(6), 751 -767.

Martin Hilpert
Freiburg Institute for Advanced Studies
martin.hilpert@frias.uni-freiburg.de

Watching language change: A new visualization technique and its theoretical implications

Introduction

In recent years, many historical corpora have become available that allow the study of linguistic change in real time (de Smet 2005, Davies 2008, 2010, inter alia). Although these resources have caused a surge of interest in diachronic corpus linguistics, many studies still limit themselves to analyses of rising or declining text frequencies. Meanwhile, diachronic corpora have much more information to offer. The present paper illustrates this with a new corpus-linguistic application – the flipbook technique.

The flipbook technique creates dynamic visualizations of language change. Like its real-world model, a flipbook of language change consists of a series of pictures in which each picture slightly differs from the previous one. Based on data from COHA (Davies 2010), two case studies illustrate recent changes in American English. The first study visualizes structural change in the behaviour of complement-taking predicates; the second one shows semantic change in the paradigm of modal auxiliaries.

Methodologically, flipbooks are created on the basis of data representing the same linguistic phenomenon across fifteen sequential decades. The data for each decade is submitted to a multi-dimensional scaling analysis (Kruskal & Wish 1978) that produces a two-dimensional plot. The overall result is a series of fifteen plots that gradually change from one to the next, thus revealing how a linguistic phenomenon has developed diachronically. It is argued that flipbooks of MDS plots offer an effective representation of complex linguistic changes that can not only guide exploratory analyses, but also speak to theoretical issues.

An illustration: complement-taking predicates

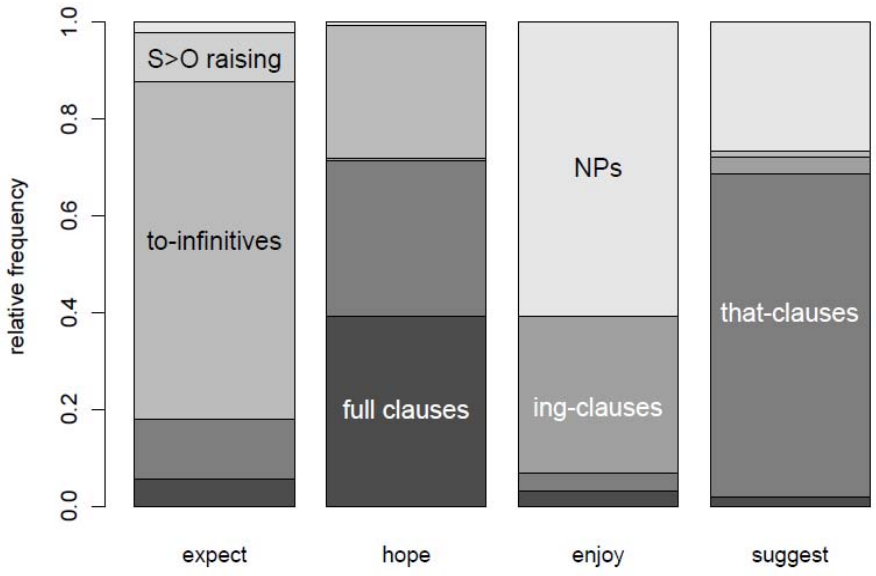
Verbs such as *expect*, *like*, or *imagine* project a syntactic complement structure that may take different shapes. The phrase *I expect* can be followed by phrases such as *a visitor*, *to hear from John*, or *that John will win*, amongst others. The present analysis distinguishes between the six different subcategorization frames shown in (1).

(1)

full clauses	I suggest we do nothing.
that-clauses	I think that John will win.
ing-clauses	I enjoy knitting sweaters.
to-infinitives	I expect to hear from John.
subject-to-object raising	I want John to be our next president.
noun phrases	I hate broccoli.

For 46 different verbs, the relative frequencies of these subcategorization frames have been retrieved from COHA. Figure 1 shows that complement-taking predicates exhibit a particular profile with regard to their preferred subcategorization frames: *expect* chiefly occurs with *to*-infinitives, whereas most examples of *enjoy* contain either an NP or an *ing*-clause.

FIGURE 1. RELATIVE FREQUENCY PROFILES FOR FOUR COMPLEMENT-TAKING PREDICATES.

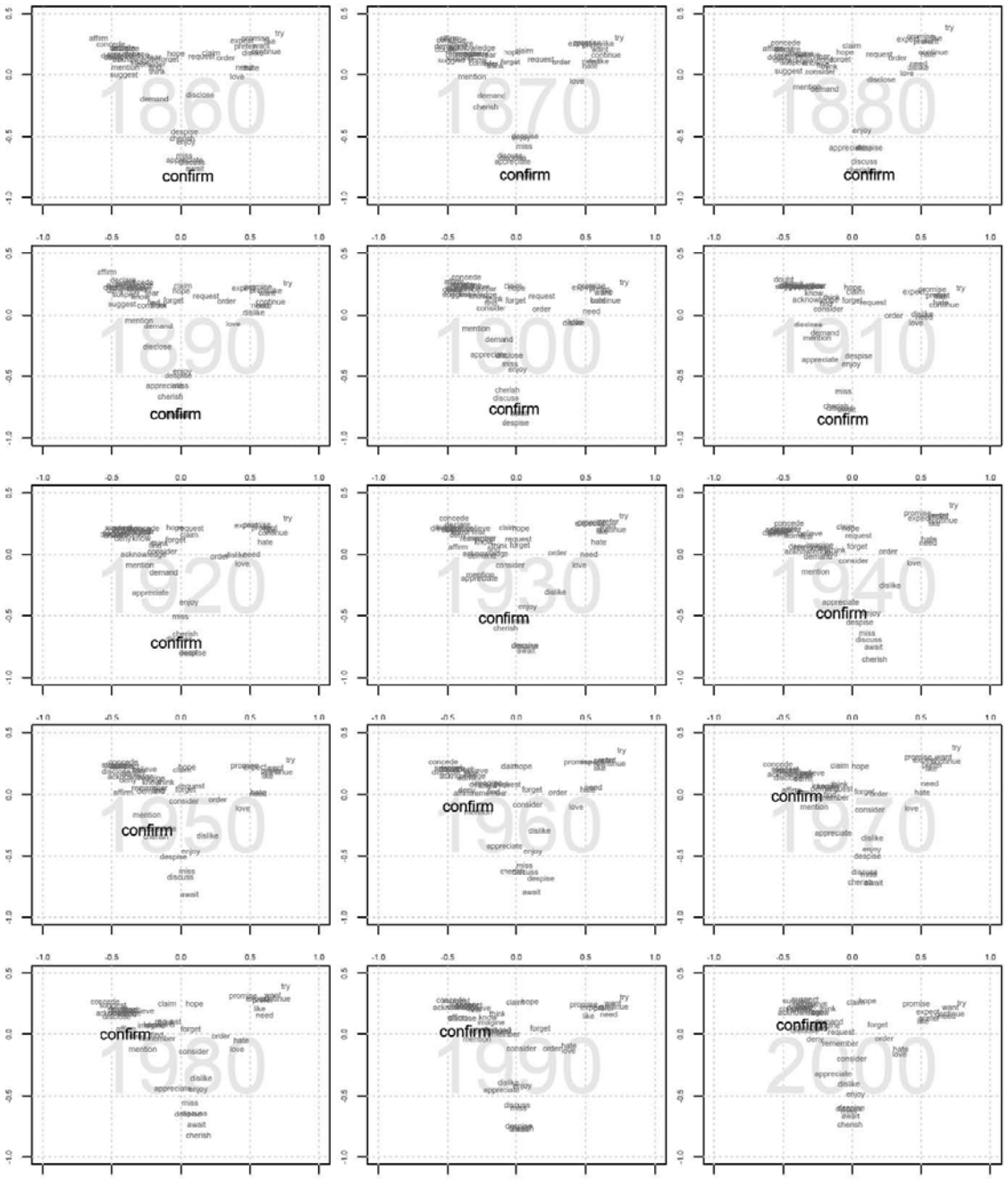


With such frequency profiles, MDS can determine the relative similarities between larger sets of complement-taking predicates. This is done separately for the data from each COHA decade, such that the overall result is the series of 15 MDS maps shown in Figure 2. The graphs clearly distinguish between verbs with a preference for *that*-clauses (in the upper left), verbs with a preference for *to*-infinitives (upper right), and verbs that primarily take NP complements (lower middle). Amongst other things, the flipbook shows that the verb *confirm* abandons a preference for NP complements in favour of *that*-clauses from the 1930s onwards.

Theoretical implications

Flipbook representations allow the viewer to take in a large amount of information in an effective way. The fine detail that is available can be put into the service of theoretical generalizations, as the researcher is required to understand individual patterns of stability and change in relation to the broader scenario in which they are embedded. Flipbooks can thus help address questions of linguistic theory: Do individual linguistic units change spontaneously, or as the result of paradigmatic pressure? Is a grammatical domain becoming more fragmented or more tightly structured? How are new members integrated into an existing grammatical paradigm? The flipbook technique is thus a tool that can advance diachronic corpus linguistics both methodologically and theoretically.

FIGURE 2. FROM *CONFIRM NP* TO *CONFIRM THAT-CLAUSE*.



References

Davies, M. 2008. The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. http://www.americancorpus.org.

Davies, M. 2010. The Corpus of Historical American English (COHA): 400+ million words, 1810-2009. http://corpus.byu.edu/coha.

De Smet, H. 2005. A corpus of Late Modern English texts. *ICAME Journal* 29, 69-82.

Kruskal, J. and M. Wish. 1978. *Multidimensional Scaling*. London: Sage.

Graham Katz, Roberto Zamparelli
Georgetown University, Universitá degli Studi di Trento
egk7@georgetown.edu, roberto.zamparelli@unitn.it

Meaning-shifting plurality and the Count/Mass Distinction

Introduction

The semantics of plurality is a recurring theme in the formal linguistics literature (Link 1983; Krifka 1989; Schwarzschild 1996; Landman 2000; Heycock & Zamparelli 2005). With the exception of a few frequently-cited idiosyncratic cases (*brain/brains*, *glass/glasses*, see (Acquaviva 2008)) it is assumed that the meaning of a plural noun is closely related to that of the corresponding singular. The goal of this talk is to test this assumption using the quantitative tools of vector space (or "distributional") semantic models (DSM) (Landauer & Dumais 1997; Baroni & Lenci 2010). When applied to the singular and plural forms of the same word, these methods clearly show that there is a large proportion of nouns whose distribution in the singular and the plural differs, in terms of neighboring content words. In these nouns, plurality seems to come with a meaning shift that goes beyond number, and which is detected by the DS methods.

Further analyses show that the shift in the interpretation of plurals correlates with the *countability* of the noun. Nouns with closely related plural and singular forms tend to be count nouns (nouns that occur in canonical *count* contexts (Baldwin & Bond 2003)), nouns with meaning-shifted plurals tend to be mass nouns. The experiment thus offers a novel angle for examining the notoriously elusive count-mass distinction (Quine 1960; Pelletier & Schubert 1989; Chierchia 2010), and a clear example of what many DSMs miss by studying word distributions in terms of their *lemma* alone.

Singular/Plural Distance

To study noun-number semantic differences, we built a vector space model of the 100 million token UKWAK-1 corpus (Baroni et al. 2009) for the 20,000 most frequent content words in the corpus, using the COALS algorithm (Rohde et al. 2005). Our model maintained both part of speech and lemma information. We analyzed the 2131 noun-types which appeared in both singular and plural forms in the model.

We chose two ways to examine semantic proximity. First using a vector cosine measure and then using a word-based measure. The 25 nouns with lowest and highest singular/plural cosine-similarity are listed below:

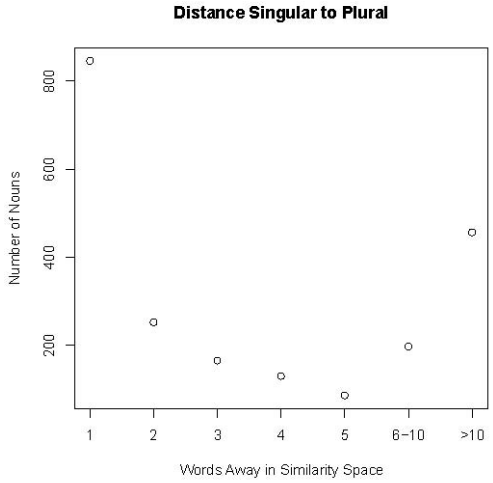
Low plural/singular similarity: *leave make creator con humanity extreme good disadvantage toddler strength fortune horizon total story hip mouse dozen tip monkey security term medium support manner custom*

High plural/singular similarity: *phone therapist resort impairment reaction list century speaker cookie engine locomotive pool sector cancer bomber venture guitar beach int examiner thou squirrel tale surgeon acid*

As we see, the nouns with lowest sing/plural similarity contain ambiguous nouns where different meanings are associated with one of the two forms (e.g. *mouse/mice*, *good/goods*, *security/securities*).

To investigate this contrast, we used the cosine-similarity metric to determine how many words in the model were distributionally closer to the singular than the corresponding plural was. In our model, *boxes* was the nearest neighbor of *box*, but *flight* comes between *airline* and *airlines*, and *arms* was more then 10 words away from *arm*. In Figure1 we display the distribution of nouns by distance in words:

FIGURE1. WORD DISTANCE.



For more than a third of the nouns under study the closest element in the model to the singular form is, as expected, the corresponding plural form (we call these the "near" group). Interestingly, however, for more than a fifth of the nouns the plural isn't among the 10 closest neighbors of the singular (we call these the "far" group). The average sing/plural cosine similarity of "far" group is 0.18, while the average similarity for "near" group is 0.50.

Mass/Count

In the formal literature, it is often noted (Carlson 1977; Chierchia 1998; Rothstein 2010) that mass nouns such as *wine* undergo a semantic shift when they are used in count contexts or pluralized (*two wines* means two kinds of wine). This suggests the hypothesis that the "far" nouns would be more count nouns and that the near nouns would be more mass nouns. To investigate this, we examined the rate of occurrence of the nouns in the sample in typical mass contexts (e.g. with *much*) and in typical count contexts (e.g. with *every*).

Looking first at mass contexts we find that the average rate of mass contexts occurrence (by noun type) for the ’far’ nouns is significantly higher than that for ’near’ nouns (0.0028 vs. 0.0014,p = 0.003041). Even more dramatic is the difference in the distribution of count-context rates between the two groups: The difference between group means (0.192 vs. 0.149) is highly significant (p = 7.09e-09). In short, the ’near’ group contains more predominantly count nouns than the ’far’ group.

FIGURE 2. RATE OF COUNT CONTEXTS FOR ‘NEAR’ NOUNS.

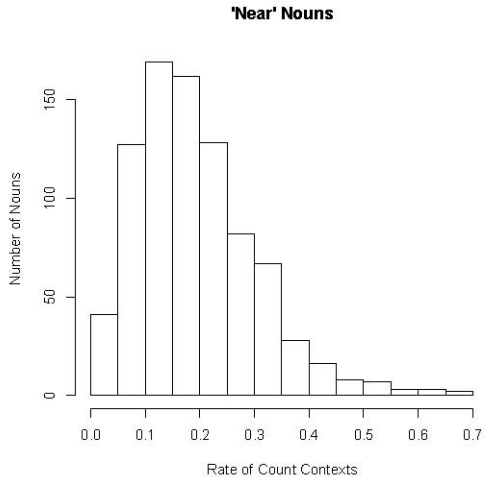
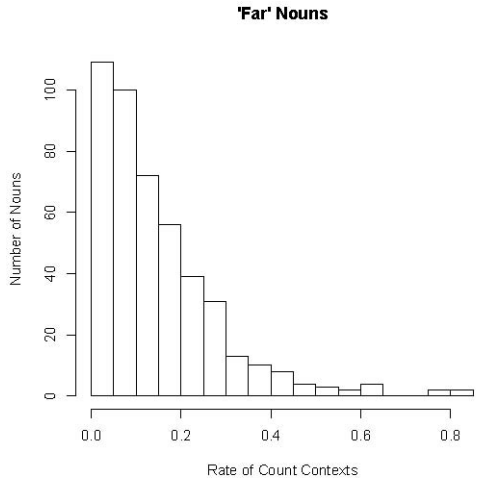


FIGURE 3. RATE OF COUNT CONTEXTS FOR ‘FAR’ NOUNS.



Conclusions

It is clear, then, that in building vector-space semantic models, some word-form information should be maintained. Collapsing lemmas in vector-space models ignores important semantic information. This is particularly the case for the class of mass nouns.

References

Acquaviva, P. 2008. *Lexical plurals: a morphosemantic approach*. Oxford University Press, USA.

Baldwin, T. and Bond, F. 2003. Learning the countability of English nouns from corpus data. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 463-470, Morristown, NJ, USA. Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209-231.

Baroni, M. and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*.

Carlson, G. N. 1977. *Reference to Kinds in English*. PhD thesis, University of Massachusetts, Amherst, Amherst, MA.

Chierchia, G. 1998. Plurality of mass nouns and the notion of semantic parameter. In *Events and grammar*, Rothstein, S. (ed.). Kluwer Academic, Dordrecht.

Chierchia, G. 2010. Mass nouns, vagueness and semantic variation. *Synthese* 174(1), 99-149.

Heycock, C. and Zamparelli, R. 2005. Friends and colleagues: Plurality, coordination, and the structure of DP. *Natural language semantics* 13(3), 201-270.

Krifka, M. 1989. Nominal reference, temporal constitution and quantification in event semantics. In *Semantics and Contextual Expression*, Bartsch, R., van Benthem, J.,and van Emde Boas, P. (eds.). Foris Publications, Dordrecht.

Landauer, T.K. and Dumais, S.T. 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.

Landman, F. 2000. *Events and plurality: The Jerusalem lectures*. Kluwer Academic Publishers.

Link, G. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. *Meaning, Use, and Interpretation of Language*, 302-323. de Gruyter, Berlin.

Pelletier, F. and Schubert, L. 1989. Mass expressions. *Handbook of philosophical logic*, 4:327-407.

Quine, W.V.O. 1960. *Word and Object*. The MIT Press, Cambridge, MA.

Rohde, D., Gonnerman, L. and Plaut, D. 2005. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. Unpublished Manuscript. Available online at http://tedlab.mit.edu/~dr/Papers/RohdeGonnermanPlaut-COALS.pdf [March 05, 2011]

Rothstein, S. 2010. Counting and the Mass/Count Distinction. *Journal of Semantics* 27(3), 343.

Schwarzschild, R. 1996. *Pluralities*. Dordrecht: Kluwer.

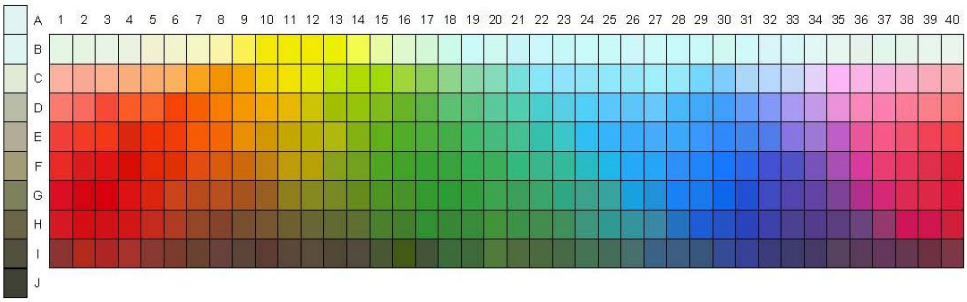
Gerhard Jäger
University of Tübingen, Institute of Linguistics
gerhard.jaeger@uni-tuebingen.de

Color naming universals: A statistical approach

The paper describes a quantitative investigation of the distribution of color naming systems across the languages of the world, using the data from the World Color Survey (WCS, see Cook et al. 2005 for details). Working with data from individual speakers, we found that (1) different categorization systems are distributed according to a power law, and (2) the systems of possible color naming patterns proposed in the literature so far only provide an imperfect description of the data. We (3) propose an alternative system of universals that provide a better fit of the empirical findings.

The WCS researchers collected field research data for 110 unwritten languages, working with an average of 24 native speakers for each of these languages. During this investigation, the Munsell chips were used, a set of 330 chips of different colors that cover 322 colors of maximal saturation plus eight shades of gray. Figure 1 displays them in form of the Munsell chart.

FIGURE 1. THE MUNSELL CHART.



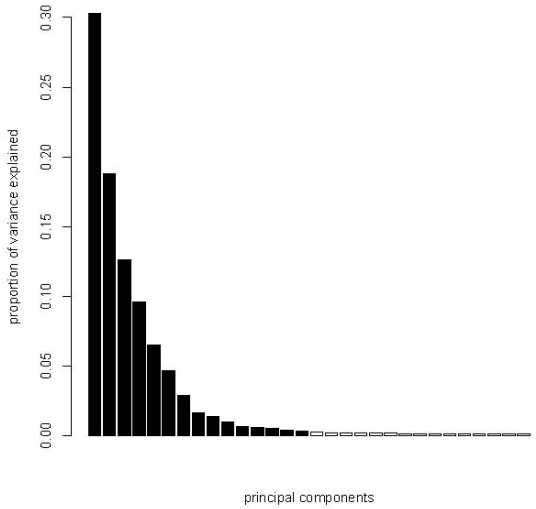
For the WCS, each test person was "asked to name each of 330 Munsell chips, shown in a constant, random order" (quoted from the WCS homepage). The data from this survey are freely available from the WCS homepage http://www.icsi.berkeley.edu/wcs/data.html.

For each informant, the outcome of the categorization task defines a partition of the Munsell space into disjoint sets — one for each color term from their idiolect.
For the present study, these data were organized in a contingency table. It has 1,601 rows — one for each term that was used by at least one of the 1,771 test persons, and 330 columns — one for each Munsell chip. Each cell contains the number of test persons that used the term corresponding to the row to name the chip corresponding to the column.

To normalize vector lengths while at the same time preserve the statistical weight of often-used color terms, each row was divided by the number of speakers that used the corresponding term at least once, and afterwards copied as many times as there were speakers that used the corresponding term.

Figure 2 depicts the proportion of the total variance in the data that are explained by the principal components. The first 15 principal components jointly explain about 91.6% of the total variance in the data. After applying the Varimax algorithm, the resulting 15 extracted features have a clear interpretation: green, white, red, yellow, black, blue, purple, pink, brown, light blue, olive green, gray, orange, violet, pastel green. Further details on the feature extraction procedure can be found in Jäger (2010).

FIGURE 2. PROPORTION OF TOTAL VARIANCE EXPLAINED BY PRINCIPAL COMPONENTS.



The first six features thus extracted correspond to the primary colors that play a central role in Kay et al.'s (1997) system of color naming universals. They assume that a system of basic color terms partitions the color space into disjoint but jointly exhaustive regions. It is claimed that almost all languages partition the primary colors into one of the systems given in Table 1 (above the horizontal line).

For each informant, each term is represented as a 15d vector after dimensionality reduction. Likewise, each primary color is a vector in this low-dimensional space. Thus the data for each speaker implicitly define a partition over the six primary colors: each primary color is assigned to (the vector representing) the term with which it has the largest cosine, i.e. the largest similarity.

The frequencies of the different partition types are distributed approximately according to a power law, as can be seen from Figure 3. The partition types are ordered according to their frequency. The distribution almost follows a straight line. This means that it can be approximated by a power law

$$fr \sim r^c,$$

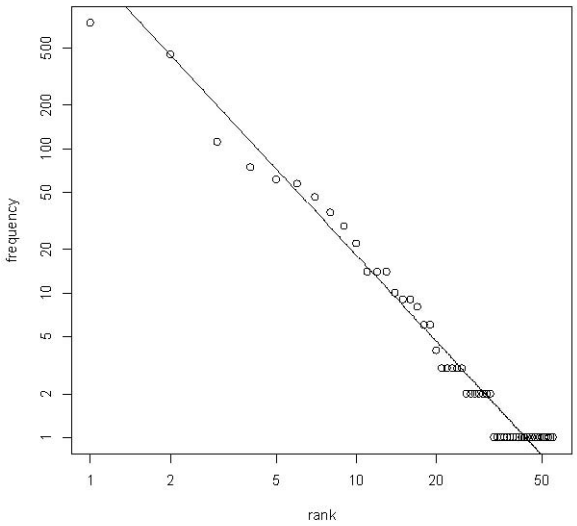
where *fr* is the frequency, *r* is the rank, and *c* is a constant coefficient. Using the methods described in Clauset et al. (2009), we estimate $c \approx -1.99$.

All of the partition types from Table 1 except the leftmost one are attested. The frequencies of each type are indicated in the table. Additionally, there are two partition types not mentioned by Kay et al. (1997) (or any other author that I would be aware of) which occur in substantial numbers. They are displayed in the table below the horizontal line.

TABLE 1. PARTITION HIERARCHY ACCORDING TO (KAY ET AL. 1997).

I	II	III	IV	V
[white/red/yellow black/green/blue] 0	[white red/yellow black/green/blue] 111	[white red/yellow green black/blue] 7		[white red yellow green blue black] 446
		[white red yellow black/green/blue] 57	[white red yellow green black/blue] 79	
		[white red yellow/green/blue black] 29	[white red yellow/green blue black] 9	
		[white red yellow/green black/blue] 22		
	[white/yellow red green/black/blue] 36	[white/yellow red green/blue black] 61		
		[white red/yellow green/blue black] 45	[white red yellow green/blue black] 737	

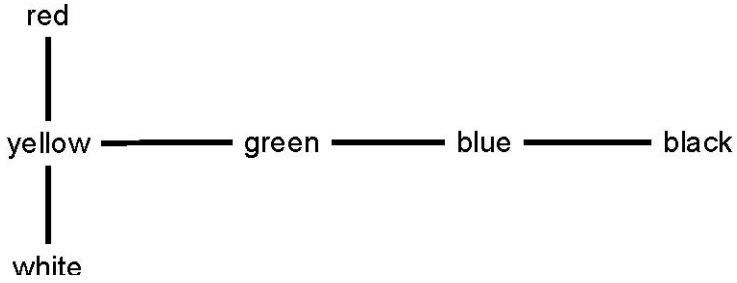
FIGURE 3. FREQUENCY DISTRIBUTION OF PARTITION TYPES OVER THE PRIMARY COLORS.



So while Kay et al.'s model is not completely off the mark, the data from the WCS do not confirm it very strongly either. Only 87.1% of all informants conform to their model. The ten partitions that, according to their model, are the only ones possible occupy the ranks 1, 2, 3, 4, 6, 7, 9, 10, 16, and 18 in the list of the attested partition types.

A closer manual inspection of the partitions in Table 1 shows a certain pattern though. All partition cells are continuous sub-graphs of the connection graph given in Figure 4. Also,

FIGURE 4. CONNECTION GRAPH.



all attested partitions in Table 1 obey the following constraints:

(1) a. All partition cells are continuous sub-graphs of the connection graph.

b. No partition cell has more than three elements.

c. *Red* and *white* only occur in cells with at most two elements.

Next to the 12 partitions in Table 1, there are three partitions obeying these constraints, which are all attested in the data:

- {green}, {white/yellow}, {red}, {black/blue} (14 occurrences)
- {green}, {white/yellow}, {red}, {black}, {blue} (8 occurrences)
- {green}, {white}, {red/yellow}, {black}, {blue} (2 occurrences)

About 94% of all data points in the WCS are captured by this model, and all partition types that are predicted to be possible are in fact attested.

References

Clauset, Aaron, Shalizi, Cosma Rohilla and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661-703.

Cook, Richard, Paul Kay and Terry Regier 2005. The world color survey database: History and use. In *Handbook of Categorisation in the Cognitive Sciences*, Cohen, Henry and Lefebvre, Claire (eds.), 223-242. Amsterdam: Elsevier.

Jäger, Gerhard 2010. Natural color categories are convex sets. In *Logic, Language and Meaning. 17th Amsterdam Colloquium*, Aloni, Maria, Bastiaanse, Harald, de Jager, Samson Tikitu and Katrin Schulz (eds.) 11-20, New York/Heidelberg: Springer.

Kay, Paul, Berlin, Brent, Maffi, Luisa and Merrifield, William 1997. Color naming across languages. In *Color categories in thought and language*, Hardin, C. L. and Luisa Maffi (eds.) 21-58, Cambridge: Cambridge University Press.

Gerrit Kentner
Institut für Linguistik
Goethe-Universität Frankfurt am Main
gerrit@lingua.uni-frankfurt.de

Rhythm-Syntax interaction in silent reading

In written text comprehension, the task of the reader is to assign the printed word sequence a sufficiently coherent syntactic structure to allow semantic analysis. At the same time, readers generate from the graphemic string an intrinsic auditory version of the text, entailing rich prosodic structure. Various reading studies have revealed that this 'implicit prosody' may affect the syntactic analysis of written text (e.g. Bader 1998, Hwang & Schafer 2009). The role of 'implicit prosody' in written sentence comprehension, however, has been described as paradoxical (Fodor 2002): on the one hand, prosody is shaped according the syntactic structure of the word string, suggesting that the syntactic analysis determines the prosodic representation; on the other hand, experimental evidence attests a clear influence of 'implicit prosody' on the syntactic analysis proper. The following research questions guide our attempt to clarify the syntax– prosody interaction in reading:

1. At what processing stage do prosodic factors constrain the incremental syntactic analysis?
2. How can the syntax–prosody interaction be embodied in a performance model.
3. How does the performance model relate to competence grammar?

In a controlled reading experiment, using sentences like (1), we tested the influence of stress-based linguistic rhythm on syntactic ambiguity resolution.

(1) a. ...nicht mehr {NACHweisen, erMITteln} kann, wer der Täter war.
...couldn't {prove, determine} anymore who the culprit was.

b. ...nicht MEHR {nachweisen, ermitteln} kann, als die Tatzeit.
...couldn't {prove, determine} more than the date of the crime.

The ambiguity concerns the word *mehr* featuring either an unaccented temporal adverbial (1-a) or an obligatorily accented, comparative complement to the verb (1-b).

The structures are disambiguated at the subordinate clause that ends the sentence. The syntactic factor was crossed with the rhythmic-prosodic environment which was systematically varied at the verb following *mehr*, featuring either initial or medial stress. Accented comparative *mehr* followed by a verb with initial stress in the citation form involves a stress clash (i.e. adjacency of two syllables carrying prosodic prominence), violating the (supralexical) prosodic constraint *Clash.

Eye-tracking data from a silent reading experiment shows that readers exhibit significantly higher processing difficulties at the disambiguating clause in comparative versions (1-b) when the verb features initial stress. This suggests that readers initially computed the unaccented – and hence temporal – *mehr* in this condition to avoid the stress clash. The results demonstrate that, at points of syntactic underspecification, the accruing prosodic representation may affect even the earliest stages of structure building in reading, viz. the analysis of syntactic features on the ambiguous word. Such an effect remains inexplicable in the context of (psycho-) linguistic theories that assume a strictly unidirectional relationship between syntactic and phonological processes, the latter merely interpreting the conditions the syntactic component imposes on it.

The performance data are modeled as an incremental constraint satisfaction process in the framework of an OT parsing account (cf. Fanselow et al. 1999, Hoeks & Hendriks, to appear). Solely making use of constraints from competence grammar, the model is capable of capturing the processing data and advocates the simultaneous application of syntactic, prosodic and syntax-phonology interface constraints in incremental processing. The OT grammar/processor integrates syntactic parsing and prosodification in reading, hence dissolving the strict separation of language production and comprehension. The model predicts that, at points of syntactic indetermination, weak prosodic constraints alone may guide syntactic structure assignment. Consequently, the OT model endorses a bidirectional relationship between syntax and phonology in grammar as well as in processing.

References

Bader, M. 1998. Prosodic influences on reading syntactically ambiguous sentences. In *Reanalysis in Sentence Processing*, Fodor and Ferreira (eds.). Dordrecht: Kluwer.

Hwang, H. and Schafer, A. 2009. Constituent length affects prosody and processing of for a dative NP ambiguity in Korean. *Journal of Psycholinguistic Research* 38(2).

Fodor, J. 2002. Psycholinguistics cannot escape prosody. In *Proceedings of Speech Prosody*. Aix-en-Provence.

Fanselow, G., Schlesewsky, M., Cavar, D. and Kliegl, R. 1999. *Optimal Parsing – Syntactic Parsing Preferences and Optimality Theory*. ROA-367.

Hoeks, J. and Hendriks, P., to appear. Optimality Theory and human sentence processing: Towards a cross-modular analysis of coordination. Ms. Center for Language and Cognition, Groningen.

Sudheer Kolachina, Taraka Rama, Lakshmi Bai B.
LTRC, IIITHyderabad and Språkbanken, Department of Swedish Language, University of Gothenburg
sudheer.kpg08@research.iiit.ac.in, taraka.rama.kasicheyanula@svenska.gu.se, lakshmi@iiit.ac.in

Maximum parsimony method in the subgrouping of Dravidian languages

Introduction

Historical linguistics has as one of its main aims, the classification of languages into *language families*. The internal classification of languages within a language family is known as *subgrouping*. Subgrouping is concerned with the way *daughter languages* within a single family are related to one another and therefore, with the branching structure of the family tree (Campbell 2003).

In recent years, there has been a rapid increase in interest in the application of phylogenetic inference methods to diachronic language data leading to the emergence of Computational historical linguistics as a distinct field within historical linguistics. The basic intuition underlying such research is that these methods which can infer phylogeny from gene sequences can do so from language data too which also consist of sequences (Atkinson & Gray 2004).

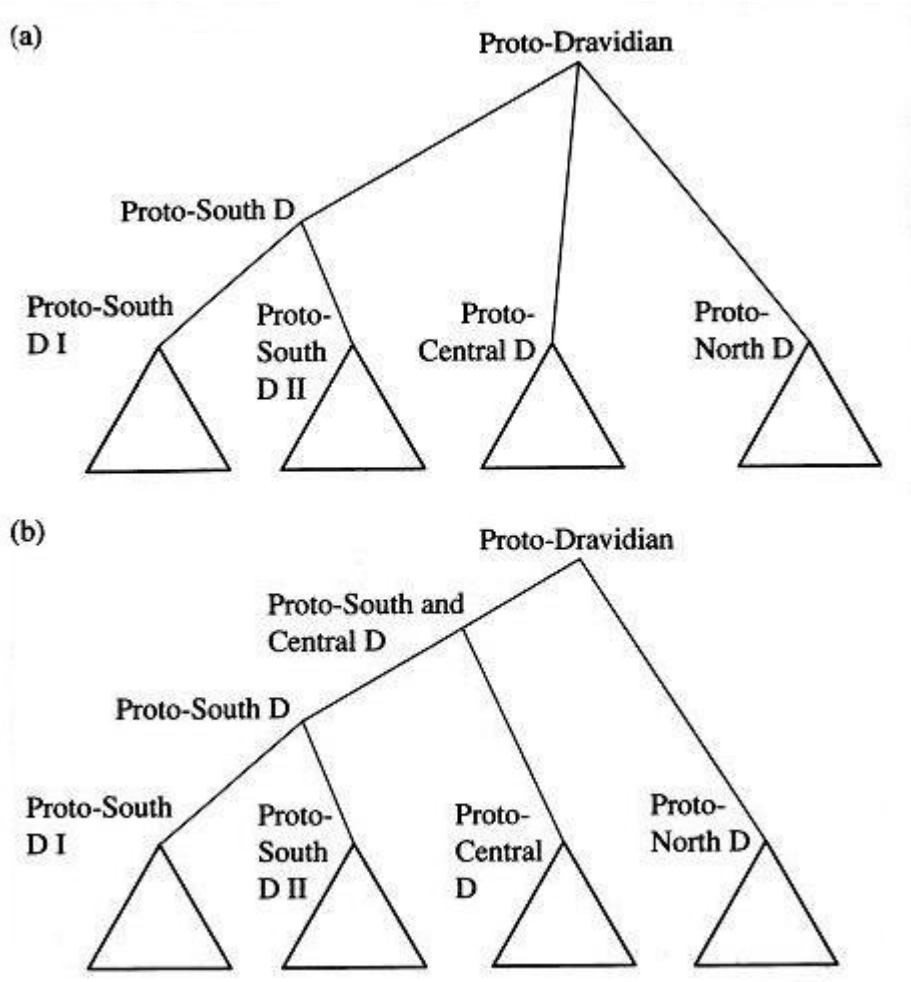
Subgrouping of Dravidian languages

The Dravidian language family consists of 26[1] languages spoken by over 200 million people in South Asia making it the world's fifth largest language family (Krishnamurti 2003). Most of these languages are geographically located in the southern and the central parts of the India with a few scattered pockets in Northern India (Kurux, Malto) and Nepal (Kurux) and a lone population in Pakistan (Brahui).

Krishnamurti (2003) is a compendious work which extensively covers various aspects of Dravidian languages. In this work, two alternative subgroupings of the Dravidian languages are proposed. These alternatives numbered (a) and (b) are shown in Figure 1. The subgrouping adopted in Krishnamurti (2003) is the alternative (a). As can be seen from Figure 1 (a), this subgrouping alternative is a ternary branching structure. In other words, Proto-Dravidian (PD) has three branches: Proto-North Dravidian (ND), Proto-Central Dravidian (CD) and Proto-South Dravidian (SD) which is further split into South Dravidian I (SD I) and South Dravidian II (SD II). This subgrouping is established on the basis of isogloss maps constructed using 27 features from comparative phonology and morphosyntax. It is possible to conceive of a binary division of Proto-Dravidian (shown as alternative (b) in Figure 1) into Proto-North Dravidian (ND) and Proto-South and Central Dravidian (SCD). In this regard, Krishnamurti (2003) notes that although in general, a binary division of a speech community is more likely than a ternary, there is lean evidence to set up a common stage of South and Central Dravidian.

[1] 27 if Naikri is treated as distinct from Naiki.

FIGURE 1. DRAVIDIAN SUBGROUPING ALTERNATIVES.



The aim of this present work is to address this specific question of ternary versus binary branching of Proto-Dravidian through the application of the maximum parsimony method for phylogenetic inference to the comparative feature dataset used by Krishnamurti (2003) in the subgrouping. Evidence from this feature data in support of the binary division is claimed to be lean but that is only when the traditional method of subgrouping based on isogloss maps is followed. It is possible to hypothesize that the application of a different subgrouping method (which relies on a different kind of 'evidence') to the same feature data can result in the setting up of a Proto-South and Central Dravidian stage.

Maximum Parsimony method for phylogenetic inference

The Maximum Parsimony (MP) method is a well-known discrete character-based method which takes as input character sequences. MP is an optimization problem which seeks a tree on which a minimum number of character state changes occurs (Nakhleh et al. 2005b). MP is an NP-hard problem and therefore, exact solutions cannot be guaranteed within polynomial time. As such, heuristics need to be applied to find good (though not provably optimal) solutions. There can be many equally good solutions and a single solution is obtained by applying a consensus method to the Kbest output. Maximum parsimony methods have been claimed to be the most efficient for inferring

the phylogenetic tree that is closest to the traditional standard tree (Ringe et al. 2006; Nakhleh et al. 2005b; Barbancon et al. 2007).

Since the aim of our study is to address the specific issue of binary versus ternary branching of Proto-Dravidian, a method to be qualified for use in our study must be one that at least searches over the sets of binary and ternary branching trees to find the most parsimonious tree. After going through implementations of maximum parsimony available in the PAUP* (Swofford 2002) and Phylip (Felsenstein 2003) packages, we found that only the *pars* program in the Phylip package meets the requirements of our study as it searches over a tree space consisting of both bifurcating and multifurcating trees.

Experimental Results and Discussion

Krishnamurti (2003) contains four sets of features from comparative phonology, morphology and syntax that support the subgrouping adopted in that work (Figure 2(a)). We use the same datasets in our experiments[2]. In order to apply any phylogenetic inference method to these datasets, the feature data are encoded as character data (summarized in Table 1 below). Following Maddison (1993), we make a distinction between missing characters (feature not relevant) as opposed to missing data and code them using two distinct characters.

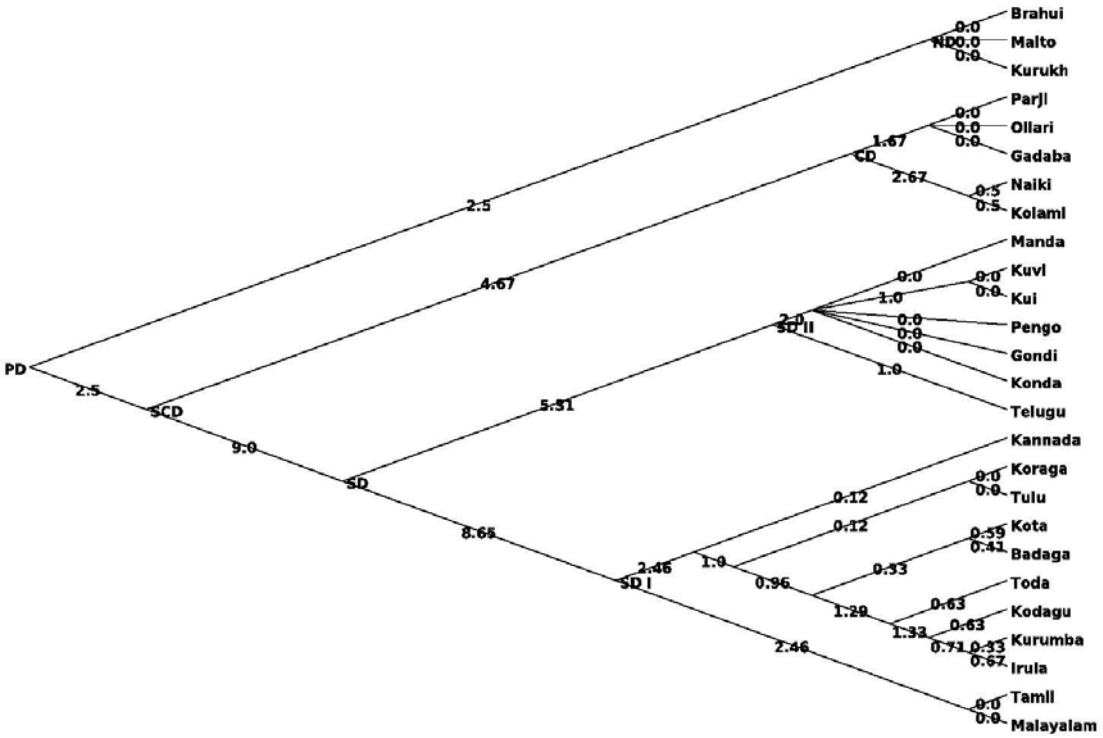
TABLE 1. FEATURES USED IN OUR EXPERIMENTS.

Feature type	Phonology	Nominal morphology	Verbal morphology	Syntax
# features	13	9	13	5

In order to guard against statistical bias, bootstrapping procedure was run for 10000 times with 'sampling with replacement' using the *seqboot* program in PHYLIP. The *pars* program was applied to these multiple datasets to find the most parsimonious trees from each set. The consensus tree was estimated from all these parsimonious trees using majority consensus (*consense*). The consensus tree obtained is the single most parsimonious tree containing edges annotated with their support values. Next, branch lengths on the consensus tree were re-estimated using the *pars* program. Finally, the unrooted tree returned by *pars* was rooted using the North Dravidian (ND) clade as the outgroup. The rationale behind using ND as the outgroup is that both the subgrouping alternatives agree on ND being the first to diverge from Proto-Dravidian. The phylogenetic tree inferred from the character sequences after rooting is shown below in Figure 2. The branches in the tree are annotated with the branch lengths returned by the *pars* program. The internal nodes are labeled with names of known subgroups they represent.

[2] The datasets are available at the following link: https://docs.google.com/leaf?id=0B6U29M4CJtXXODJkZWM4MGQtOThiMC00NDZiLTgxZGYtMTRiMGYwMTE 0YjVj&hl=en

FIGURE 2. PHYLOGENETIC TREE OF DRAVIDIAN LANGUAGES INFERRED USING MAXIMUM PARSIMONY.



On the question of binary versus ternary branching of Proto-Dravidian, the phylogenetic tree shows a binary branching structure. However, it must be noted that caution needs to be exercised while discriminating a binary branching from a ternary branching. This is because a ternary branching may also be shown as a binary branching. In such a situation, the clue for determining whether a branching is ternary or binary can come from the number of state changes that take place along the branches, indicated by the branch lengths estimated by *pars*. If the number of state changes along two branches coming out of an internal node shown with binary branching are equal, such an internal node can be eliminated and the tree structure at that point can be considered as a ternary branching from its parent node. Now, applying this criterion to the internal node corresponding to Proto-South Central Dravidian (SCD) in the tree (Figure 2 above), we observe that the difference in branch lengths between SCD and SD, and SCD and CD is 4.33. This shows that the internal node corresponding to SCD cannot be eliminated. This in turn, implies a binary rather than a ternary split of Proto-Dravidian and is in agreement with the general understanding about evolution of speech communities.

References

Atkinson, Q.D. and Gray, R.D. 2005. Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54, 513-526.

Barbancon, F.T., Warnow, T., Evans, S.N., Ringe, D. and Nakhleh, L.. 2007. *An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods.* Technical Report 732. Department of Statistics, University of California, Berkeley.

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

Felsenstein, J. 2003. *Inferring Phylogenies.* Sunderland, MA: Sinauer Associates, 664.

Krishnamurti, Bh. 2003. *The Dravidian Languages.* Cambridge University Press, Cambridge, UK.

Maddison, Wayne P. 1993. Missing data versus missing characters in Phylogenetic analysis. *Systematic Biology*. 42(4), 576-581.

Nakhleh, L., Warnow, T., Ringe, D. and Evans, S.N. 2005b. A Comparison of Phylogenetic reconstruction Methods on an Indo-European Dataset. *Transactions of the Philological Society* 3(2), 171-192.
Nichols, J. and Warnow, T. 2008. Tutorial on computational linguistic phylogeny. *Linguistics and Language Compass* 2(5), 760-820.
Ringe, D., Warnow, T. and Evans, S. 2006. Polymorphic characters in Indo-European. In *Indo-European Languages, Languages and Genes*, September 2006.
Swofford, D.L. 2002. PAUP*. *Phylogenetic analysis using parsimony (and other methods).* Version 4. Sunderland, MA: Sinauer Associates.

Karin Madlener
Albert-Ludwigs-Universität Freiburg,
Research training group GRK DFG 1624/1 "Frequency effects in language"
karin.madlener@frequenz.uni-freiburg.de

Developing productivity with a new construction: Are there frequency effects in instructed second language acquisition (SLA)?

Frequency effects in language

Usage-based models of language have shown that for first languages, implicit learning - via mostly unconscious domain-general processes like entrenchment[1], distributional tallying of form-function mappings, schematization and categorization -from the input is crucial for processing, storing and acquisition (Behrens 2009, Goldberg 2009, Tomasello 2003).

This assumption predicts significant effects for input features like frequency distributions on all aspects of processing, storage and acquisition (Bybee 2006, Diessel 2007, Ellis 2002). The development of productivity in first language acquisition and artificial language learning seems to be specifically dependent on input features such as type variability and skewed input[2] (Boyd/Goldberg 2009, Goldberg/Casenhiser 2008, Suttle/Goldberg to appear).

However, the question whether the development of productivity in instructed SLA is bound to the same mechanisms and frequency effects is only beginning to be seriously investigated (Ellis/Ferreira-Junior 2009, Year/Gordon 2009, McDonough/Kim 2009).

Frequency effects in second language productivity

The research project contributes to clarifying the issue of frequency effects in developing productivity in instructed adult SLA by investigating the following main research questions:

[1] The process of strengthening of the representations of frequently encountered items and sequences.
[2] With a sensibly higher token frequency for one specific type of the target construction than for the others.

(i) How do second language learners attain productivity[3] with a new schematic construction[4]?
(ii) Which effects do different input features, specifically frequency distributions, have on the development of (over-) productivity with new constructions in instructed SLA?
(iii) How can developing productivity be detected, tracked, measured and quantified over time in a learner data set?

If statistically significant frequency effects on productivity can be demonstrated in instructed SLA settings under implicit *focus on form*[5] conditions, this will not only provide evidence for implicit learning mechanisms still working in SLA under certain conditions, but also open up new perspectives for instruction, namely for improved input structuring for optimal input processing (cf. Ellis 2009).

I expect different types of frequency effects in specific conditions of input enhancement[6] on the speed and depth of the development of productivity with a new schematic construction, namely effects of specifically manipulated (i) overall type frequencies[7] (high, mid, low); and (ii) type-token ratios[8] (balanced vs. skewed). At the same time, potentially negative (temporary or long-lasting) side effects in the domain of overgeneralizations/overproductivity are expected for certain conditions.
In order to corroborate these assumptions, training studies in adult second language classes are conducted. A first training study with academic learners of German as a second language (levels B.1/B.2) focuses on the acquisition of a specific participle construction, featuring five treatment conditions and a control group. The daily training sessions feature differently enhanced audio input without explicit grammar instruction over two weeks during regular class time.

The input texts are based on the analysis of the target construction in two native speaker corpora of German, focusing on (i) the availability and distinctiveness of the target construction with different verbs and (ii) natural co-and contexts of the target construction in native speaker discourse.

Qualitative data as a potential window on learning processes are gathered through daily tasks in the classroom and learner diaries/homework. Quantitatively exploitable data on learning outcomes (i.e. productivity, overproductivity) are gathered through a pretest, a posttest and two delayed posttests.

The poster presents first results from a pilot study focusing on the effects of overall type frequencies and discussing:
(i) trends and inter-group differences in developing productivity and development of/retreat from overgeneralizations; and

[3] Defined here as the increasingly accurate command of an increasing number of different types of the new construction in an increasing number of adequate contexts, be it through generalization/abstraction of higher-level schemas or through analogy to low-level patterns.
[4] „stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns“ (Goldberg 2003:221).
[5] „focus on form [...] overtly draws students‘ attention to linguistic elements as they arise incidentally in lessons whose overriding focus is on meaning and communication“ Long (1991: 45f.).
[6] Defined here in a narrow sense as specific manipulations of type/token frequency distributions, e.g. input flood.
[7] i.e. the number of different verb lemmas instantiating the construction in the input.
[8] i.e. the number of instantiations/tokens per type.

(ii) methodological issues in identifying and quantifying productivity in second language learner data, namely theoretical problems influencing data elicitation and sampling and empirical issues impacting the interpretation of elicited data and theory building, e.g. how to operationalize the concepts of entrenchment, productivity and overproductivity when evaluating second language learner data; how to deal with individually highly variable amounts of self-priming during the training period, arising from quantitatively and qualitatively varying learner productions in the tasks and learner diaries.

References

Barðdal, Jóhanna 2008. *Productivity. Evidence from Case and Argument Structure in Icelandic*. Amsterdam/Philadelphia: John Benjamins.

Behrens, Heike 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics* 47(2), 383-411.

Boyd, Jeremy K. and Goldberg, Adele E. 2009. Input effects within a constructionist framework. *Journal of Modern Language* 93(3), 418-429.

Bybee, Joan 2006. From usage to grammar: the mind‘s response to repetition. Language 82(4) 711-733.

Bybee, Joan 2008. Usage-based Grammar and Second Language Acquisition. In Robinson, Peter and Ellis, Nick C. (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 216-236. New York/London: Routledge.

Diessel, Holger 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25(2), 108-127.

Goldberg, Adele E. and Casenhiser, Devin 2008. Construction Learning and Second Language Acquisition. In Robinson, Peter and Ellis, Nick C. (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 197-215. New York/London: Routledge.

Ellis, Nick C. 2002. Frequency Effects in Language Processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition. Studies in Second Language Acquisition 24(2), 143-188.

Ellis, Nick C. 2009. Optimizing the input: Frequency and sampling in Usage-based and Form-focussed Learning. In Long, Michael H. and Doughty, Catherine J. (eds.), *Handbook of Language Teaching*, 139-158. Oxford: Blackwell.

Ellis, Nick C. and Ferreira-Junior, Fernando 2009. Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal* 93(3), 370-385.

Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences* 7(5), 219-224.

Goldberg, Adele E. 2009. The Nature of Generalization in Language. *Cognitive Linguistics* 20(1), 93-127.

Long, Michael H. 1991. Focus on form: A design feature in language teaching methodology. In Bot, Kees de, Ginsberg, Ralph B. and Kramsch, Claire (eds.), *Foreign Language Research in Cross-cultural Perspective*, 39-52. Amsterdam/ Philadelphia: John Benjamins.

McDonough, Kim and Kim, Youjin 2009. Syntactic Priming, Type Frequency, and EFL Learners‘ Production of Wh-Questions. *Modern Language Journal* 93(3), 386-398.

Suttle, Laura and Goldberg, Adele E. to appear. The Partial Productivity of Constructions as Induction. Linguistics.

Tomasello, Michael 2003. *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge et al.: Harvard University Press.

Year, Jungeun and Gordon, Peter 2009. Korean Speakers’ Acquisition of the English Ditransitive Construction: The Role of Verb Prototype, Input Distribution, and Frequency. *Modern Language Journal* 93(3), 399-417.

Thomas Mayer[1], Christian Rohrdantz[2], Frans Plank[1], Miriam Butt[1], Daniel A. Keim[2]
Department of Linguistics[1], Department of Computer Science[2] University of Konstanz
thomas.mayer@uni-konstanz.de, christian.rohrdantz@uni-konstanz.de

A Quantitative Approach to the Contrast and Stability of Sounds

The phonology of a language (like other components of grammar) undergoes change in the course of time. Languages differ as to which phonemic contrasts are made and also which changes their phonological system undergoes. How intimately related these two aspects of a phonological system are remains an open question, though usually some relationship is expected. In this paper we look at two aspects of phonological systems and language change: 1) the relative stability of phonemes as presumed indirectly from a cross-linguistic study of cognates; 2) expressions of phonemic contrast with respect to a cross-linguistic study of word forms which differ in only one sound.

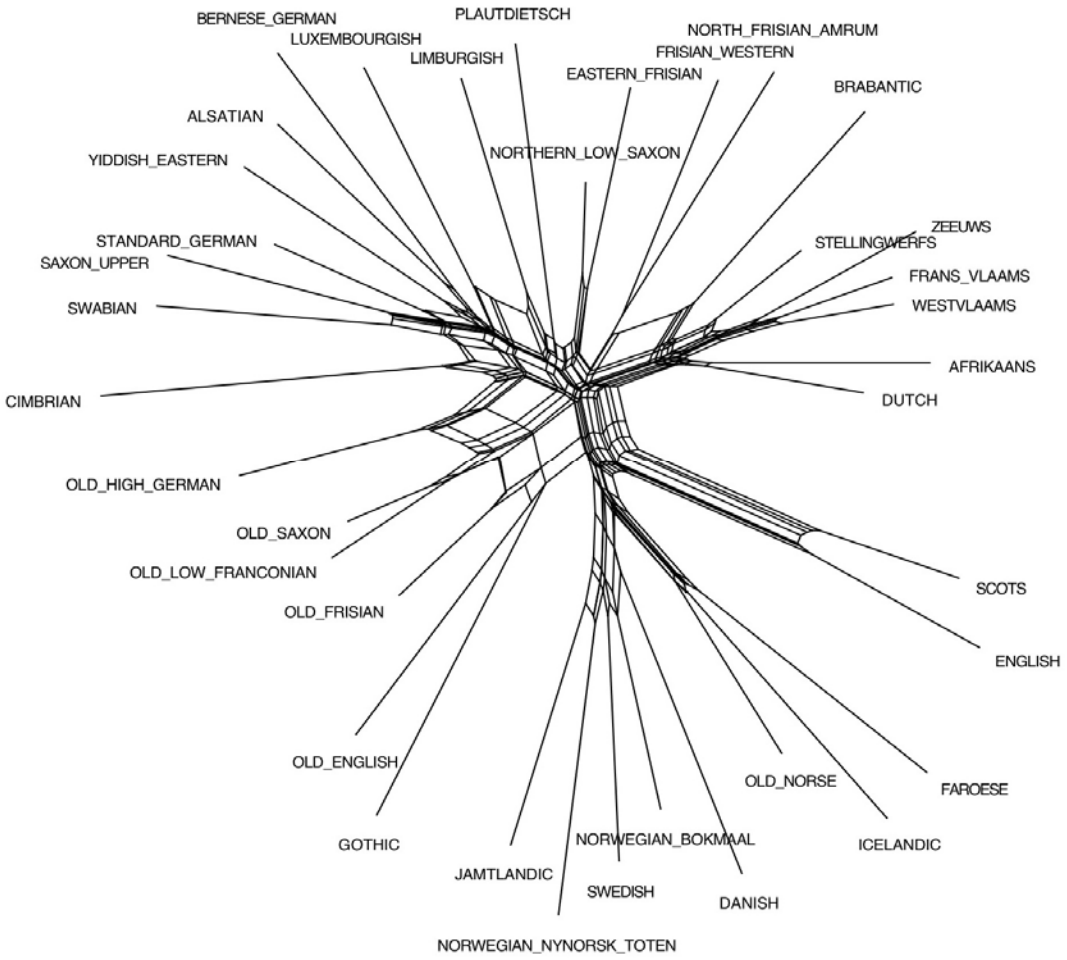
Relative stability of sounds

Ideally, we should be using diachronic data in order to directly investigate the historical stability of phonemes. However, due to the lack of suitable diachronic corpora for a cross-linguistically representative set of languages, we decided to experiment with assessing the stability of sounds indirectly by basing our work on synchronic data that is available for a wide range of languages. In particular, we decided to use the data collected as part of the ASJP (Automated Similarity Judgment Program) database (version 12, Wichmann et al. 2010), since the database includes data on a wide range of languages and in a phonetically transcribed form.

Generally, historical linguists have tacitly assumed consonants to be more reliable/stable than vowels in the search for cognates as the basis for reconstructing sound changes (Campbell 2004; Wälchli 2010). But can it be quantitatively defended that vowels are generally less stable than consonants? And is there a general stability cline in the sounds of the languages (either for individual families or universal)? In addressing these questions, we experimented with automatically comparing items in related languages. Since the vocabulary items in the Swadesh list are expected to be culturally neutral and stable over time, areal influence is kept to a minimum and diachronic conclusions are potentially justified. We further make a simplifying assumption that the same Swadesh item in related languages is a cognate. This is not true for all items (e.g., English tree and German Baum are not cognates, yet fill the corresponding slot in the Swadesh list), but across languages in our approach cases like this can be considered to be noise in the data.

Our experiments show that setting up genealogical relationships with synchronic data on Swadesh list items yields reasonably accurate results when comparing a restricted set of languages. In Figure 1, for example, an automatically created neighbor net based on the Levenshtein distance of corresponding Swadesh items groups languages in accordance with expert classifications (see also Brown et al. 2008 for similar results). So despite of the sparse data available for individual languages we assume that interesting conclusions can be drawn when comparing languages within language families.

FIGURE 1. NEIGHBOR NET OF ALL GERMANIC LANGUAGES IN THE ASJP DATABASE ON THE BASIS OF THEIR LEVENSHTEIN DISTANCE (CREATED WITH SPLITSTREE, CF. HUSON & BRYANT 2006).



In order to investigate the historical stability of sounds via automatic methods, we compared each Swadesh item in the ASJP database for all languages within a language family with its corresponding item. For each comparison we counted the substitutions that are required for each word pair with respect to its Levenshtein distance. This gives us an approximation of the sound changes that might have taken place. It is a only a rough approximation because the method assumes that one of the sounds must have been in the respective form of the proto-language from which the other sound diverted. The identified substitutions were then statistically analyzed as to their association strength. For this purpose we used the φ value (the normalized χ^2 value, see Manning & Schütze 1999) in order to be able to compare language families with differing number of pairs. The direction of the sound change cannot be determined with synchronic data, therefore all substitutions must necessarily be considered to be sound correspondences rather than changes. Table 1 shows the top and bottom consonant correspondences for the Germanic and Romance language families, respectively. It is easy to find examples for the top sound correspondences in the languages of the respective families (e.g., English stone [st@Un] vs. German Stein [Staın]).

TABLE 1. TOP AND BOTTOM SOUND CORRESPONDENCES FOR GERMANIC (LEFT) AND ROMANCE (RIGHT) LANGUAGES IN THE ASJP DATABASE. IF SYMBOLS IN THE ASJP ORTHOGRAPHY REPRESENT MORE THAN ONE SOUND, ALL CORRESPONDING IPA SYMBOLS ARE LISTED. THE SOUND CORRESPONDENCES HAVE BEEN SORTED ACCORDING TO THEIR SIGNED φ VALUES.

rank	corresponding sounds	φ value	rank	corresponding sounds	φ value
1	[S].[s]	0.4511388	1	[v].[b,β]	0.419241
2	[t].[d]	0.4202489	2	[ɯ,u].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	0.4096724
3	[t].[ts,dz]	0.3814132	3	[ɨ,ə,ə,ɜ,ʉ,ɵ,ɞ].[ɐ]	0.2465891
4	[v].[w]	0.3765362	4	[n].[ɲ]	0.2305777
5	[b,β].[p,φ]	0.3727747	5	[ʃ].[s]	0.2179642
6	[s].[z]	0.3529081	6	[L,ɭ,ʎ].[h,ɦ]	0.1868974
7	[ɯ,u].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	0.2538024	7	[k].[ʒ]	0.1449062
8	[k].[x,ɣ]	0.2249017	8	[g].[x,ɣ]	0.1354777
…	…	…	…	…	…
660	[t].[i,ɪ,y,ʏ]	-0.04844407	555	[e,ø].[.g]	-0.04514492
661	[s].[ɯ,u]	-0.04988262	556	[ɯ,u].[a,æ,ɛ,œ]	-0.049493
662	[s]. [ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.05132239	557	[i,ɪ,y,ʏ].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.056765
663	[t].[a,æ,ɛ,œ]	-0.05507378	558	[ɤ,ʌ,ɑ,ɒ,o,ɔ].[a,æ,ɛ,œ]	-0.0583884
664	[s].[ɐ]	-0.05950131	559	[ɯ,u].[ɐ]	-0.06042425
665	[t].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.06332589	560	[e,ø].[ɯ,u]	-0.06334957
666	[t].[ɐ]	-0.0693204	561	[ɤ,ʌ,ɑ,ɒ,o,ɔ].[ɐ]	-0.06435476

If we grant that synchronic comparison can inform us about historical factors, then our data indicate that the top ranked sounds should be historically less stable than ones that are lower ranked. How these values can be related to the stability of vowels vs. consonants is currently a matter of on-going investigation.

Cross-linguistic study of sound contrast

Our initial results show that an automatic analysis of cognates across languages is successfully able to identify language relatedness and provides information about which sounds are most likely to be changed. One factor in sound change is the expression and/or preservation of phonemic contrast in a language. We therefore looked at word forms which differ in only one sound across languages in the ASJP database to see whether one could automatically identify patterns among sounds based on their distribution across maximally large contexts on the word level. The substitutions in these cross-linguistic minimal pairs have been counted and statistically analyzed with the φ value (see above).

Some results are shown in Figures 2 and 3. Figure 2 shows that vowels form a group that can be differentiated very clearly from consonants. This is to be expected, since vowels should mainly be contrasting with one another. However, the results in Figure 3 are unexpected. Figure 3 focuses on just the consonant patterns and removes the main effect of the vowels. Once this main effect is removed, a clear pattern with respect to the consonants emerges. The consonants appear to fall into two major groups. This division is unexpected as we do not see it following from distinctions established so far in the phonological literature.

FIGURE 2. SOUND CORRESPONDENCES WITHIN MINIMAL PAIRS ACROSS LANGUAGES. VOWELS CAN BE CLEARLY DIFFERENTIATED FROM CONSONANTS. ROWS AND COLUMNS HAVE BEEN SORTED AUTOMATICALLY ACCORDING TO THE SIMILARITY OF THE SOUNDS. FOR THE SYMBOLS SEE TABLE 2.

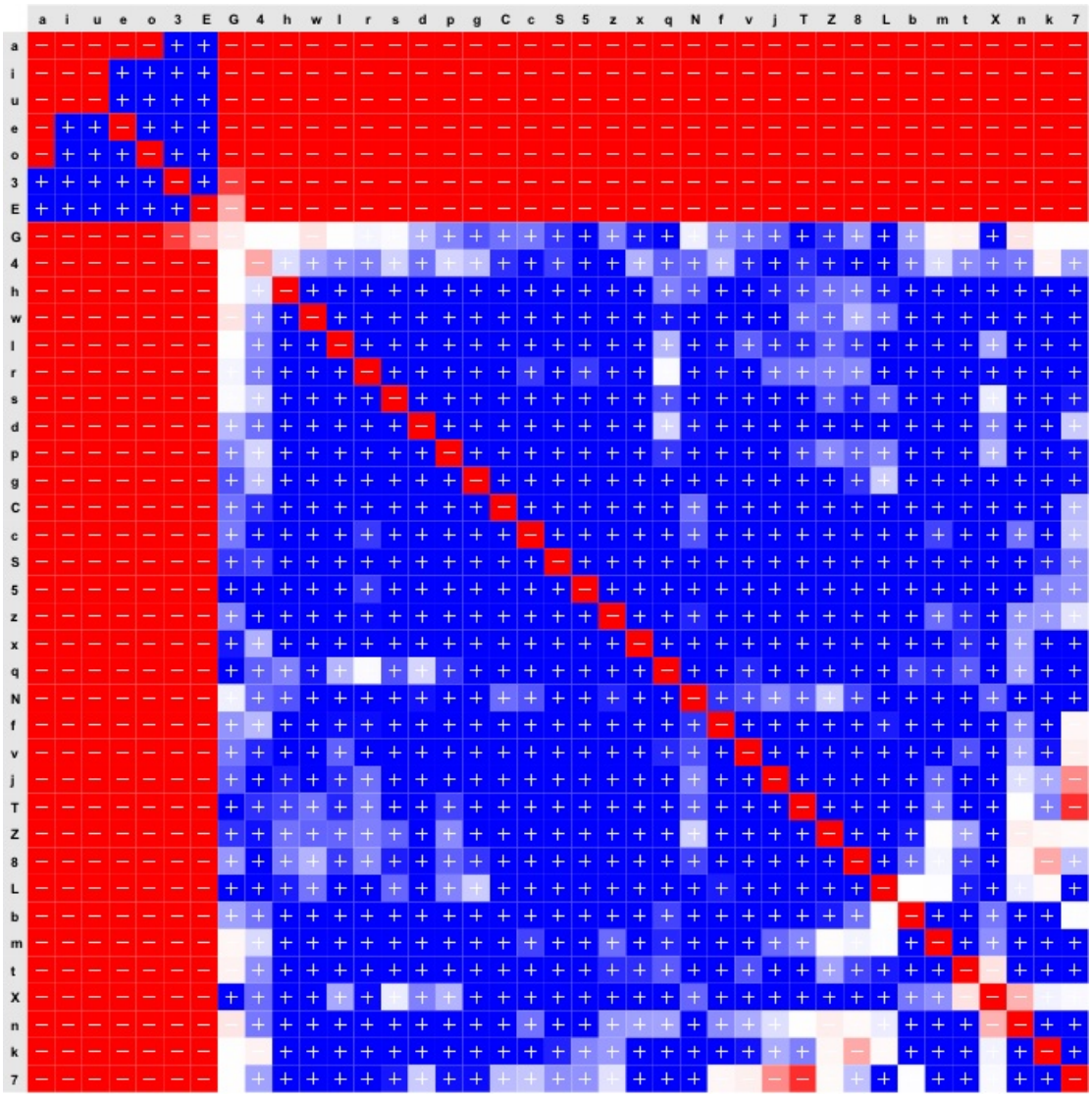
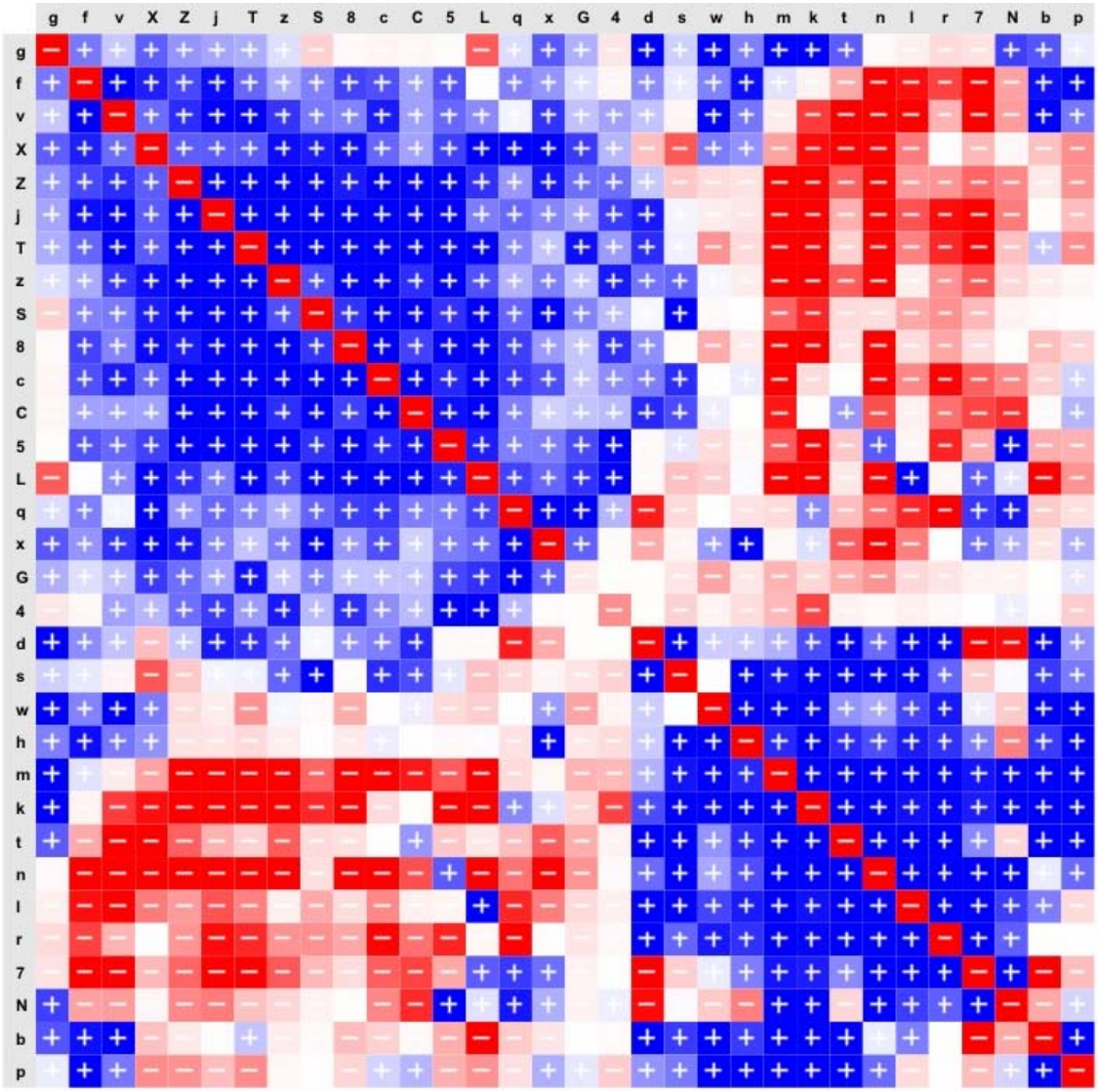


FIGURE 3. SOUND CORRESPONDENCES WITHIN MINIMAL PAIRS ACROSS LANGUAGES (CONSONANTS ONLY). THE CONSONANTS SEEM TO FALL INTO TWO MAJOR GROUPS (LEFT TOP AND BOTTOM RIGHT CORNER). ROWS AND COLUMNS HAVE BEEN SORTED AUTOMATICALLY ACCORDING TO THE SIMILARITY OF THE SOUNDS. FOR THE SYMBOLS SEE TABLE 2.



The results presented here are part of a larger on-going effort to introduce methods from visual analytics (Thomas & Cook 2005; Keim et al. 2008) into quantitative linguistic analyses. In this paper, we show that the automatic examination of sound patterns across languages can be used to further our understanding of sound change (phoneme stability) and phonemic distinctions. In particular, the results in Figure 3 have brought to light a new linguistic pattern which can now be explored further in terms of a fruitful interaction between theoretical and quantitative approaches.

TABLE 2. ASJP ORTHOGRAPHY (CF. BROWN ET AL. 2008).

ASJP symbol	IPA symbol(s)	ASJP symbol	IPA symbol(s)
i	[i,ɪ,y,ʏ]	S	[ʃ]
e	[e,ø]	Z	[ʒ]
E	[a,æ,ɛ,œ]	C	[ʧ]
3	[ɨ,ɘ,ə,ɜ,ʉ,ɵ,ɞ]	j	[ʤ]
a	[ɐ]	T	[c,ʃ]
u	[ɯ,u]	5	[ɲ]
o	[ɤ,ʌ,ɑ,ɒ,o,ɔ]	k	[k]
p	[p, φ]	g	[g]
b	[b, β]	x	[x,ɣ]
m	[m]	N	[ŋ]
f	[f]	q	[q]
v	[v]	G	[G]
8	[θ,ð]	X	[χ,ʁ,ħ,ʕ]
4	[ɳ]	7	[ʔ]
t	[t]	h	[h,ɦ]
d	[d]	l	[l]
s	[s]	L	[L,ɭ,ʎ]
z	[z]	w	[w]
c	[ts,dz]	y	[j]
n	[n]	r	[r,R,etc.]

References

Brown, Cecil H., Holman, Eric W., Wichmann, Søren and Velupillai, Viveka. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF Language Typology and Universals* 61, 285-308.

Campbell, Lyle. 2004. *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Huson, Daniel H. and Bryant, David 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23, 254-267.

Keim, Daniel A., Mansmann, Florian, Schneidewind, Joern, Thomas, Jim and Ziegler, Hartmut. 2008. Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, Lecture Notes in Computer Science*, 76-91. Springer.

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

Thomas, James J. and Cook, Kristin A. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.

Wälchli, Bernhard. 2010. The consonant template in synchrony and diachrony. *Baltic Linguistics* 1, 137-166.

Wichmann, Søren, Müller, André, Velupillai, Viveka, Brown, Cecil H., Holman, Eric W., Brown, Pamela, Urban, Matthias, Sauppe, Sebastian, Belyaev, Oleg, Molochieva, Zarina, Wett, Annkathrin, Bakker, Dik, List, Johann-Mattis, Egorov, Dmitry, Mailhammer, Robert and Geyer, Helen. 2010. The ASJP Database (version 12). URL: http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm.

Paul Metzner
University of Potsdam
paul.metzner@uni-potsdam.de

Can the asymmetric effect of implicit causality be explained through usage patterns?

Introduction

To understand complex utterances, people have to build a coherent mental representation of the described events. An essential part of such representations are causal relations that are typically constructed with the help of overt markers like conjunctions and covert properties like *implicit causality* (Garvey & Caramazza 1974; Garvey, Caramazza, & Yates 1975), which designates some verbs' propensity to impute the cause of an event to either subject or object. According to the respective causality bias, people will continue *Oskar annoyed Emma, because…* with *he* and *Oskar noticed Emma, because…* with *she*. This pattern is very consistent and has been observed in numerous studies (e.g., Au 1986; Brown & Fish 1983). Furthermore, causality-biased verbs like *annoy* and *notice* have been shown to influence online sentence comprehension. Bias-incongruent pronouns are harder to integrate, resulting in longer reaction and reading times (Garnham, Traxler, Oakhill, & Gernsbacher 1996; Stewart, Pickering, & Sanford 2000), more regressions and longer fixations (Featherstone & Sturt 2010; Koornneef & Van Berkum 2006; Metzner 2010) plus electrophysiological signs of syntactic integration problems (Van Berkum, Koornneef, Otten, & Nieuwland 2007). Consider sentences 1a (congruent) and 1b (incongruent) for clarification:

(1) a. *Oskar noticed Emma, because she waved at him.*
b. *Oskar noticed Emma, because he looked up.*

Greene and McKoon (1995) and Long and De Ley (2000) report an effect of congruence only for NP2-biased verbs like *notice*. Whereas Greene and McKoon ascribe this to the recency of NP2, Long and De Ley assume that it is due to differences in everyday language. They hypothesize that NP2-biased verbs yield more useful predictions for upcoming anaphora and support this assumption with a corpus study. They find sentences with NP2-biased verbs to be congruent with the verb's implicit causality more often than sentences with NP1-biased verbs like *annoy*. However, Long and De Ley's (2000) corpus is a comparably small and regionally very constrained newspaper corpus that does not reflect verbal communication and thus might not enable sophisticated generalizations about language comprehension.

Method and Results

To validate Long and De Ley's (2000) results, sample sentences for strongly biased verbs (determined in a preliminary study and taken from Long & De Ley 2000, respectively) were randomly chosen from a corpus of written German and a corpus of spoken English. It was recorded whether a reference to subject or object of the main clause was present, and, given that one had occurred, if it was congruent with the verb's implicit causality bias. The investigation of *Tübingen's Partially Parsed Corpus of Written German* (TüPP-D/Z: Eberhard Karls Universität Tübingen 1986-1999) was supposed to substantiate Long and De Ley's findings with results from another

language. Moreover, the analysis of the spoken component of *BYU-BNC: The British National Corpus* (Davies 2004–) could have strengthened the usage-based account with results from a different modality. The patterns did, however, not resemble Long and De Ley's. First, NP1- and NP2-biased verbs were equally good predictors for upcoming anaphora in the German corpus ($\chi 2(1) = 0.003$*;* $p > .05$). Second, the data from the BNC revealed the exact opposite pattern to Long and De Ley (2000), with NP1-biased verbs being more reliable predictors than NP2-biased verbs ($\chi 2(1) = 7.71$*;* $p < .01$). Consequently, our results do not support a usage-based explanation of the diverging magnitudes of NP1- and NP2-based effects.

Discussion

The picture remains diffuse. Greene and McKoon (1995), Long and De Ley (2000), and Featherstone and Sturt (2010) found stronger effects for NP2-biased verbs in probe-recognition tasks and eye-tracking experiments in English. Koornneef and Van Berkum (2006) and Metzner (2010), however, report the opposite pattern in eye-tracking experiments on Dutch and German. Also, NP1-biased verbs were better predictors for future reference in a corpus of spoken English and NP2-biased verbs were better predictors in a corpus of written English, but there was no difference in a corpus of written German. If the asymmetric effects of NP1- and NP2-biased verbs were based on usage-based heuristics as proposed by Long and De Ley (2000), we would have expected to see a corresponding pattern in the German corpus, which we did not.

There are theoretical reasons why NP1-biased verbs would elicit a stronger effect. Incongruence with a bias towards NP1 always coincides with a violation of the first-mention advantage (Gernsbacher & Hargreaves 1988) and almost always with the assumptions of the parallel-function hypothesis (Grober, Beardsley, & Caramazza 1978), which posits that a discourse entity will preferably take the same grammatical role in both main and subordinate clause. This is not the case with verbs that are biased towards NP2.

An account for the difference between Dutch/German and English and between written and spoken English might highlight cross-linguistic and cross-modal variation. However, apart from the asymmetries, all experiments have brought roughly the same results for different languages. Also, we would rather expect NP2-biased verbs to play a more important role in verbal communication, where parsimony is instrumental. The most convincing explanation for Long and De Ley's (2000) and similar results thus appears to be recency, because their probe-recognition task is by design prone to capture memory effects. However, Featherstone and Sturt (2010) also report stronger effects for NP2-biased verbs in an eye-tracking experiment, so there must be more to it than just recency.

Obviously, different experimental sentences and methods were used across experiments.

Although they have in common that their verbs are biased towards NP1 or NP2, they differ in terms of overall meaning, bias strength, language, participants, and method.

Featherstone and Sturt (2010, p. 10) argue that there might not be a systematic difference between NP1- and NP2-biased verbs, but that the asymmetric effects might merely reflect the above-mentioned variation in the material and the different methods. Given the seemingly random pattern of results, we are very inclined to follow this reasoning.

Conclusions

The present study casts doubt on a usage-based explanation for implicit causality, but this doubt is mainly grounded in a null result from the German corpus. To confidently reject such an explanation, one would have to find diverging evidence from experimental and corpus data (e.g., NP2-biased verbs to be better predictors in the German corpus). Moreover, the present results come with a big caveat. They are based on very small samples, so an investigation of the topic with a larger corpus is certainly indicated. Also, sentences in the German preterite tense were extracted to grant comparability with Long and De Ley's (2000) results. Since this form is becoming more and more unusual in everyday German, another study with sentences in perfect or present tense appears like another useful extension.

References

Au, T. K.-F. 1986. A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language* 25(1), 104-122.

Brown, R. and Fish, D. 1983. The psychological causality implicit in language. *Cognition* 14(3), 237-273.

Davies, M. 2004–. BYU-BNC: The British National Corpus.

Eberhard Karls Universität Tübingen. 1986-1999. *Tübingen's Partially Parsed Corpus of Written German – TüPP-D/Z*.

Featherstone, C. R. and Sturt, P. 2010. Because there was a cause for concern: an investigation into a word-specific prediction account of the implicit-causality effect. *The Quarterly Journal of Experimental Psychology* 63(1), 3-15.

Garnham, A., Traxler, M. J., Oakhill, J. and Gernsbacher, M. 1996. The locus of implicit causality effects in comprehension. *Journal of Memory and Language* 35(4), 517-543.

Garvey, C. and Caramazza, A. 1974. Implicit causality in verbs. *Linguistic Inquiry* 5(3), 459-464.

Garvey, C., Caramazza, A. and Yates, J. 1975. Factors influencing assignment of pronoun antecedents. *Cognition* 3(3), 227-243.

Gernsbacher, M. and Hargreaves, D. 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* 27(6), 699-717.

Greene, S. B. and McKoon, G. 1995. Telling something we can't know: Experimental approaches to verbs exhibiting implicit causality. *Psychological Science* 6(5), 262-270.

Grober, E., Beardsley, W. and Caramazza, A. 1978. Parallel function strategy in pronoun assignment.
Cognition 6(2), 117-33.

Koornneef, A. W. and Van Berkum, J. J. A. 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language* 54(4), 445-465.

Long, D. and De Ley, L. 2000. Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language* 42(4), 545-570.

Metzner, P. 2010. Working memory and implicit causality. An eye-tracking study on individual differences in language comprehension. Unpublished master's thesis, Humboldt-Universität zu Berlin.

Stewart, A., Pickering, M. J. and Sanford, A. J. 2000. The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language* 42(3), 423-443.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M. and Nieuwland, M. S. 2007. Establishing reference in language comprehension: an electrophysiological perspective. *Brain Research* 1146, 158-171.

Benjamin J. Molineaux
Faculty of Linguistics, Philology and Phonetics, Oxford University
benjamin.molineaux@ling-phil.ox.ac.uk

The acquisition of English particle placement: A case study in developing productivity

Expanding upon work by Diessel & Tomasello (2005, hereafter D&T) and Gries (to appear),[1] this study explores particle positioning in child language as a gradual cognitive-pragmatic development based on statistical distribution of input. The phenomenon of particle position alternation is peculiar insofar as it defines a category of combinations which show syntactic variation without concomitant semantic variation.

Place alternation in verb-particle constructions

a. Tip over$_{[\text{PARTICLE}]}$ the cup. (continuous)
b. Tip the cup over$_{[\text{PARTICLE}]}$. (discontinuous)

Such a category is distinct from verbs followed by prepositional phrases, which lack alternation, as well as phrasal stress on the preposition (see Dehé 2002 for a review).

Lack of place alternation in verbs followed by prepositional phrases

a. Climb over$_{[\text{PREPOSITION}]}$ the wall.
b. ?Climb the wall over$_{[\text{PREPOSITION}]}$.

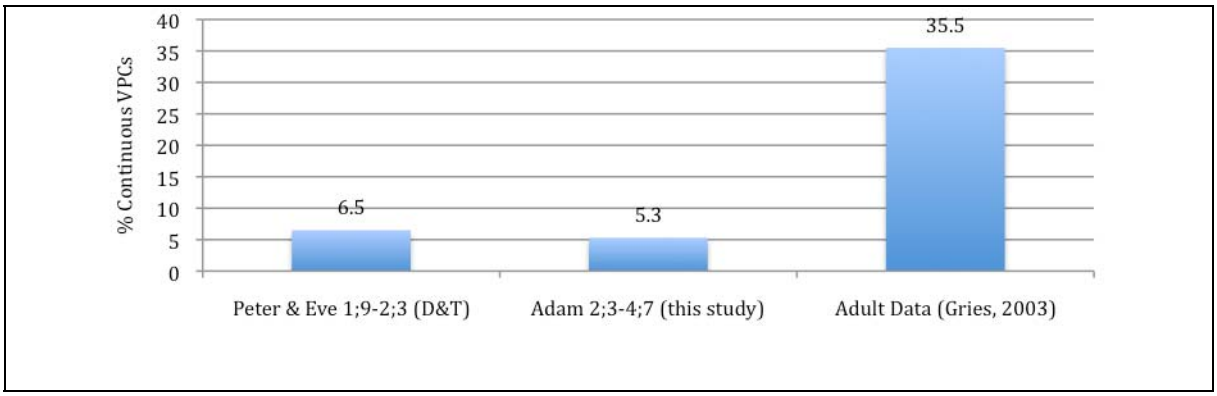
Overall, D&T find that early child language has no default particle order (as claimed by Hyams et al. 1993; Bennis et al. 1995 and more recently by Snyder 2007), and that some of the contextual and processing factors determining particle placement distribution (Gries 2003) are already present in children's earliest multi-word utterances. This study goes further, attempting to establish the course of development for adult-like particle ordering. For these purposes, it focuses, not on the earliest stages of syntax, but on the developmental period immediately following that studied by D&T (1;6 – 2;3), that is, the period between 2;3 and 4;10, as represented by Adam from the Brown Corpus (1973).

In contrast with Gries' study (to appear), which analyses multiple semantic, syntactic and phonological factors across the entirety of the developmental corpus data for three children, my goal is to observe the corpus-internal development of particle distribution within the 31 months of the available data. A view is provided into how a single child gradually develops a principled use of particle order, as a function of contextual factors such as object type and meaning. More suggestively, based on a sampling of data from Adam's mother, arguments are provided so as to consider child particle ordering a fundamentally lexically-based process (an idea also examined in Gries, to appear), which gradually transitions into a principled, probabilistic system.

The initial contrast of D&T's data (based on Eve (Brown 1973) and Peter (Bloom et al. 1974-5)) with my own, is unilluminating as regards particle position, since it shows a decrease in the continuous order when viewed from the perspective of adult distributions (data from Gries 2003).

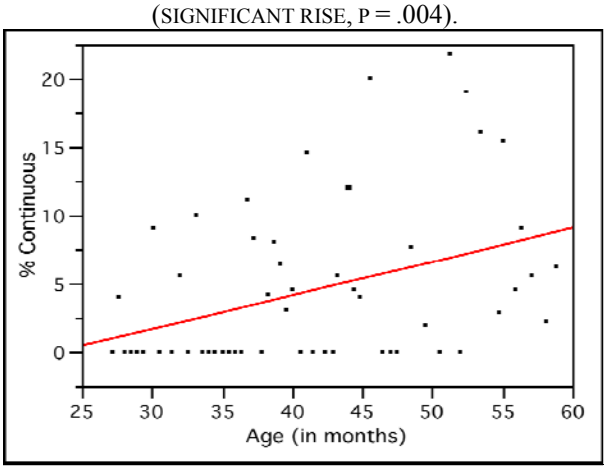
[1] The author would like to thank an anonymous reviewer for pointing out the imminent publication of Gries (to appear), which substantially changes the landscape of English particle-verb acquisition studies.

FIGURE 1: PERCENTAGE OF CONTINUOUS VPCS BY STUDY.



A closer look at the internal development of Adam's data alone, provides a very different picture.

FIGURE 2. BIVARIATE FIT FOR PERCENTAGE OF CONTINUOUS VPCS AGAINST ADAM'S AGE (SIGNIFICANT RISE, P = .004).



Internally, the corpus data for Adam shows a gradual growth in the continuous order, which predicts that, if sustained, adult-like ordering may not be acquired until after the age of twelve. Furthermore, in line with D&T, we may say that there is no one moment in the data that attests to the setting of a parameter licensing the adult distributions.

As in the case of D&T's and Gries' (2003, to appear), our study finds a statistical correlation (here a simple, monofactorial chi-square distribution) between VPC order and its meaning as well as its object-type. In other words, we find a significant relation ($\chi^2 = 52.224$; $p < .0001$) between the ordering of the particle and the object NP's being a pronoun (*I wrap it around*, 4;1) or a noun (*gobble up the food*, 3;8). Such a relation also holds for particle order and the meaning of the VPC ($\chi^2 = 14.829$; $p < .0001$), which may be spatial (i.e. directional, *pick dirt up,* 2;4) or non-spatial (i.e. idiomatic, *made it all up*, 2;5, or telic aktionsart, *eat you up*, 2;7). Again, here I add to D&T and Gries' (forthcoming) analyses by focusing on the internal development of these factors within the corpus. In separate linear bivariate fit analyses, the proportion of nominal objects is shown to decrease over time, as does the percentage of spatial VPCs.

FIGURE 3. BIVARIATE FIT FOR PERCENTAGE OF NOMINAL NPS AGAINST ADAM'S AGE (SIGNIFICANT DROP, P=.008).

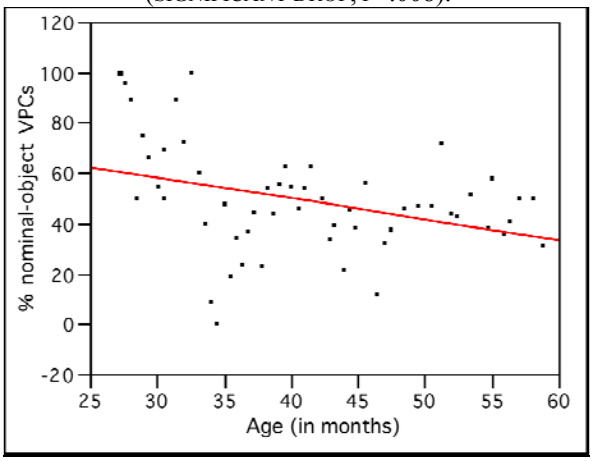
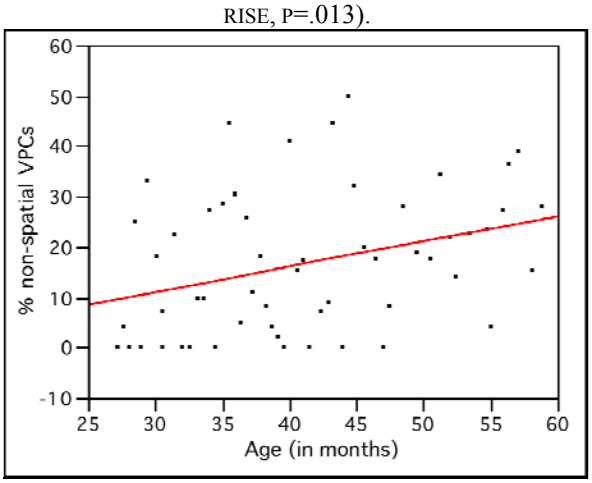


FIGURE 4. BIVARIATE FIT FOR THE PERCENTAGE OF NON-SPATIAL NPS AGAINST ADAM'S AGE (SIGNIFICANT RISE, P=.013).



This is precisely what we would expect if we consider that adult-like VPC distribution is the result of the acquisition of principles and processing strategies (such as those outlined by Gries 2003), and the gradual growth in their frequency of use by children. Two of the factors that appear to contribute to adult distributions are the decrease in nominal objects – which correlates to the shrinking percentage of discontinuous VPCs – and the growth of non-spatial forms – which correlates to the growth of the continuous VPC.

I argue that my analysis is in line with the 'Cognitive Commitment' proposed by Lakoff (1991) insofar as the use of pronominal forms, as well as non-spatial semantics imply a greater development of abstraction, which ultimately shows the mapping of more complex cognition to particular linguistic structures, which in turn grow in frequency. The development of productive particle ordering in phrasal verbs is another case of grammatical developments mirroring the growth of cognitive structure, both linguistic and non-linguistic (cf. Bowerman 2007; Behrens 2001).

Finally, the study examines whether particular particle-verb combinations are more or less likely to surface in the continuous or discontinuous order, finding that, of the 213 VPC types in the corpus, 19 appeared (in a chi-square test) as significantly different in terms of their overall ordering ($p < .05$). I argue that these items are strongly

associated to a particular ordering, which is not a default, but an item-based generalization. Further evidence for this comes from the close match found between the child-data and the particle ordering used by Adam's mother, in the same corpus.

TABLE 1. COMPARISON OF IRREGULARLY AND REGULARLY DISTRIBUTED ITEMS IN THE ADAM CORPUS.

Sample Irregularly Distributed Items					Sample Regularly Distributed Items				
	Adam		Mother			Adam		Mother	
VPC item	Cont	Disc	Cont	Disc	VPC item	Cont	Disc	Cont	Disc
Run over	11 (0.58)	0 (10.48)	4	1	Turn on	1 (0.99)	18 (18.01)	0	8
Find out	3 (0.16)	0 (2.84)	2	0	Take off	3 (5.65)	105 (102.35)	0	47
Chop down	3 (0.16)	0 (2.84)	0	0	Take out	2 (3.77)	70 (68.23)	0	49
Cut out	7 (0.94)	11 (17.05)	1	0	Turn off	1 (0.63)	11 (11.37)	0	4
Put in	0 (5.02)	96 (90.97)	2	102	Bring down	0 (0.05)	1 (0.95)	0	1

*figures in parenthesis correspond to expected (H_0) values for the item in its particular order

Although the number of attested data for each VPC type is limited, I suggest that Adam's data represents an intermediate stage between item-dependent and principle-based particle placement. That is, based on the predominant distribution of each combination in the ambience language, particle placement may be determined for some (or all) VPCs in children's early language, only to gradually be replaced by the contextual-pragmatic factors used by adults.

The data and analyses of VPC ordering provide new evidence for a usage-based perspective on language acquisition (Tomasello 2003; Diessel 2004). Particle alternation in child language appears to behave neither randomly nor on the basis of a single set of pre-established, monolithic principles, but is rather the result of interaction between the child's observation of adult frequency, and the gradual systematization and abstraction of pragmatic, cognitive, and formal principles governing particle placement.

References

Bennis, Hans, den Dikken, Marcel, Jordens, Peter, Powers, Susan and Weissenborn, Jürgen 1995. Picking up Particles. In *Proceedings of the 19th Annual Boston University Conference on Language Development,* Malaughlin, Dawn and McEwan, Susan (eds.), 70-81. Somerville: Cascadilla Press.

Behrens, Heike 2001. Cognitive-conceptual development and the acquisition of grammatical morphemes: the development of time concepts and verb-tense. In *Language Acquisition and Conceptual Development,* Bowerman, Melissa and Levinson, Stephen. (eds.), 450-474. Cambridge: Cambridge University Press.

Bloom, Lois, Hood, Lois and Lightbown, Patsy 1974. Imitation in language development: If, when and why. *Cognitive Psychology* 6(3), 380-420.

Bloom, Lois, Lightbown, Patsy and Hood, Lois 1975. Structure and variation in child language. *Monographs of the Society for Research in Child Development*, 40(2), [Serial No. 160].

Bolinger, Dwight 1971. *The Phrasal Verb in English*. Cambridge, MA: Harvard University Press.

Brinton, Laurel 1985. Verb Particles in English: Aspect or Aktionsart? *Studia Linguistica*, 39(2), 157-168.

Brown, Roger 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Bowerman, Melissa 2007. Containment, support, and beyond. In *The Categorization of Spatial Entities in Language and Cognition.* Aurnague, Michael Hickmann, Maya and Vieu, Laure (eds.), 177-203. Amsterdam: John Benjamins.
Dehé, Nicole 2002. *Particle Verbs in English: Syntax, Information Structure and Intonation*. Amsterdam: John Benjamins.
Diessel, Holger 2004. *The Acquisition of Complex Sentences*, Cambridge: Cambridge University Press.
Diessel, Holger and Tomasello, Michael 2005. Particle placement in early child language: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 1(1), 89-112.
Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum.
Gries, Stefan Th. to appear. Acquiring particle placement in English: A corpus-based perspective. In *Morphosyntactic alternations in English: functional and cognitive perspectives*, Pilar Guerrero Medina (ed.), 235-263. London & Oakville, CT: Equinox.
Hyams, Nina, Schaeffer, Jeannette and Johnson, Kyle 1993. *On the Acquisition of Verb Particle Constructions*. Unpublished manuscript. University of California at Los Angeles and University of Massachusetts, Amherst.
Lakoff, George 1991. Cognitive versus Generative Linguistics: How Commitments Influence Results. *Language and Communication* 11(1-2), 53-62.
MacWhinney, Brian 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
Snyder, William 2007. *Child Language: The parametric approach*. Oxford: Oxford University Press.
Tomasello, Michael 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA; London: Harvard University Press.

Hagen Peukert
University of Hamburg
hagen.peukert@uni-hamburg.de

Proving poverties of the stimulus wrong: A computational and corpus-based case in point

The Argument from the Poverty of the Stimulus (APS) still dominates a substantial part of the research in theoretical linguistics. While in generative approaches, the APS provides the foundation of the innateness hypothesis; in other paradigms (usage-based, interactionist, cognitivist) the argument is central because it is argued against it.

Poverty of the stimulus arguments are indeed theoretical because they do little to empirically back their premises. Following Pullum & Scholz (2002) the logic of the argument has two premises. First, language is either genetically encoded or learned data-driven. The second premise, however, is meant to be empirical, that is, infants learn what is not given in the input. Since, by definition, data driven means that everything can be learned from the input, the empirical premise excludes data driven strategies as an option. Now, only one alternative remains in the initial premise. Thus one can deduce modus tollens the genetically encoded strategy as the correct alternative via exclusion. Although the conclusion is correct, the result can be questioned because the premises are far from plausible nowadays. Accepting that the two alternatives postulated in the first premise are somewhat sufficient, the empirical premise seems more problematic. Also known as lack of evidence, this premise is a simple assertion without a systematic and comprehensive proof. The assumption was certainly still acceptable at the time of its conception. Today, however, we have collected specific and representative corpora, developed a body of sophisticated methodologies for quantitative analyses, and used the computational power to carry out complex

calculations on several variables at a time. This development enables us to re-evaluate the argument from the poverty of the stimulus.

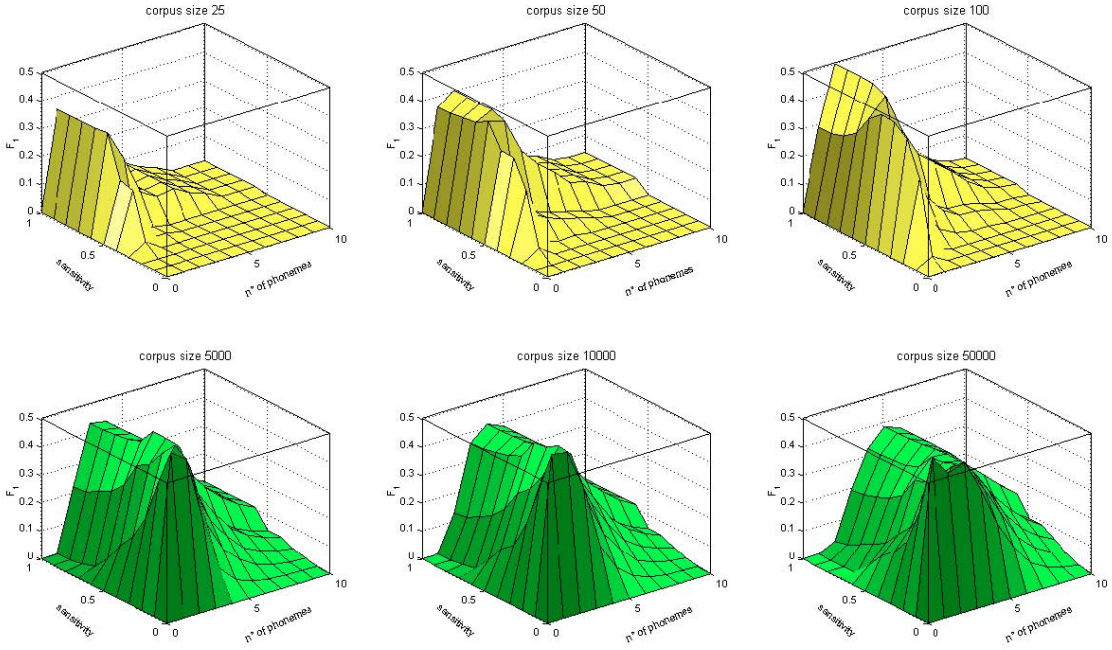
Even though the discussion on the APS centered around the syntactic analysis of phrase structures within the linguistic community (Clark 2003:18), all of the three other cognitive capabilities originally identified by Chomsky (Chomsky 1977:152; see also Dittmann 2006:74) – segmenting the speech stream, identifying word categories and syntactic categories – still deserve equal attention. Up to now, word segmentation belongs to the set of phenomena in language acquisition that cannot be solved since the input is impoverished. As a case in point, for a representative corpus of English child-directed speech, we will show that the speech stream contains enough information, although hidden, so that words and phrases can be reliably located. We model a new algorithm for word segmentation that only includes general cognitive abilities of six to eight months-olds. The model will not hide any language specific information; it is data-driven, bottom-up, usage-based. The guiding question is how do six to eight month-olds segment a stream of sounds into words? Our suggestion is to build a model around the specific distribution of sounds of a language, that is, transitional probabilities. The model at hand thus bridges Phonology and Semantics in the acquisition process.

The constraints defined in this model are all based on general cognitive abilities. First, the most important constraint is the inclusion of transitional probabilities in the model. Second, it is important to pin down the exact form of the unit of perception. From this unit, calculations of transitional probabilities will be accomplished (Saffran et al. 1996). Here the phoneme seemed best suited. Using phonemes allows for a more elaborated inquiry by controlling both possibilities: syllables as well as all other existing combinations of phonemes. Third, it is clear that babies have to memorize words for a longer period of time (Jusczyk & Aslin 1995; Jusczyk & Hohne 1997), so that, fourth, the most frequent ones (Shi et al. 2006; Jusczyk et al. 1994) can be mapped top-down (Bortfeld et al. 2005) onto an unknown speech input.

The last implementation of the model to be presented takes representative samples of a controlled size from CHILDES (MacWhinney 1995), converts them into an IPA-format, deletes all stress information and white spaces, calculates the transitional probabilities for each combination of phoneme chains ranging from 1 to 10 phonemes (n-gram model) and marks white spaces when a threshold value is reached. The threshold value is a variable manipulating the sensitivity of the child and is defined between 0 and 1. Thus, a second loop runs through ten values of a threshold limit and outputs its maximum for each phoneme chain combination.

Once the value at the optimum of both variables (partial derivative of length of phoneme chain and sensitivity of the child), is calculated, the most frequent items (differing between 5 and 30 ‘words’) are selected and saved in a list. Eventually the next corpus is input and processed as described above.

FIGURE 1. SEGMENTATION PERFORMANCE (F1) FOR SPECIFIED PARAMETER COMBINATIONS (LENGTH OF PHONEME CHAIN AND CHILD'S SENSITIVITY) AT GIVEN CORPUS SIZES.



The simulation allows measuring the exact size of the corpus that is necessary to make a maximum of correct segmentations as a function of the child's sensitivity and a corresponding unit of perception (Figure 1). The most important discovery of the simulation is an important property of the segmented corpus. It is indeed true that only about one third of all segmented words are correct (Brent 1999) and that this number cannot possibly account for a starting point in the segmentation process. At first sight it seems convincing because the child cannot know which words are correct segmentations (Gambel/Yang 2005). Searching the corpus for more detail, it becomes clear that the wrongly segmented corpora all encode some more important information, which is not at all obvious looking at it superficially. The most frequent segmentations of the wrongly segmented corpus happen to be lexical items.

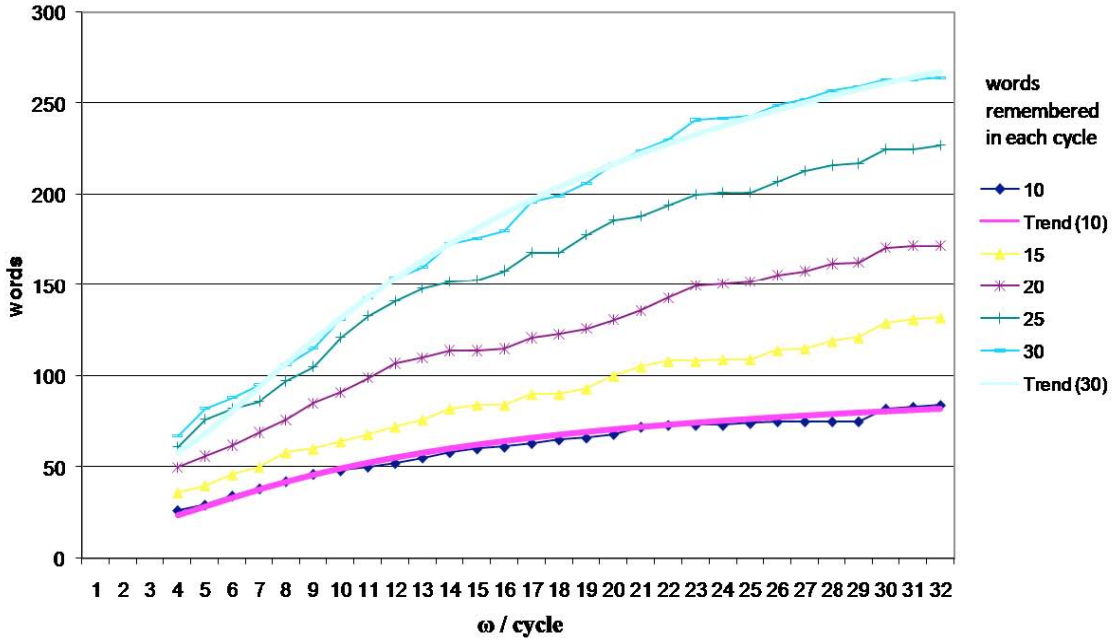
Only that finding allows solving the segmentation problem since out of each representative corpus a certain number of words can now be extracted reliably (Table 1). This number is described by a function and encodes the number of words an infant should be able to memorize (Figure 2).

$$f(\omega) = \frac{12{,}51\xi - 35{,}1}{e^{\frac{\xi}{5\omega} + \frac{8}{\omega}}} + \frac{31}{20}\xi + 1{,}5$$

(for $2 \leq \omega \leq \infty$, $10 \leq \zeta \leq 30$, whereas ζ is the function specifying the words to be memorized and ω is the number of corpora or cycles passed without regress to a lexicon)

In addition, the set can serve as 'learning material' for language specific rules (prosody, allophonic variation, phonotactics).

FIGURE 2. GROWTH OF WORDS PER CYCLE (ACTUAL AND IDEAL).



For arguing against the argument from the poverty of the stimulus, it is sufficient to show that the information is actually existent in the input since the APS is itself theoretical and does not provide empirical evidence. The model to be presented goes a step further and shows that infants of the respective age possess the necessary cognitive abilities as well. The next step would be an experimental setting confirming the interrelation and order of the suggested processes.

TABLE 1. PRELEXICAL DEVELOPMENT FOR $\omega = 6$ CYCLES AND $\zeta = 10$ ITEMS REMEMBERED DURING EACH CYCLE.

Rank	$\omega = 1$		$\omega = 2$		$\omega = 3$		$\omega = 4$		$\omega = 5$		$\omega = 6$	
1	ðət	20	ðət	34	ju	46	ðət	58	ðət	73	ðət	93
2	wʌt	19	ðer	32	ðət	46	ju	57	ju	71	ju	89
3	jə	19	mɔrgʌn	24	ðer	45	ðer	45	jɔr	47	ənd	56
4	lɪtʌl	14	ju	24	lets	31	ɒrju	37	wʌt	46	ðəts	48
5	ju	13	wʌt	19	ənd	26	wʌt	33	ðer	45	jɔr	47
6	ðer	13	jə	19	ɒrju	26	lets	31	ənd	44	wʌt	46
7	ɒrju	13	wer	14	hɪr	25	jɔr	28	hɪr	37	ðer	45
8	ənd	12	lɪtʌl	14	mɔrgʌn	24	ənd	26	ɒrju	37	mɔrgʌn	44
9	duŋ	12	ðets	13	wʌt	19	wer	25	lets	31	hɪr	37
10	ðerjugəʊ	12	ɒrju	13	jə	19	hɪr	25	ðəts	31	ɒrju	37
11			ənd	12	əʊkeɪ	17	mɔrgʌn	24	wer	25	lets	31
12			duŋ	12	wer	14	həv	23	mɔrgʌn	24	wer	25
13			ðerjugəʊ	12	lɪtʌl	14	jə	19	həv	23	ŋ	24
14			mɒmi	11	kən	14	əʊkeɪ	17	jə	19	həv	23
15			lets	11	wi	13	red	16	əʊkeɪ	17	jə	19
16			jɔr	11	ðəts	13	lɪtʌl	14	red	16	əʊkeɪ	17
17			hɪr	11	ŋ	12	kən	14	lɪtʌl	14	red	16
18					həv	12	wi	13	k?n	14	lɪtʌl	14
19					duŋ	12	tu	13	wi	13	kən	14
20					ðerjugəʊ	12	ðəts	13	tu	13	wi	13
21					mɒmi	11	ŋ	12	kʌdlitedi	13	tu	13
22					jɔr	11	duŋ	12	lʊk	12	prɪti	13
23							ðerjugəʊ	12	kʌm	12	kʌdlitedi	13
24							mɒmi	11	ŋ	12	wɔtɜ	12
25							ðʌ	11	hənd	12	ʌp	12
26									duŋ	12	lʊk	12
27									ðerjugəʊ	12	kəm	12
28									mɒmi	11	ɪts	12
29									ðʌ	11	hənd	12
30											duŋ	12
31											ðerjugəʊ	12
32											ɒr	12
33											mɒmi	11
34											ðʌ	11

References

Bortfeld, Heather, Morgan, James L. und Golinkoff, Roberta Michnick 2005. Mommy and me. *Psychological Science* 16(4), 298-304.

Brent, Michael R. 1999. Speech segmentation and word discovery. In *Trends in Cognitive Sciences* 3(8), 294-301.

Chomsky, Noam. 1977. *Reflections on Language*. New York: Pantheon Books.

Chomsky, Noam 1988. *Language and problems of knowledge*. Cambridge: MIT Press.

Clark, Eve V. 2003. *First language acquisition*. Cambridge: Cambridge University Press.

Dittmann, Jürgen 2006. *Der Spracherwerb des Kindes*. München: Beck.

Gambell, Timothy and Yang, Charles D. 2005. *Mechanisms and constraints in word segmentation*. ms, Yale University.

Jusczyk, Peter W. and Aslin, Richard N. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* 29(1), 1-23.

Jusczyk, Peter W. and Hohne, Elizabeth A. 1997. Infants' memory for spoken words. *Science* 277, 1984-1986.

Jusczyk, Peter W., Luce, Paul A. und Luce, Jan Charles 1994. Infants sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*. 33(5), 630-645.

MacWhinney, Brian 1995. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale: Lawrence Erlbaum.

Pullum, Geoffrey K. and Scholz, Barbara C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1), 9-50.

Saffran, Jenny R., Aslin, Richard N. and Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294), 1926-1928.
Shi, Rushen S., Cutler, Anne, Werker, Janet and Cruickshank, Marisa 2006. Frequency and form as determinants of functor sensitivity in English-acquiring infants. *Journal of the Acoustical Society of America* 119(6), El61-El67.

Tom Ruette, Dirk Speelman, Dirk Geeraerts
Quantitative Lexicology and Variational Linguistics, University of Leuven
tom.ruette@arts.kuleuven.be, dirk.speelman@arts.kuleuven.be,
dirk.geeraerts@arts.kuleuven.be

Aggregating and interpreting lexical alternation variables. Benefits of Weighted Multidimensional Scaling for lectal categorization

Research question

In this corpus-based sociolectometric study, we quantify the (lexical) difference between language varieties by aggregating a large number of lexical alternation variables with a technique that prevents variable-level details from being obscured by the aggregation step. This study consists of two parts: (a) finding an abundance of lexical alternation variables, and (b) carefully aggregating variables to find lectal patterns and to investigate the individual behavior of the variables. For step (a), we will fall back on a pre-made list of lexical alternation variables (Martin 2005). In the broader frame of this study, the lexical alternation variables are automatically generated on the basis of the similarity measures of semantic Vector Space Models (Turney & Pantel 2010; Peirsman et al. 2010). This aspect of the study will receive less attention in this paper, in which we focus on the aggregation procedure of step (b).

An inherent problem of aggregation techniques, as used in dialectometry (Goebl 1982; Nerbonne 2006) and sociolectometry, is that information of the individual variables is obscured or lost. This seems to be the price to pay for being able to make general claims, based on large amounts of data. Although a common metaphor is that the dialectometricist wants to see THE FOREST FOR THE TREES (see e.g. Szmrecsanyi, to appear), the loss of the "trees" (i.e. the linguistic variables) is problematic for our current purpose. We want to link the lectal patterns back to (types of) linguistic variables. Previous studies from both dialectometry and sociolectometry relied on manual scrutiny of complex in-between steps (Geeraerts et al. 1999; Soares da Silva 2010) or statistical comparison of aggregation solutions (Spruit et al. 2009) for assessing this link. The current study now proposes an aggregation method that grants more transparent access to the behavior of the individual variables.

Methodology

The cornerstone of the proposed method lies in the application of Weighted Multidimensional Scaling (WMDS, also referred to as "Individual Differences Multidimension Scaling", see Cox & Cox (2001)). Traditionally, all variables are aggregated into one distance matrix that averages the behavior of every individual variable in each lect. Then, a dimension reduction technique takes this single distance matrix and identifies a coordinate for every lect on the retrieved dimensions, allowing for direct visualizations (Speelman et al. 2003) or further analysis (Szmrecsanyi 2011). WMDS, on the other hand, has no (theoretical) restrictions on the amount of distance

matrices that can be used. As a consequence, we are not forced to aggregate all variables into one distance matrix — exactly this caused the loss of information at the variable level in a traditional approach — and we can create a distance matrix for every single variable (or for pre-defined subgroups of variables). Using multiple distance matrices in a lectometric study is not new (see e.g. Spruit et al. 2009, mentioned above); the application of WMDS in a sociolectometric study, however, is innovative.

The outcome of a WMDS analysis consists of two parts. On the one hand, a single reduced space is returned, which is very similar to a typical MDS solution. Therefore, the WMDS approach is compatible with existing methodologies. On the other hand, WMDS also returns a *weight configuration space*, which gives information on the importance (weight) of every single input distance matrix for every dimension. This weighting coefficient is thus the key to an interpretation of the behavior of the variables.

Results and conclusions

In order to show the application of WMDS, we set out to verify the classification of "typical Belgian Dutch words" in the "Referentielijst Belgisch Nederlands" (Reference List Belgian Dutch, RBBN), described in Martin (2005). For a quick overview of the Dutch situation, see Section 4.1 of Geeraerts (2003). The verification here will be based on a very large corpus that combines spontaneous conversations, Usenet posts, popular and quality newspaper articles and official government announcements from both Belgium and The Netherlands. The incorporation of registers in this regionally patterned corpus will allow us to see the multivariate strength of a WMDS-based sociolectometric study.

The RBBN classified more than 4000 Belgian Dutch words manually into categories. Here, we will focus on the categories "colloquialisms" and "unique variants". If we perform a traditional sociolectometric study (cf. Speelman et al. 2003), we see that texts in our corpus are divided in two dimensions on the basis of a strong regional difference and a slightly weaker register difference. From these results, one might assume that the "unique variants" and the "colloquialisms" are together responsible for the regional dispersion, but that the "colloquialisms" alone cause the register dimension. This assumption — which is basically the assumption that the categorization in the RBBN is accurate — can now be tested with WMDS.

In the WMDS approach, every variable on the RBBN (a combination of a Belgian Dutch word and its Netherlandic Dutch counterpart, hence a lexical alternation variable) is used to create its own distance matrix. These matrices are the input for the actual WMDS calculation. The first part of the WMDS outcome, the single, reduced space, shows the region and register variation on separate dimensions, just like the traditional approach. From the weight configuration space, however, we can see which variables "have more weight" on each of these dimension. It appears that the RBBN classification in "colloquialisms" and "unique variants" is distinctly related to the register and regional dimensions, yet not in a categorical fashion. From this, we conclude that the WMDS approach adds transparency and interpretability to a sociolectometric study. The comparison of the weight configuration space with the manual RBBN categorization adds trust in the accuracy of the methodology.

References

Cox, Trevor and Cox, Michael 2001. *Multidimensional Scaling*. Chapman & Hall.

Geeraerts, D., Grondelaers, S. and Speelman, D. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding-en voetbaltermen*. Amsterdam: Meertens Instituut.
Geeraerts, Dirk 2003. Cultural models of linguistic standardization. *Cognitive Models in Language and Thought. Ideology, Metaphors and Meanings* Dirven, René, Frank, Roslyn and Pütz, Martin (eds.), 25-68. Berlin: Mouton de Gruyter.
Goebl, Hans 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie.* Wien: Oesterreichische Akademie der Wissenschaften.
Martin, Willy 2005. *Het Belgisch-Nederlands anders bekeken: het Referentiebestand Belgisch-Nederlands (RBBN)*. Tech. rept. Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.
Nerbonne, John 2006. Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21, 463-476.
Peirsman, Yves, Geeraerts, Dirk and Speelman, Dirk 2010. The Automatic Identification of Lexical Variation between Language Varieties. *Natural Language Engineering* 16(4), 469-490.
Soares da Silva, Augusto 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In *Advances in Cognitive Sociolinguistics*, Geeraerts, Dirk, Kristiansen, Gitte and Peirsman, Yves (eds.). Berlin/New York, Mouton de Gruyter.
Speelman, Dirk, Grondelaers, Stefan and Geeraerts, Dirk 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317-337.
Spruit, Marco Rene, Heeringa, Wilbert and Nerbonne, John 2009. Associations among Linguistic Levels. *Lingua* 119(11), 1624-1642.
Szmrecsanyi, Benedikt 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1).
Szmrecsanyi, Benedikt to appear. Aggregate data analysis in variationist linguistics. In *Research Methods in Language Variation and Change*, Krug, Manfred and Schlüter, Julia (eds.). Cambridge: Cambridge University Press.
Turney, Peter and Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Cristina Sánchez-Marco, Stefan Evert
Universitat Pompeu Fabra, Barcelona, Institute of Cognitive Science, University of Osnabrück
cristina.sanchezm@upf.edu, stefan.evert@uos.de

Measuring semantic change: The case of Spanish participial constructions

An empirical study of semantic change

The goal of this paper is to explore different quantitative measures of semantic change and thus contribute to the development of statistical techniques for diachronic studies of meaning. Specifically, we focus on semantic change in the Spanish participial constructions *haber* 'have' + participle, *ser* 'be' + participle, *estar* 'be/stay' + participle, *tener* 'have/possess' + participle.

In 20th-century Spanish these constructions have clearly distinct meanings: they express the perfect, the verbal passive, the resultant state passive and a stative meaning, respectively. In contrast, the meanings of these participial constructions before the 17th century were quite different. *Ser* + participle was mainly used as a passive, but could also be a perfect for some intransitive and reflexive verbs, as well as a resultant state passive. *Haber* + participle could also have a stative meaning similar to the construction *tener* + participle in contemporary Spanish. *Estar* + participle and *tener* + participle emerged for the first time in the 13th century, but it took them more than 300 years to

generalise their meanings as they are today. The following examples illustrate the use of *ser* + participle as a passive nowadays and its use as a perfect in the earlier centuries.

(1) *Su monotonía es interrumpida solamente por algunas llamadas quejumbrosas.*
His monotony is interrupt$_{PTCP.F.S}$ only by some calls plaintive
'His monotony is interrumpted only by a few plaintive calls.' (20th c.)

(2) *Myo Çid Ruy Diaz a Alcolçer es venido.*
Myo Çid Ruy Diaz to Alcolçer is come$_{PTCP.M.S}$
'Myo Çid Ruy Diaz has come to Alcolçer.' (12th c.)

Approach

There is a large body of literature on semantic change, proposing different explanations for the change in Spanish participial constructions (Mendeldoff 1964, Vincent 1982, Pountain 1985, Aranovich 2003, Copple 2009). However, we are not aware of any empirical study that supports these claims with quantitative evidence. Our goal is to fill this gap, using data from a representative diachronic corpus of Spanish (see the section Data below).

The empirical study reported here addresses two basic questions:

A. Is there a *significant change* in the frequency and usage of all four participial constructions?
B. If a significant change took place, *how* and *why* did it take place?

We argue that quantitative data can play an important role in answering such questions. In particular, we explore the following approaches in order to measure and explain semantic change:

1. *Frequency and productivity of participial constructions.* We track changes in the frequency and productivity of all four constructions from the 12th to the 20th century. Our hypothesis is that *specialization* of *ser* + participle as the passive leads to a decrease both in frequency and productivity, as the auxiliary is applicable in fewer contexts and can be combined with fewer types of predicates. Conversely, *grammaticalization* of *haber* + participle as a perfect should lead to an increase in frequency and productivity.
2. *Distributional measures of semantic variability.* Following Sagi et al. (2009), we define the semantic density of a construction as the average distance between distributional representations of its instances in the corpus. If the construction undergoes specialization, it should be applicable in a narrower range of contexts and hence its semantic density should increase (*mutatis mutandum* for grammaticalization). Expanding on this token-based method, we also compare distributional representations of the types generated by the construction with those of the corresponding base predicates.
3. *Similarity as an explanation for semantic change.* We use distributional methods to determine the semantic similarity between the auxiliaries *estar*, *ser*, *tener* and *haber* in different centuries. Our hypothesis is that semantic similarity of *estar* with *ser* and of *tener* with *haber* might have contributed to the emergence of the

Data

The data for this study have been retrieved from a Spanish diachronic corpus consisting of 651 documents from the 12th to the 20th century, with a total size of more than 40 million words. The corpus comprises a variety of genres (fiction and nonfiction) and the documents come from different sources: Data from the 12th century to the 1950s were collected from the electronic texts transcribed and compiled by the *Hispanic Seminary of Medieval Studies*[1], the *Gutenberg project*[2] and the *Biblioteca Cervantes*[3]. This part of the corpus has been annotated automatically with linguistic information (morphosyntactic tag and lemma), using an expanded version of the Freeling morphological analyzer (Sánchez-Marco et al. 2010). Tagging accuracy in the oldest texts of the corpus yields 92%, which is sufficient to make reliable statistical generalizations.[4] Additional texts from the years 1975 to 1995 were obtained from the *Lexesp corpus* (Sebastian-Gallés 2000).

In order to facilitate statistical analysis of the data, we divided the corpus into four main periods, following the customary division determined by external historical events: Middle Spanish (1100-1492), Modern Spanish (1493-1788), Contemporary Spanish (1789-1974), and Late Contemporary Spanish (1975-2000).

Results

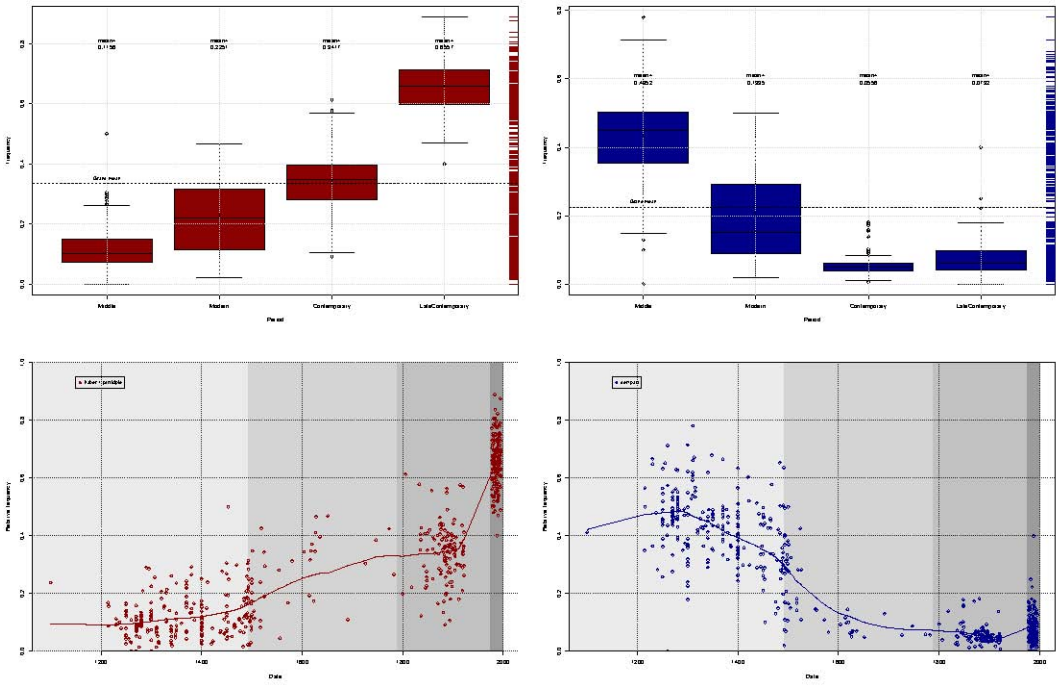
So far, we have completed a detailed study of changes in the usage frequency of each participial construction from the 12th to the 20th century. Frequency counts were obtained using the IMS Open Corpus Workbench[5] and analyzed with the open-source statistical software R (R Development Core Team 2010).

Figure 1 illustrates the development of *haber* + participle (left column) and *ser* + participle (right column). Each point in the two bottom panels corresponds to a single text from the corpus, showing time of composition on the x-axis and the relative frequency of the corresponding construction on the y-axis. From these graphs, it is obvious that the frequency of *haber* + participle increases continuously from the 12th to the 20th century; the change accelerates from the 15th century on. The frequency of *ser* + participle decreases, especially during the 15th and 16th centuries.

The boxplots in the top row of Figure 1 compare pooled data for the four main periods. All differences between Middle, Modern and Contemporary Spanish are highly significant (Generalized Linear Model with binomial family and logit link, $p < .001$). We interpret these findings as evidence for the grammaticalization of *haber* + participle and the specialization of *ser* + participle.

[1] See Corfis et al. [1997], Herrera and de Fauve [1997], Kasten et al. [1997], Nitti and Kasten [1997], O'Neill [1999], Sánchez et al. [2003].
[2] http://www.gutenberg.org/wiki/Main_Page.
[3] http://www.cervantesvirtual.com/.
[4] Sánchez-Marco, p. c.
[5] http://cwb.sourceforge.net/.

FIGURE 1. FREQUENCIES OF *HABER* + PARTICIPLE AND *SER* + PARTICIPLE FROM THE 12TH TO THE 20TH CENTURY.



Work in progress

We are currently working on the other quantitative approaches described in the section Approach above. In particular, we have compiled distributional semantic models for each century, which allow us to derive (i) *type vectors* representing each type generated by a specific construction in a particular century and (ii) *context vectors* representing token instances of each construction in a particular century. Based on these data sets, we will perform experiments with methods 2 and 3.

We are also collecting type-token statistics for the participial constructions. These data will be analysed with LNRE models of quantitative productivity (Baayen 2001) in order to complete our application of method 1. We intend to use the open-source R package *zipfR* (Evert & Baroni 2007) for this purpose.

References

Aranovich, Raúl 2003. The semantics of auxiliary selection in Old Spanish. *Studies in Language*, 27:1-37.

Baayen, R. Harald 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Copple, Mary T. 2009. *A Diachronic Study of the Spanish Perfect(ive): Tracking the Constraints on a Grammaticalizing Construction*. PhD thesis, University of New Mexico.

Corfis, Ivy A., O'Neill, John and Beardsley, Theodore S. Jr. (eds.) 1997. *Early Celestina Electronic Texts and Concordances*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

Evert, Stefan and Baroni, Marco 2007. *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29-32, Prague, Czech Republic, June.

Herrera, María Teresa and González de Fauve, María Estela (eds.) 1997. *Concordancias electrónicas del corpus médicos español*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.

Kasten, Llyod, Nitti, John and Jonxis-Henkemens, Wilhemina (eds.) 1997. *The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio*. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.
Mendeldoff, Henry 1964. The passive voice in old spanish. *Romanistisches Jahrbuch*, 15:269-287.
Nitti, John and Kasten, Llyod (eds.) 1997. The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts. Hispanic Seminary of Medieval Studies, Madison, Wisconsin.
O'Neill, John 1999. Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.
Pountain, Christopher 1985. Copulas, verbs of possession and auxiliaries in Old Spanish: The evidence for structurally interdependent changes. *Bulletin of Hispanic Studies*, 62:337-355.
R Development Core Team 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Sagi, Eyal, Kaufmann, Stefan and Clark, Brady 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens, Greece.
Sebastián-Gallés, Núria 2000. *LEXESP: léxico informatizado del español*. Edicions Universitat Barcelona.
Nieves Sánchez, Maria, Herrera, María Teresa and Zabía, María Purificación 2003. *Textos medievales misceláneos*. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.
Sánchez-Marco, Cristina, Boleda, Gemma, Fontana, Josep Maria and Domingo, Judith 2010. Annotation and representation of a diachronic corpus of Spanish. In *Proceedings of LREC,* Valletta, Malta. Online at http://www.lrec-conf.org/proceedings/lrec2010/pdf/535_Paper.pdf [March 05, 2011].
Vincent, Nigel 1982. The development of auxiliaries *habere* and *esse* in Romance. In *Studies in the Romance Verb*, Vincent, Nigel and Harris, Martin (eds.), 71-96. London: Croom Helm.

Ankelien Schippers, Machteld Brands
University of Groningen, Groningen
a.schippers@rug.nl, m.brands@student.rug.nl

Effects of complexity on the acceptability of long-distance wh-dependencies

This study concerns a grammaticality judgment task in Dutch investigating the effects of complexity on the acceptability of long-distance (LD) and partial wh-movement (PM) constructions, exemplified in (1) and (2), respectively. Both constructions are attested in Dutch, although LD-movement is more common (cf. Barbiers et al. 2008, Schippers 2008, Strik 2009).

Dutch
(1) [$_{CP}$ *Wie zei Mark* [$_{CP}$ t_{wie} *dat hij* t_{wie} *had bedankt*?]]
Who said Mark that he had thanked
'Who did Mark say he thanked?'

Dutch
(2) [$_{CP}$ *Wat zei Mark* [$_{CP}$ *wie hij* t_{wie} *had bedankt*?]]
What said Mark who he had thanked
'Who did Mark say he thanked?'

PM-constructions have the same interpretation as their LD-movement counterparts, but differ from them syntactically. Instead of moving the wh-phrase to its matrix scope position as in (1), it is moved no further than the intermediate CP. The matrix CP is

instead occupied by a so-called scope marker *wat* 'what'. Nonetheless, the intermediate wh-phrase does take matrix scope. PM-constructions are either analyzed as involving two separate movement chains (cf. Dayal 1994), which are only linked semantically, or as a continuous chain with intermediate spell-out of the wh-phrase (cf. McDaniel 1989).

Many Germanic varieties, including German, Dutch and Frisian have both LD and PM constructions in the language. One open question is what determines the choice for one construction over the other, both within and between particular languages. In this study, we investigated the role of the complexity by experimentally manipulating the length of the dependency.

It is well known that long-distance wh-dependencies are difficult to process. This is usually contributed to the relatively large distance between the wh-phrase and its gap (cf. Phillips et al. 2005). An uncontroversial assumption in formal syntax is that wh-movement leaves intermediate traces at clause edges. It appears that such intermediate traces may facilitate processing, as suggested in a study by Gibson & Warren (2004). An important difference between LD-constructions and PM-constructions is that the intermediate chain links in LD-movement constructions are deleted at PF. In case of partial movement, however, the long-distance dependency is spelled out strictly local. This may play an important factor in favoring PM-movement over LD-movement once the complexity of the dependency increases. We tested this hypothesis by means of a grammaticality judgment task in which 34 native speakers of Dutch were asked to rate wh-dependencies on a scale from 1-10, were 10 represents 'very acceptable'. Constructions as in (1) and (2) with one embedding were compared to the more complex cases in (3) and (4) with two embeddings.

Dutch

(3) [CP *Wie zei Jantine* [CP t_{wie} *dat zij dacht* [CP t_{wie} *dat zij* t_{wie} *had herkend?*]]]
Who said Jantine that she thought that she had recognized?
'Who did Jantine say that she thought that she had recognized?'

Dutch

(4) [CP *Wat zei Jantine* [CP *wat zij dacht* [CP *wie zij* t_{wie} *had herkend?*]]]
What said Jantine what she thought who she had recognized
'Who did Jantine say that she thought that she had recognized?'

To control for length effects and to abstract away from irrelevant lexical differences, declarative counterparts to (1)-(4) were also tested, and both the raw scores of the wh-dependencies and the difference scores (computed by subtracting the scores of the wh-dependencies from their declarative counterparts) were analyzed in a repeated measures ANOVA.

The raw scores showed that LD-constructions were judged significantly better than PM-constructions, and that constructions with one embedding were rated higher than those with two embeddings. However, there was no interaction between type of movement and length of the dependency, contrary to what is to be expected if partial movement increases the acceptability of longer wh-dependencies. Interestingly, however, on closer inspection it turned out that two groups of speakers could be distinguished, namely PM-speakers and LD-speakers. PM-speakers were those who rated PM constructions with one embedding higher than their LD-counterparts. For this group, there was a length effect, but only for the PM-constructions: PM-constructions

with two embeddings (PM-2) were rated lower than those with one embedding (PM-1), while there were no significant differences between the LD-movement constructions (see Figure 3). The LD-movement speakers showed a reverse result: they judged LD-movement with two embeddings (LD-1) lower than the LD-constructions with one embedding (LD-2), but there were no significant differences between the PM-constructions (see Figure 1). Similar effects were observed for the difference scores, illustrated in Figure 2 and 4.

FIGURE 1. RAW SCORES LD-SPEAKERS.

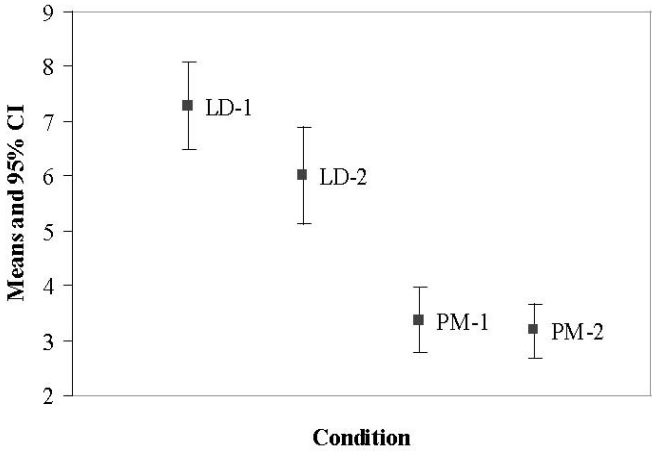


FIGURE 2. DIFFERENCE SCORES LD-SPEAKERS.

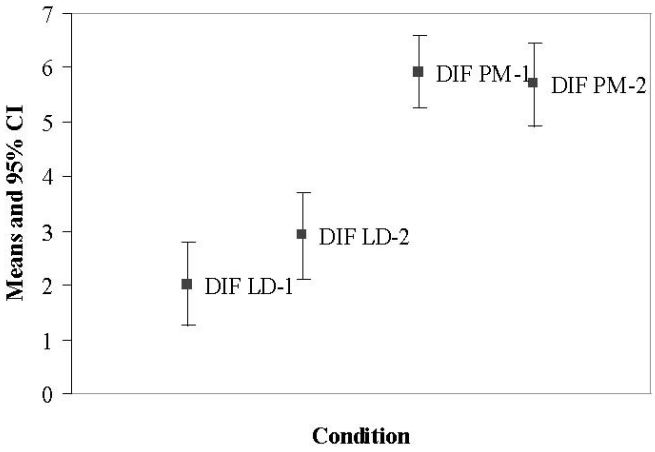


FIGURE 3. RAW SCORES PM-SPEAKERS.

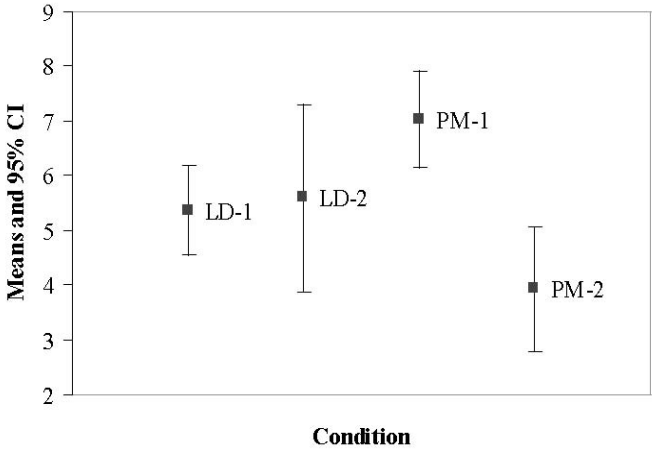
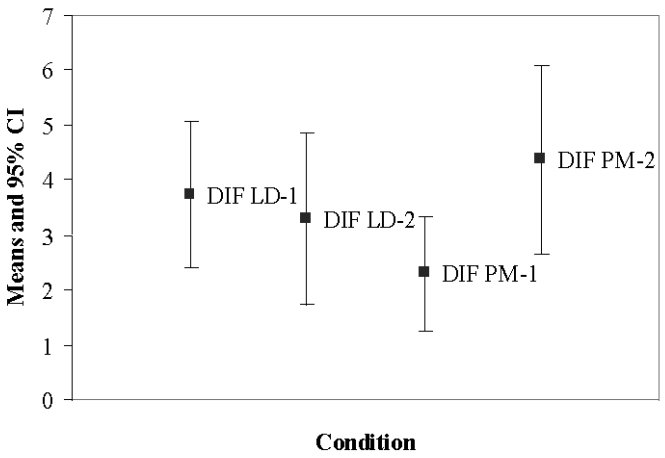


FIGURE 4. DIFFERENCE SCORES PM-SPEAKERS.



A number of interesting conclusions can be drawn from these results. First of all, it turns out that grammaticality judgment scores reflect the degree of complexity of a construction, even though the participants in this experiment were specifically instructed not to pay attention to the complexity or length of the items. Second, this effect is persistent even when the difference scores between the extraction sentences and equally long declaratives are being considered. The data therefore form important evidence for the hypothesis that acceptability judgments (amongst other things) reflect syntactic complexity (cf. Gibson 1998; Fanselow & Frisch 2006).

Third, the current study provides no evidence for the hypothesis that partial wh-movement chains improve the acceptability of longer movement chains: for the LD-speakers the complexity effect was only observed for LD-constructions, while for the PM-speakers the PM-constructions with two embeddings were worse than the ones with one embedding.

Finally, the data discussed here also provide evidence for the idea that partial wh-movement and LD-movement are in complementary distribution: separate groups of speakers can be distinguished that clearly have a preference for one construction over the other. Furthermore, these groups are able to differentiate between complex and simpler variants only for the most preferred constructions. The fact that for the PM-speakers the difference between LD and PM constructions is less pronounced than it is for the LD-speakers is very likely due to underlying frequency differences (cf. Ford & Bresnan 2010), since LD-movement is still much more prevalent in Dutch.

References

Barbiers, Sjef, Koeneman, Olaf and Lekakou, Marika 2008. Syntactic Doubling and the Structure of Chains. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Chang, Charles B. and Haynie, Hannah J. (ed.), 77-86. Somerville MA: Cascadilla Proceedings Project.

Bresnan, Joan and Ford, Marilyn 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86, 186 -213.

Dayal, Veneeta 1994. Scope Marking as Indirect Wh-Dependency. *Natural Language Semantics* 2, 137-170.

Fanselow, Gisbert and Frisch, Stefan 2004. Effects of processing difficulty on judgments of acceptability. In *Gradience in grammar,* Fanselow, Gisbert, Féry, Caroline, Schlesewsky, Matthias and Vogel, Ralf (eds.), 291 -316. Oxford: Oxford University Press.

Gibson, Edward 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1-76.

Gibson, Edward and Warren, Tessa 2004. Reading time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7, 55-78.
Phillips, Collin, Kazanina, Nina and Abada, Shani 2005. ERP Effects of the Processing of Syntactic Long-distance Dependencies. *Cognitive Brain Research* 22, 407-428.
Schippers, Ankelien 2008. Partial wh-movement and wh-copying in Dutch: Evidence for an indirect dependency approach. Ms., Groningen: University of Groningen.
Strik, Nelleke 2009. *Syntaxe et Acquisition des Phrases Interrogatives en Français et en Neerlandais : une Etude Contrastive*. PhD dissertation, Université Paris 8 – Saint Denis.

Pavel Štichauer
Charles University in Prague, Czech Republic
pavel.stichauer@ff.cuni.cz

The quantitative approach to morphological productivity in a diachronic perspective

Introduction: the aim of the paper

The quantitative corpus-based approach to morphological productivity, based on Baayen's work (1992; 2001; 2008), has become a major paradigm in the synchronic studies of the productivity of word formations processes (e.g. Plag 1999; 2006; Gaeta & Ricca 2002; 2003; 2006; Dal 2003). However, there are still not many studies which would apply the method diachronically, i.e. on diachronic corpora covering, preferably, more periods (cf. Lüdeling & Evert 2005; Scherer 2007; Säily & Suomela 2009; Štichauer 2009). This is obviously due to the fact that, with diachronic corpora, there are more problems to be solved before one can proceed to the calculation of the productivity of a given process (cf., e.g., Baayen 2009:909-910). My aim is to formulate some methodological prerequisites and to show the results on one concrete example.

The quantitative notion of morphological productivity

It is well known that Baayen's corpus-based approach conceives of productivity as the likelihood of observing a new type when sampling a sufficiently large corpus. The gradually increasing number of new types (type frequency, V) may in fact be seen as a function of token frequency (N): with the increasing number of tokens (given by the corpus size), the number of types will also increase (cfr. Baayen 1992:113). This relation gives rise to the definition of *vocabulary growth curve* and to the notion of *vocabulary growth rate*, the latter being captured by the proportion of *hapax legomena* (V_1) to the overall number of tokens. The obvious fact that this measure cannot be used for different-sized corpora can now be easily overcome by two techniques (*binomial interpolation* and *extrapolation* based on LNRE models of word frequency distributions, cfr. Baayen 2001; Evert 2004), which are implemented in the package *zipfR*, recently developed by Marco Baroni and Stefan Evert (Baroni & Evert 2006; http://zipfr.r-forge.r-project.org/).

Principles of lemmatization

However, before one can use these techniques, some work of linguistic *pre-processing* is necessary. In diachronic corpora, the process of *lemmatization* is particularly tricky because of two facts. First, the lemmatization requires a particular attention to

orthographic and phonomorphological variants. Second, one needs to eliminate all the types which have nothing to do with the word formation rule in question. If this is not done *manually*, the "extraction noise" (Evert 2005:63) will probably distort all possible outcomes. Furthermore, there are problems with the *inhomogeneity* of the underlying corpus. Diachronic corpora (as opposed to large synchronic corpora) tend to be *inhomogeneous* for at least two reasons: first, there may be different text types with different proportions across the corpora one wishes to compare (cf. Baayen 2009:910); second, there is a strong tendency (perhaps, a stronger one than for synchronic corpora) to the so-called *clustering / repetition effects* (cf. Evert 2005:59). In other words, diachronic corpora tend to display a stronger *non-randomness* than could possibly be corrected for within a statistical model (cf. Evert 2006; Baayen 2009:910).

Case study: suffixes *-mento / -zione* in Old Italian

I wish to put forward a partial solution to these problems by presenting one concrete example. I intend to show the diachronic development of two Italian deverbal suffixes *-mento / -zione* within the time span that goes from the 13th to the 16th century. The data have been sampled from four different-sized subcorpora created out of the known corpus LIZ 4.0. All the formations in *-mento* and *-zione* have been lemmatized manually, they have been checked against their particular contexts and they were "filtered" by using five major "type elimination" criteria put forward, among others, by Gaeta & Ricca 2002; 2003; 2006 (*strong opacity, baseless formations, nominal bases / different semantic instruction, derivational inner cycles, specific borrowings*). In order to show the diachronic aspect of the productivity of these two suffixes, the above mentioned tools of lexical statistics, implemented in the package *zipfR*, will be used, especially the technique of extrapolation, as the four subcorpora are of gradually increasing sizes. I will show that the suffix *-mento* tends to be constant in its productivity across the four periods in question, while *-zione* displays interesting diachronic variability.

This diachronic variability is particularly interesting because it has important repercussions on the present-day situation of these two suffixes, as described by Gaeta - Ricca 2002; 2003; 2006. First, it is the diachronically increasing token frequency of *-zione* derivatives. In the present-day Italian, *-zione* is by far the most frequent suffix (cf. Gaeta & Ricca 2003; 2006) and with respect to *-mento* it is, of course, of more limited productivity. Second, it is the proportion of *hapax legomena* to the overall number of types. I wish to show how this quantitative complementarity can be traced back to the 15th and 16th century where the Latin influence leads probably not to a systematic change in the derivational paradigm (*-mento* vs. *-zione*), but to a massive borrowing process. It is only later that *-zione* settles down as an independent and „available“ (in Corbin's sense) affix, being largely dependent upon other changes in the derivational paradigms (especially the increasing productivity of *-izzare* verbs which are an exclusive input to the *-zione* derivatives (cf. Gaeta & Ricca 2006:70).

References

Baayen, Harald 1992. Quantitative aspects of morphological productivity. In Booij, Geert, Marle, Jaap van (eds.). *Yearbook of Morphology 1991*, 109-149. Dordrecht: Kluwer.

Baayen, Harald 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, Harald 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, H. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke, Merja,

Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, 899- 919, Berlin, Mouton de Gruyter.
Baroni, Marco and Evert, Stefan 2006. The *zipfR* package for lexical statistics: A tutorial introduction (http://www.cogsci.uni-osnabrueck.de/~severt/zipfR/).
Dal, Georgette 2003. Productivité morphologique: définitions et notions connexes. *Langue française*, 140, 3-23.
Evert, Stefan 2004. A simple LNRE model for random character sequences. *Proceedings of JADT* 2004, 411-422.
Evert, Stefan 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004, published in 2005; available from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/ .
Evert, Stefan 2006. How Random is a Corpus? The Library Metaphor. *Zeitshrift für Anglistik und Amerikanistik* 54(2), 177-190.
Gaeta, Livio and Ricca, Davide 2002. Corpora testuali e produttività morfologica: i nomi d'azione in due annate della *Stampa*. In *Parallela IX. Testo – variazione – informatica. Text – Variation – Informatik*, Bauer, R. - Goebl, H. (a cura di), 223-249. Wilhelmsfeld: Gottfried Egert Verlag..
Gaeta, Livio and Ricca, Davide 2003. Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Italian Journal of Linguistics/Rivista di Linguistica* 15, 1, 63-98.
Gaeta, Livio and Ricca, Davide 2006. Productivity in Italian word formation: A variable- corpus approach. *Linguistics* 44, 1, 57-89.
LIZ 4.0 2001. *Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*, a cura di Pasquale Stoppelli ed Eugenio Picchi. Bologna: Zanichelli.
Lüdeling, Anke and Evert, Stefan 2005. The Emergence of Non-Medical -itis. Corpus Evidence and Qualitative Analysis. In *Linguistic evidence. Empirical, Theoretical, and Computational Perspectives*, Kepser, S. and Reis, M. (eds.), 315-333. Berlin: Mouton de Gruyter.
Plag, Ingo 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.
Plag, Ingo 2006. Productivity. In *The Handbook of English Linguistics*, Aarts, B. and McMahon, A. (eds.), 537-557. Oxford: Blackwell.
Säily, Tanja and Suomela, Jukka 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In *Corpus Linguistics: Refinements and Reassessments* [Language and Computers: Studies in Practical Linguistics 69], Renouf, A. and Kehoe, A. (eds.), 87-109. Amsterdam: Rodopi,
Scherer, Carmen 2007. The role of productivity in word-formation change. In *Historical Linguistics 2005: Selected papers from the 17th International Conference on Historical Linguistics, Madison, Wisconsin, 31 July - 5 August 2005*, Salmons, Joseph C. and Shannon Dubenion-Smith (eds.), 257-271. Amsterdam: John Benjamins.
Štichauer, Pavel 2009. Morphological productivity in diachrony: the case of the deverbal nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th century. In *Selected Proceedings of the 6th Décembrettes*, Fabio Montermini, Gilles Boyé and Jesse Tseng (eds.), 138-147. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2241.

Daphne Theijssen, Joan Bresnan, Marilyn Ford
Radboud University Nijmegen, Stanford University, Griffith University
d.theijssen@let.ru.nl, bresnan@stanford.edu, m.ford@griffith.edu.au

In a land far, far away… The dative alternation in British, American and Australian English

Introduction

In many situations, one can choose between several syntactic options that are equally grammatical, but that may differ in their acceptability in the given context. An example

is the dative alternation in English, for which speakers and writers can choose between structures with a prepositional dative (1) or double object structure (2).

(1) The evil queen gave the poisoned apple to Snow White.
(2) The evil queen gave Snow White the poisoned apple.

Previous research has shown that there is a general tendency to place animate before inanimate nouns, shorter before longer, pronouns before non-pronouns and definite before indefinite. Notably, these observations have been made both in psycholinguistic research (e.g. Prat-Sala & Branigan 2000), in corpus studies (e.g. Snyder 2003, Szmrecsanyi 2006, Bresnan et al. 2007), and in studies that combine the two (e.g. Arnold et al. 2000, Rosenbach 2005, Bresnan & Ford 2010).

In this paper, we compare the distributional properties of the dative alternation in different variants of the same language: English. More specifically, we compare American, British and Australian English. Previous studies have shown that there are distributional differences in the dative alternation across these variations and across time (Rohdenburg 2007, Grimm & Bresnan 2009, Bresnan & Ford 2010). Given the history of English in the three countries, we would expect to see the most differences between American and Australian English. These varieties have originated from British English separately from each other, at different points in time. Since Australian English is the younger variety, we expect that the difference between Australian and British English is smaller than that between American and British English. Moreover, various studies have shown that when it comes to fine-grained syntactic changes over time, American English seems to lead the way: 'Americanization' (e.g. Leech & Smith 2006, Szmrecsanyi 2010).

Method

We replicated the judgment study of Bresnan and Ford (2010), but included British English, used a wider age range (20 to 65 years), and conducted it through a website instead of on paper. Participants read a short passage followed by two possible continuations: one with a double object construction, one with a prepositional dative. They were asked to rate the naturalness of both options by dividing 100 points between them: the more points, the more natural. We used the same 30 items taken from the (American English) Switchboard corpus as in Bresnan and Ford (2010), localizing them to British and Australian English by replacing American-specific vocabulary and place names. All items were presented in random order, and the order in which the two options were presented was alternated. Table 1 shows the characteristics of the participants.

TABLE 1. CHARACTERISTICS OF PARTICIPANTS IN JUDGMENT STUDY.

	Female				*Male*			
	N	av age	min age	max age	N	av age	min age	max age
British	22	32.0	21	61	18	31.5	21	63
American	22	37.3	21	65	13	32.1	21	61
Australian	20	34.0	23	63	17	32.1	20	64

Using linear regression, we modeled the participants' ratings for the prepositional dative variant. We included the two grouping factors, speaker and item, as random effects in the model. The predictors we included as fixed effects were definiteness of the recipient

(*Snow White* in the example), definiteness of the theme (*the poisoned apple*), animacy of the recipient, pronominality of the theme and the log ratio of the length of the theme to the length of the recipient[1]. Moreover, we looked at age, sex, and variety, and their interaction with the other predictors. We also controlled for the order of the items, the order of the two options and the rating assigned for the previous item by including these as fixed effects.

We first built three separate models for the three varieties. Only the main effects and interactions that were significant[2] in (at least one of) these models were then included in the pairwise models, with interactions between variety.

Results

The significant effects in the three models are presented in Table 2.

TABLE 2. SIGNIFICANT EFFECTS AND THEIR ESTIMATES IN THE PAIRWISE LINEAR REGRESSION MODELS. POSITIVE ESTIMATES FAVOR THE PREPOSITIONAL DATIVE, NEGATIVE ONES THE DOUBLE OBJECT CONSTRUCTION.

	Am vs. Aus	*Br vs. Am*	*Br vs. Aus*
Length ratio (th/rec)	-23.15	-18.77	-19.59
Recipient = inanimate	43.46	34.48	42.30
Theme = indefinite	-13.31		-12.45
Theme = pronominal : Age		-0.23	
Variety = US : Length ratio	9.33		
Variety = US : Length ratio : Age	-0.14		
Variety = US	6.10	4.28	
Previous rating		-0.04	
Variety = US : Previous rating	-0.05		
(Intercept)	47.78	47.99	47.78

We see that length ratio and animacy of recipient are significant across the different varieties, and they are in line with the findings in previous research: animate precedes inanimate and shorter precedes longer. The same is true for the definiteness of the theme, although it just misses significance in the model with British and American judgment: definite precedes indefinite.

For British and American speakers, there is also an effect of the pronominality of the theme in interaction with age. It thus seems that older British and American people are less reluctant to use the double object construction when the theme is a pronoun (e.g. *give the man it*) than younger people. It has been shown in previous research that some British dialects show different patterns when the theme is pronominal: reversed double constructions such as *give it him* are also possible (Haddican 2010). There are also significant interactions between the length ratio, variety and age. Compared to Australians, Americans award higher naturalness scores to instances that violate the principle of end weight (measured by length ratio), as also

[1] In fact, seeing the correlation between the length ratio and the definiteness of the recipient and the theme, and the pronominality of the theme, we residualized the length ratio on these three predictors, and included the residuals in the regression model.

[2] We applied several variable selection methods, and the models presented are representative of what we found. They are established by first removing all interactions for which the regression estimate was below twice the standard error, and then a stepwise backward elimination with the remaining predictors until all p-values (based on the t-value) were below 0.05.

demonstrated in the judgment study by Bresnan and Ford (2010). In addition, our study shows that this is especially true for younger Americans, which could mean 'Americanization' is taking place: American English is leading in this change.

The only significant interactions with variety and linguistic predictors are in the model that compares American and Australian English. This is exactly what we expected: the largest differences exist between these two varieties, while British English, from which they originated, seems to be in the middle. There is a main effect of variety for the comparison between British and American English, but not between British and Australian English. This is again as expected: the younger variety Australian English shows no (significant) differences with British English. Finally, the rating that the participant assigned to the previous instance has an effect in the combined British and American English data, but not in Australian English

Conclusion

This paper shows that there are distributional differences between the dative alternation in American, Australian and American English. The effects we found in this judgment study are consistent with the history of the varieties and with 'Americanization'. In the near future, we plan to compare the dative alternation across the varieties in corpus data, and relate the results to this study.

References

Arnold, Jennifer, Wasow, Thomas, Losongco, Anthony and Ginstrom, Ryan 2000. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76(1), 28-55.

Bresnan, Joan and Ford, Marilyn 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86(1), 186-213.

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana and Baayen, R. Harald 2007. Predicting the Dative Alternation. In *Cognitive Foundations of Interpretation*, Bouma, Gerlof, Kraemer, Ineke and Zwarts, Joost (eds.), 69-94. Amsterdam: Royal Netherlands Academy of Science.

Grimm, Scott and Bresnan, Joan 2009. Spatiotemporal variation in the dative alternation: A study of four corpora of British and American English. *Third International Conference Grammar and Corpora*, 22-24 September 2009, Mannheim, Germany.

Haddican, William 2010. Theme-Goal Ditransitives and Theme Passivisation in British English Dialects. *Lingua* 120(10), 2424-2443.

Leech, Geoffrey and Smith, Nicholas 2006. Recent grammatical change in written English 1961-1992: Some preliminary findings of a comparison of American with British English. In *The changing face of corpus linguistics,* Renouf, Antoinette and Kehoe, Andrew (eds.), 185-204. Amsterdam and New York: Rodopi.

Prat-Sala, Mercè and Branigan, Holly P. 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language* 42(2), 168-82.

Rohdenburg, Günter 2007. Grammatical divergence between British and American English in the 19th and early 20th centuries. *The third Late Modern English Conference*, University of Leiden, September 1, 2007.

Rosenbach, Anette 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3), 613-44.

Snyder, Kieran 2003. *The relationship between form and function in ditransitive constructions*. PhD dissertation, University of Pennsylvania, Philadelphia.

Szmrecsányi, Benedikt 2006. *Morphosyntactic persistence in Spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Mouton de Gruyter.

Eline Zenner, Dirk Speelman, Dirk Geeraerts
FWO Flanders/University of Leuven, University of Leuven, University of Leuven
eline.zenner@arts.kuleuven.be, dirk.speelman@arts.kuleuven.be,
dirk.geeraerts@arts.kuleuven.be

A concept-based approach to measuring the success of loanwords

Introduction

This paper presents a multivariate, quantitative, corpus-based study in which we examine, by means of mixed effect regression modeling, the combined effect of a set of structural, conceptual, extra-linguistic and contact-related features affecting the success of 125 English person reference nouns (*manager, babysitter* …) in Dutch.

Background

THEORETICAL ACCOUNTS ON LEXICAL BORROWING often focus on the demarcation of types of loanwords (e.g. Haugen 1950, Filipovic 1977, Field 2002). Classifications are typically based on (1) objectively defined structural characteristics of the loanword (e.g. morphological adaptation); (2) subjective distinctions serving normative purposes, like the division between the tolerated "necessary" loanwords (i.e. those naming a new concept) and the denounced "luxury" loanwords (i.e. those introduced as alternative lexicalization for an existing concept) (see Onysko & Winter-Froemel *forthcoming* for discussion).

Current QUANTITATIVE CORPUS-BASED ANGLICISM RESEARCH aims at counting the number of English loanwords within the proposed classifications, trying to identify those classes which are most susceptible to foreign influence (e.g. Yang 1990, Nettmann-Multanowska 2003). Typically, the focus is restricted to classifications based on structural characteristics (i.e. the first class presented above). Furthermore, most studies show several methodological shortcomings, which have hitherto remained unaddressed. As such, a fully comprehensive empirical attempt at pin-pointing the locus of foreign influence is still missing.

The *aim of our paper* is twofold. First, we wish to encourage the quantitative perspective in loanword research by presenting ways to clear the recurring methodological hurdles, which we will present below. Second, we wish to open up the theoretical scope of current research by presenting objective operationalizations for the so far neglected distinctions (e.g. luxury vs. necessary loans) and by introducing new conceptual and extra-linguistic features that possibly influence variation in the success of anglicisms.

Methodology: Clearing the hurdles

DATA-COLLECTION: As most studies extract anglicisms manually from their sources, existing corpus-based anglicism research has typically relied on relatively small datasets (e.g. Fink 1997). However, empirical analyses dealing with lexicology and lexical variation require sufficiently large data collections (Armstrong 2001, Geeraerts 2010). For our analyses on variation in the success of 125 English loanwords we rely on a lemmatised and syntactically parsed Dutch newspaper corpus containing over 1.5 billion words. The corpus comprises material from 1999 to 2005 and represents the two national varieties of Dutch (Belgian Dutch and Netherlandic Dutch).

SUCCESS MEASURE: The second hurdle to be cleared is the definition of the success of a loanword. So far, this success has simply been equated with the token count for the borrowed item in a given corpus. However, as topic specificity can easily distort these counts and hence disfigure the results (Speelman *et al.* 2003), a more advanced success measure is required (e.g. Van Hout & Muysken 1994).

We propose a concept-based approach, taking alternative expressions into account. Adapting the profile-based method introduced by Geeraerts *et al.* (1999), we define the success of an anglicism as the corpus frequency of that anglicism, relative to the total frequency of the loanword and its synonyms (see example in Table 1).

TABLE 1. CONCEPT-BASED SUCCESS MEASURE.

Synonymous expressions for the concept BABYSITTER	Token frequencies in the Dutch corpus
babysit(ter)	1031
kinderoppas	456
success rate for *babysit(ter)*: 1031 / (1031 + 465) = **69.33 %**	

Our conceptualization of the success of a loanword is thus the extent to which it succeeds in "fighting off" alternative lexicalisations for a given concept. As such, maximal success is the situation in which the loanword is (or has become) the only occurring lexicalization for a given concept.

DATA COLLECTION: For this study we focus on 125 English person reference nouns (*fulltimer, freak*…) used in Dutch. Only lexical items which are clearly identifiable as English loanwords by native speakers of Dutch are included as anglicisms. Denotational synonyms for the loanwords were identified using multiple lexicographical sources, complemented with results from automatic synonym detection based on word space models (see Lin 1998 and Peirsman *et al.* 2007*)*. To avoid unreliable success-rates due to semantic specialization of the items in a profile, we specifically seek out profiles with a high degree of synonymy between the lexical items (avoiding fuzziness and near-synonymy as much as possible). Next, automated methods are introduced to retrieve tokens and remove noise (e.g. proper names, longer stretches of English) from the dataset. Then, we semi-automatically disambiguate remaining polysemous items (e.g. *chicken*). Using the resulting database, we finally measure the success rates for the set of 125 English nouns across eight subsections of the corpus (see example in Table 2). This way, we acquire 1000 success rates (8 per loanword), based on over 30 million tokens of person reference nouns.

OBJECTIVE DEFINITION OF PREDICTORS: As was mentioned above, current quantitative approaches only focus on the influence of structural features on variation in the success of loanwords. If we want to get a finer grip on the actual impact of the different classifications that have been proposed, a wider variety of features should be discussed. In this study, we therefore not only incorporate structural features, but also include (1) operationalizations for the distinction between necessary and luxury loans; (2) extra-linguistic features (e.g. regional variation - see Table 2); (3) conceptual features (e.g. entrenchment of the concept); (4) contact-related features (e.g. "travel time" from donor language to recipient language – Chesley & Baayen 2010).

STATISTICAL ANALYSES: A final step to a comprehensive empirical view on the success of loanwords is to inquire into the combined effect of the different influential features on the success measure. Using a logarithmic transformation of the success rate,

we perform mixed-effect linear regression analyses to study the interplay between the variables.

TABLE 2. SUCCESS-RATES ACROSS SUBSECTIONS OF THE CORPUS.

subcorpus:	success-rate for *babysit(ter)* in the subcorpus:
Belgian Dutch; 1999-2000; popular newspapers	348 / (348 + 135) = 0.720
Belgian Dutch; 1999-2000; qualitative newspapers	102 / (102 + 48) = 0.680
Belgian Dutch; 2001-2002; popular newspapers	339 / (339 + 95) = 0.781
Belgian Dutch; 2001-2002; qualitative newspapers	130 / (130 + 74) = 0.637
Netherlandic Dutch; 1999-2000; popular newspapers	33 / (33 + 17) = 0.660
Netherlandic Dutch; 1999-2000; qualitative newspapers	34 / (34 + 24) = 0.586
Netherlandic Dutch; 2001-2002; popular newspapers	19 / (19 + 30) = 0.388
Netherlandic Dutch; 2001-2002; qualitative newspapers	26 / (26 +33) = 0.441

Results and Conclusion

The results show the overall usefulness of the multivariate approach: features from all distinguished classes (structural, conceptual, extra-linguistic and contact-related features) contribute to the explanation of variation found in the success rates of the 125 anglicisms under scrutiny. Furthermore we find that the most important factor in the model is the distinction between luxury and necessary loans: the status of a loanword at the time of introduction in a language has a long-term effect on its success rate. We discuss two converging interpretations for this effect: (1) the influence of language planning; (2) the effect of conceptual entrenchment. Overall, the results from our empirical study indicate how current structuralist theories of the borrowing process need to be complemented with a variationist, usage-based perspective.

References

Armstrong, Nigel 2001. Social and stylistic variation in spoken French: a comparative approach. Amsterdam: John Benjamins.

Chesley, Paula, Baayen, R. Harald 2010. Predicting new words from newer words: Lexical borrowings in French. *Linguistics* 48(6).

Field, Fredric W. 2002. *Linguistic borrowing in bilingual contexts.* Amsterdam: John Benjamins.

Filipovic, Rudolf 1977. English words in European mouths and minds. *Folia Linguistica* 11, 195-206.

Fink, Hermann 1997. Von Kuh-Look bis Fit for Fun: Anglizismen in der heutigen deutschen Allgemein- und Werbesprache. Frankfurt am Main: Peter Lang.

Geeraerts, Dirk 2010. Lexical variation in space. In *Language and Space I. Theories and Methods*, Auer, Peter, Schmidt, Jürgen Erich (eds.), 820-836. Berlin/New York: Mouton de Gruyter.

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

Haugen, Einar 1950. The analysis of linguistic borrowing. *Language* 26(2), 210-231.

Lin, Dekang 1998. Automatic Retrieval and Clustering of Similar Words. *Proceedings of COLING-ACL98*, Montreal (Canada).

Nettmann-Multanowska, Kinga 2003. *English Loanwords in Polish and German after 1945: Orthography and Morphology*. Frankfurt am Main: Peter Lang.
Onysko, Alexander and Winter-Froemel, Esme forthcoming. Necessary loans? Luxury loans? Explaining the pragmatic dimension of borrowing.
Peirsman, Yves, Heylen Kris and Speelman, Dirk 2007. Finding semantically similar words in Dutch. Co-occurrences versus syntactic contexts. *Proceedings of the CoSMO workshop*, 9-16.
Speelman, Dirk, Grondelaers, Stef and Geeraerts, Dirk 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317-337.
Van Hout, Roeland and Muysken, Pieter 1994. Modeling lexical borrowability. *Language Variation and Change* 6(1), 39-62.
Yang, Wenliang 1990. *Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins der Spiegel*. Tübingen: Niemeyer.