New Developments in the Quantitative Study of Languages

Book of abstracts

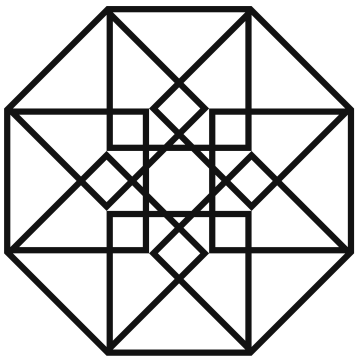
Organized by the *Linguistic Association of Finland*
http://www.linguistics.fi

28–29 August 2015

House of Science and Letters ("Tieteiden talo")
Kirkkokatu 6, 00170 Helsinki
http://www.linguistics.fi/quantling-2015/

Acknowledgements

Financial support from the Federation of Finnish Learned Societies is gratefully acknowledged.



FEDERATION OF FINNISH
LEARNED SOCIETIES

Contents

Part I.

Keynotes

Advances in computer-assisted quantitative historical reconstruction

Michael Cysouw

Philipps University Marburg

Historical-comparative linguistics is one of the hallmarks of linguistic research, and arguably the origin of linguistics as a discipline in the 19th Century. Throughout the 20th Century there have been various attempts to quantify historical reconstruction (cf. early approaches like Kroeber & Chrétien 1937 or Swadesh 1952), but none of those attempts really managed to take hold in the linguistic methodological toolbox. In the last decade, there is again a rising interest in applying computational methods in historical linguistics, with ever more advanced mathematical approaches (cf. Bouckaert et al. 2012, Bouchard-Côté et al. 2013, List 2014), but again it looks like the methods are not used by historical linguists.

One of the main reasons why more 'traditional' historical linguists do not seem to be impressed by the recent methodological advances is that recent computational approaches mostly focus on a historical tree (or network) as the central result. What is mostly missing is the actual decisions on cognacy, sound correspondences, and reconstruction for all individual words. Further, although many modern methods provide reasonably good results, there still remain many details that are not captured correctly. Although some computation method might be, say, 80% correct (impressive by computational standards), the historical linguist will argue that the missing 20% are mostly the more difficult and more interesting cases.

The vision for a more cooperative future is clearly that the computational methods should provide results that are more easily interpretable for historical linguists and allow for manual correction and enhancement. In this talk I will present various ongoing efforts of pipelines of historical reconstruction that combine computational methods with manual intervention (cf. Steiner et al. 2011).

References

Bouchard-Côté, Alexandre, David Hall, Thomas L Griffiths & Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *PNAS* 110(11). 4224-4229.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard & Quentin D Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957-960.

Kroeber, A L & C Douglas Chrétien. 1937. Quantitative Classification of Indo-European Languages. *Language* 13(2). 83-103.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. (Dissertations in Language and Cognition). Düsseldorf: Düsseldorf University Press.

Steiner, Lydia, Peter F Stadler & Michael Cysouw. 2011. A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change* 1(1). 89-127.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4). 452-463.

More and better regression analyses: what they can do for us and how

Stefan Th. Gries

University of California, Santa Barbara

This talk is essentially a plea for more and better regression modeling in linguistics. On the one hand, there is still a large body of work that does not yet use regression methods and, to some extent, pays a huge price for using older/simpler techniques when more powerful regression methods have been available for quite some time. On the other hand, some areas of linguistics, in particular corpus, psycho-, and sociolinguistics, have seen more applications of regression modeling but even in those one often just finds fairly 'standard' applications of (generalized) linear (mixed-effects) modeling that do not utilize all that comprehensive regression modeling has to offer. In this talk, I will essentially discuss a range of applications of statistical methods, showing in each case how a regression approach in general or a specific aspect of a particular regression approach leads to better statistical analyses; the examples will involve applications from learner corpus research, first language acquisition, alternation studies in English varieties research, and others.

Part II.

Section papers

Extraction of Estonian particle verbs from text corpus using statistical methods

Eleri Aedmaa

University of Tartu

Multiword expressions (MWEs) are problematic phenomena in natural language processing tasks (e.g. Sag *et al*. 2002). From semantic point of view, a multiword expression can be more or less opaque with respect to the meaning of their constituents (e.g. Bott & Schulte im Walde 2014). The current study focuses on one type of MWE – particle verbs. In order to distinguish the variation in extracting different types of Estonian particle verbs, lexical association measures (AMs) are compared.

An Estonian particle verb consists of a verb and a particle. According to Rätsep (1978) the verb-particle combination can be compositional or idiomatic. The components of compositional particle verbs are understood with their literal meaning, but the meaning of an idiomatic particle verb cannot be inferred from the literal meanings of its verb and particle, so it is idiosynctratic. Estonian lacks a study of distinction of particle verbs, so I tried to divide particle verbs into two groups – idiomatic and compositional. This is complex task because the list of particle verbs is not closed and often a single particle verb can have features of both idiomatic and compositional type. For instance, in example (1) particle verb *ette nägema* is of compositional type, but in example (2) *ette nägema* has features of the idiomatic type.

(1) Udu tõttu ei **näe** autojuht kaugele **ette**.
Fog due not see driver far ahead.
'Due to the fog driver doesn't see far ahead.'

(2) Ta ei **näinud** probleemi **ette**.
She not see problem before.
'She didn't foresee the problem.'

It is well-known fact that nearly all frequent words have multiple senses (e.g. Lewandowsky, Dunn, Kirsner 2014), and frequent Estonian particle verbs make no exception. This also adds complexity to the current task. Therefore, three groups of particle verbs are formed: idiomatic, compositional, and idiomatic and compositional (particle verbs that have features of both types).

In order to compare results with the previous work (Aedmaa 2014), the same AMs and data are used in this study. I evaluate following methods: t-test, mutual information (MI),

chi-square measure, log-likelihood function, minimum sensitivity (MS), and co-occurrence frequency of a verb and a verbal particle in one clause. Study is based on the newspaper part of Estonian Reference Corpus[1], which is morphologically analyzed and disambiguated, and annotated with clause boundaries. The list of particle verbs I study is the list of particle verbs presented in the Explanatory Dictionary of Estonian.[2]

I tested the hypothesis that t-test and frequency (as the best AMs in previous study (Aedmaa 2014)) perform better than others in extraction particle verbs which have features of both types. Also, I prove the hypothesis that there is difference in extraction of different type of particle verbs: MI works better for extraction of compositional particle verbs than idiomatic particle verbs. In addition I demonstrate how the results change as the number of candidate pairs increases.

References

Aedmaa, Eleri 2014. "Statistical methods for Estonian particle verb extraction from text corpus". *Proceedings of the ESSLLI 2014 Workshop: Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations*, 17–22.

Bott, Stefan, Sabine Schulte im Walde 2014. "Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality". *Proceedings of the 9th Conference on Language Resources and Evaluation, Reykjavik, Iceland.*

Lewandowsky, Stephan, John C Dunn, Kim Kirsner 2014. *Implicit memory: Theoretical issues*. Psychology Press.

Rätsep, Huno 1978. *Eesti keele lihtlausete tüübid*. Tallinn: Valgus.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger 2002. "Multiword expressions: A pain in the neck for NLP". *Computational Linguistics and Intelligent Text Processing*, 1–15. Springer.

[1] http://www.cl.ut.ee/korpused/segakorpus/index.php

[2] http://www.eki.ee/dict/ekss/

New methods for causal inference in the language sciences

Damian Blasi

Max Planck Institute for Mathematics in the Sciences

A well established doctrine of XX century statistics is that the different species of correlational analyses are not informative with respect to the actual underlying causes or mechanisms operating behind the data under study, and that statistical analyses alone are simple an ancillary tool that need to be complemented with experiments or theory-driven reasoning (Ladd, Roberts and Dediu 2015). Mistaking correlations for causes produced a host of putative relations between variables that are likely to be spurious—as for instance in the tongue-in-cheek correlation between number of traffic accidents and linguistic diversity (Roberts and Winters 2013).

However, this methodological situation is problematic. There is a rich number of problems in the language sciences of which we have no direct, ethical or accessible way of performing experiments or where our theoretical understanding is not mature enough to produce robust predictions. Some of these problems include the spatial and temporal distribution of typological variables, the relation between verbal behaviours and rare cases of aphasia, and the entangled heap of psycholinguistic indices that are massively correlated with each other.

Fortunately, the last decades witnessed an increased effort towards the development of causal models of observational data (Pearl 2000, Mooij et al. 2014). These models might or might not depend on classic correlations, but they aim to detect not only the space of all potential associations between variables but only those mediated by a reasonable causal logic. As an illustration: given three variables A, B and C and the sequential causal model A$\rightarrow$ B $\rightarrow$ C (where the $\rightarrow$ symbol stands for "causes") it will be reasonable to ask that A does not provide any information about C once B is known. Such constraints have proven to be useful beyond the mere assessment of causal relations, for instance for the task of defining measures of causal influence and for the elicitation of hidden structure in the data.

In this presentation I will illustrate the application of this family of methods using a large database of lexical variables from English words (Blasi, Roberts and Maathuis, in prep.). Beyond a number of interesting findings relevant for psycholinguistics, I will focus on highlighting the differences in reasoning, implementation and computation of causal analyses in contrast to correlational analyses.

References

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015). "Correlational studies in typological and

historical linguistics". *Annual Review of Linguistics*, 1, 221-241.

Roberts, S. G., & Winters, J. (2013). "Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits". *PLOS ONE*, 8(8): e70902.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Cambridge: MIT press.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2014). "Distinguishing cause from effect using observational data: methods and benchmarks". *arXiv preprint*, arXiv:1412.3773.

Blasi, D. E, Roberts. S. G., & Maathuis, M. (in preparation) *Causal relations in the lexicon*.

Investigating grammtical space in a parallel corpus

Östen Dahl

Stockholm University

This paper presents an on-going project where a massive parallel corpus consisting of Bible translations into approximately 1200 languages is used to study the structure of what we call "grammatical space". Grammatical space can be said to be one step more abstract than the more well-known notion of semantic space as displayed in "semantic maps". In a semantic map, a specific meaning or function of an expression is represented as a point. The total set of meanings or functions of an expression or a category will thus constitute a region in semantic space. By contrast, a grammatical item in a language will correspond to a point in grammatical space, with more closely related items being less distant to another. The empirical study of grammatical space rests on the general assumption that items with a similar semantics or pragmatics will have similar distributions in text. By comparing the distribution of grammatical items in parallel corpora, it is possible to establish cross-linguistic types of such items, which will be represented as clusters in grammatical space. Although grammatical space must be seen as having a large number of dimensions, it is often possible to use techniques such as multi-dimensional scaling to represent regions of grammatical space graphically and thus obtain a view of the internal structures of and relationships between such clusters.

So far, our attempts to apply this methodology has focused on grammatical domains such as tense-aspect and negation, but we hope to be able to extend it to other phenomena such as grammatical gender. An ongoing dissertation project aims at the creation of a system for aligning massive parallel corpora at the lexical level without previous knowledge of the languages; this will open up new possibilities for a more precise analysis of the texts. On the other hand, we have seen that even a coarser approach where the distribution of an item as defined as the set of bible verses in which it occurs is often sufficient to classify it and study its relationships to other items. So far, it has been possible to obtain a robust picture of the cross-linguistic variation within the tense-aspect category of perfects. This and other examples of the methodology will be presented in the paper.

Using topic modeling to detect and quantify semantic change

Haim Dubossarsky, Uri Shalit, Eitan Grossman, and Daphna Weinshall

The Hebrew University of Jerusalem

Today's 'dynamic duo' of big data and modern computational tools is changing the field of historical linguistics. These tools allow the large-scale analysis of entire corpora, providing quantitative measures for age-old questions. The goal of this paper is to evaluate two hypotheses: (1) that frequency interacts with change in word meaning (Bybee, 2006), and (2) that different word classes (POS) change at different rates (Sagi, 2010).

We use Latent Dirichlet Allocation (LDA; Blei & Lafferty, 2007), originally developed for the classification of documents according to their latent topics, to analyze changes in word meaning throughout a historical corpus. LDA assumes that each document is comprised of a mixture of a number of topics, and that similar documents have similar topic distributions. The model learns the topic distribution for each document, and hence captures its 'meaning.'

We create pseudo-documents for a large sample of words from a historical corpus in English. Each pseudo-document combines the contexts in which a given word occurs, and produces a mixture of topics that captures that word's meaning. Crucially, meaning change is reflected in changes in this topic distribution (TD) at different times, with greater changes in TD reflecting greater change in meaning, and vice versa.

We then test the possibility that such an approach can detect change in word meaning over time. We used the Corpus of Late Modern English Texts (CLMET, 1710-1920, 34 million words(, which was originally divided into three sub-corpora, and extracted 6,000 words-of-interest, which were the most frequent words in the full corpora. For a given word-of-interest ('ring'), we retrieved all the sentences in which it appeared for each of the historical sub-corpora separately, and constructed a pseudo-document that represented the contexts of occurrence for that particular word, thus creating a pseudo-document for each word at each time period. LDA model was then trained on the pseudo-documents, generating topic distributions for each one. Evaluating each word's change in meaning was done through computing the Hellinger distance of its TD between two time periods.

The correlations between the words' log frequencies and their meaning change scores were computed (Table 1). The cosine distances of a standard term-vector model of the same pseudo-documents were computed and correlated with the words' log frequencies to serve as control condition. The negative correlations (*all p's* $<$ *.001 permutation tests*) suggest that frequent words show less change, and vice versa.

LDA		term-vector model	
1st – 2nd time period	2nd – 3rd time period	1st – 2nd time period	2nd – 3rd time period
-0.40	**-0.54**	-0.15	-0.27

Table 1. Pearson correlation coefficients between words' log frequencies and meaning change scores for the two models (LDA and term-vector) at two time period transitions. Best results are bolded.

	LDA		term-vector model	
POS	1st – 2nd time period	2nd – 3rd time period	1st – 2nd time period	2nd – 3rd time period
Adjective	0.070	0.108	0.229	0.210
Noun	0.063	0.102	0.234	0.211
Adverb	0.049	0.085	0.233	0.213
Verb	0.041	0.092	0.224	0.208

Table 2. Averages of the LDA and term-vector model meaning change scores at two time period transitions, grouped by their POS.

Table 2 depicts averages of meaning change for four POS-tag groups, showing that different POS change at different rates. Overall, the largest changes are for adjectives, followed by nouns, adverbs, and verbs. Importantly, the control condition does not show such pattern, and differ drastically from the LDA results. The results support the use of LDA as a tool for representing synchronic meaning and detecting diachronic change. They also corroborate both the inhibiting nature of word frequency, and the significant interaction between a word's change in meaning over time and its POS assignment.

References

Blei, D. M., & Lafferty, J. D. (2007). *Correction: A correlated topic model of Science*, 17–35. 10.1214/07-AOAS136

Bybee, J. (2006). *Frequency of Use and the Organization of Language* (p. 375). Oxford University Press. Retrieved from http://books.google.co.il/books?id=W20t_5AXeaYC

Sagi, E. (2010). "Nouns are more stable than verbs: Patterns of semantic change in 19th century english". *32nd Annual Conference of the Cognitive Science Society*. Portland, OR.

Deviant diachrony: Exploring new methods for analyzing language change

Jason Grafmiller

KU Leuven

We present a novel technique for analyzing change in syntactic variation within a probabilistic framework by adapting the deviation analysis of Gries and Deshors' (2014) MuPDAR (Multifactorial Prediction and Deviation Analysis with Regression) method to the investigation of diachronic data from native speakers. While traditional variationist analyses of diachronic syntactic variation (e.g. Hinrichs and Szmrecsanyi 2007; Grimm and Bresnan 2009; Wolk et al. 2013) have focused on aggregate trends in historical corpora using standard regression-with-interaction models, our approach takes a more fine-grained, outcome-centered perspective on syntactic variation in diachrony. We use multivariate statistical techniques, namely multilevel logistic regression, to investigate how the probability of a constructional variant in a specific context, e.g. *hand me the book* vs. *hand the book to me*, varies across speakers from different time periods. In essence, we ask, "Given the same grammatical choice in the same context, how would the choice(s) of speakers from one time have differed from the choice(s) of speakers at a later time?"

The innovation in the present study is that we explore how speakers' usage at earlier time periods deviates from those of later speakers in not only the cases where the speakers from different times made (or would have made) different choices, but also in those instances where they (would have) made the *same* choices. We fit regression models to data from two (or more) distinct time periods, which generate separate synchronic probabilistic grammars derived from observations at those time slices. The models/grammars from different times are then used to predict construction probabilities on the same dataset, and by comparing the changes in probability from earlier to later model(s) for each observation, we explore how the usage of specific tokens in specific contexts has changed over time.

We evaluate the method with test cases involving previous studies of recent changes in the English genitive and dative alternations (Hinrichs and Szmrecsanyi 2007; Grimm and Bresnan 2009), using data from the Brown family of corpora (Brown, Frown, LOB, and F-LOB). We show that not only does the method provide results consistent with traditional analyses, it also provides greater resolution for discerning subtle linguistic and cultural shifts. For example, we find that while the use of collective possessors in the *s*-genitive construction (*the board's approval*) has increased over time, the kinds of collective entities US and UK speakers tend

to use in this construction differs noticeably. UK speakers not only show a greater tendency to refer to places as collective entities (*North Korea's contention*), but their use of place-as-collective nouns in the *s*-genitive—relative to that of Americans—has increased substantially over time. We find a similar, though less pronounced, pattern with collective recipients in the dative alternation. Patterns such as these provide probative information for further exploration of broader stylistic changes within and across varieties. The value of this technique is thus two-fold: it offers a confirmatory method for testing hypotheses comparable to traditional multivariate techniques, while at the same greatly facilitating exploratory qualitative research by providing researchers a quantitatively robust method for homing in on the most relevant/important subsets of their data.

References

Gries, S. T. and S. C. Deshors (2014). "Using regressions to explore deviations between corpus data and a standard/target: Two suggestions". *Corpora* 9(1), 109–136.

Grimm, S. and J. Bresnan (2009). "Spatiotemporal variation in the dative alternation: A study of four corpora of British and American English". In *Grammar & Corpora 2009*, Mannheim, Germany. September.

Hinrichs, L. and B. Szmrecsányi (2007). "Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora". *English Language and Linguistics* 11, 437–474.

Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsányi (2013). "Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change". *Diachronica* 30, 382–419.

Clause-initial adverbials of time in Finnish and Russian: a quantitative approach

Juho Härme

University of Tampere

In Finnish and Russian, as, supposedly, in the majority of languages, adverbials, including adverbials of time, tend to have a variety of possible locations in a clause. My presentation focuses on the clause-initial position, which, according to traditional descriptions of Russian and Finnish grammars, seems to be among the most typical ones in both languages. In addition, the use and the functions of this adverbial position are, at least superficially, quite similar. However, quantitative comparison of Finnish and Russian seems to suggest that there is a major difference between the studied languages in the frequency of the clause-initial position. Is this really the case and what does the possible difference in frequency imply about the difference in the functions of the clause-initial position in these languages on a more general level?

The study uses two corpora of literary texts, ParFin (my subcorpus consisting of Finnish fiction from 1976–2010) and ParRus (my subcorpus consisting of Russian fiction from 1970–1995). Both are actually collections of aligned parallel texts, which makes it possible also to look at the presumed difference in the use of the clause-initial position in the light of translations. The total size (including the translations) of the subcorpora are 1170338 tokens (ParFin) and 1212031 tokens (ParRus). For the purposes of this study, the corpora are syntactically annotated using dependency parsers (for Finnish, the TDT dependency parser[1] is used; for Russian, the dependency parser by Nivre & Sharoff is used[2].

I will narrow the scope of the studied adverbials of time to a group of expressions I call the *time measuring* words. The group includes calendaric words (i.e. words like second, hour, day, year) and words expressing days of week, names of months and times of day. This builds up to a reasonable group of words to be searched in the corpora.

To collect the data for the quantitative analysis, a parallel concordance search is conducted on every lemma categorized as a time measuring expression. Utilizing the syntactic annotations, the retrieved concordances are then further analyzed to

1. take into account only the occurrences where the lemma is actually used as (a part of) an adverbial of time

[1] http://turkunlp.github.io/Finnish-dep-parser/
[2] http://corpus.leeds.ac.uk/mocky/

2. separate the clause-initial adverbials from the non-clause-initial ones.

Preliminary results based on a smaller, manually annotated set of Finnish and Russian SV-clauses suggest that approximately 40,8% of Russian time-measuring expressions are located clause-initially, whereas for Finnish the number is 24,3%. The first aim of the study is to statistically confirm or reject these results by using the larger, automatically annotated corpora described above and by taking into account all possible clause types. Secondly, my goal is to find out, what motivates the possible differences between the studied languages. For this purpose, I take advantage of the parallel nature of the corpora and investigate the translations of clauses with a clause-initial adverbial of time. Thirdly, this study aims to test the syntactic annotations of the parallel corpora in use.

New evidence from linguistic phylogenetics supports phyletic gradualism

Eric W. Holman[1] and Søren Wichmann[2]

[1]University of California, Los Angeles
[2]Max Planck Institute for Evolutionary Anthropology & Kazan Federal University

Since the early 1970s, biologists have debated whether evolution is punctuated by speciation events with bursts of cladogenetic changes, or whether evolution tends to be of a more gradual, anagenetic nature (cf. [1] for a recent contribution to the debate). A similar discussion among linguists has only barely begun, the present study being the second to address the issue of punctuated equilibrium in the evolution of language families. The differing results of this and the previous study suggest that there is also room for controversy over this issue in linguistics.

In the previous study, Atkinson et al. [2] constructed phylogenetic trees for the Bantu, Indo-European, and Austronesian language families from published matrices of cognate judgments in basic vocabulary. For each language they counted the inferred lexical changes along the path from the root of the tree, along with the number of nodes along that path. A positive correlation between the number of changes and the number of nodes was attributed to increased changes caused by branching events.

The present analyses apply different methods to a much larger dataset, and show no systematic effects of punctuational change. We compare sister groups, defined as the descendents of two branches from the same ancestral node in the phylogeny. The number of branching nodes within each sister group is inferred from the number of extant languages in the group, given that more branching events are necessary to produce more languages. Sister groups are also compared with respect to lexical change. If the sister group with more languages shows more change than the sister group with fewer languages, the comparison is scored as positive for punctuation; and if the larger sister group shows less change than the smaller one, the comparison is scored as negative.

In this analysis lexical change is defined not in terms of cognate judgments but rather by a computerized measure of similarity between pairs of wordlists in the ASJP database [3], which consists of 40-item basic vocabulary lists in standard notation from about 62% of the world's languages. Phylogenies and language counts are from the classifications in Glottolog [4] and Ethnologue [5], which include all the known languages in each of the world's language families. Sister-group tests on all families with at least 20 languages reveal no evidence for punctuational evolution. Further analyses were carried out to verify the power of the sister-

group test to identify punctuated equilibrium when it is known to occur.

References

1. Pennell MW, Harmon LJ, Uyeda JC. 2014 “Is there room for punctuated equilibrium in macroevolution?” *Trends Ecol. Evol.* 29, 23–32. http://dx.doi.org/10.1016/j.tree.2013.07.004

2. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008 “Languages evolve in punctuational bursts”. *Science* 319, 588. (doi: 10.1126/science.1149683)

3. Wichmann S, Müller A, Wett A, Velupillai V, Bischoffberger J, Brown CH, Holman EW, Sauppe S, Molochieva Z, Brown P, Hammarström H, Belyaev O, List J-M, Bakker D, Egorov D, Urban M, Mailhammer R, Carrizo A, Dryer MS, Korovina E, Beck D, Geyer H, Epps P, Grant A, Valenzuela P. 2013 *The ASJP Database* (version 16). http://asjp.clld.org.

4. Hammarström H, Forkel R, Haspelmath M, Nordhoff S. 2014 *Glottolog 2.3*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org.

5. Lewis MP, Simons GF, Fennig CD (eds.). 2014 *Ethnologue: Languages of the world*, 17th ed. Dallas, TX: SIL International. http://www.ethnologue.com.

Incremental syntactic prediction in the comprehension of Swedish

Thomas Hörberg

Stockholm university

Comprehenders need to incrementally integrate incoming input with previously processed material. Constraint-based and probabilistic theories of language understanding hold that comprehenders do this by drawing on implicit knowledge about the statistics of the language signal, as observed in their previous experience. I test this prediction against the processing of grammatical relations in Swedish transitive sentences, combining corpus-based modeling and a self-paced reading experiment.

Grammatical relations are often assumed to express role-semantic (such as Actor and Undergoer) and discourse-related (e.g., topic and focus) functions that are encoded on the basis of a systematic interplay between morphosyntactic (e.g., case and word order), semantic / referential (e.g., animacy and definiteness) and verb semantic (e.g., volitionality and sentience) information. Constraint-based and probabilistic theories predict that these information types serve as cues in the process of assigning functions to the argument NPs during language comprehension. The weighting, interplay and availability of these cues vary across languages but do so in systematic ways. For example, languages with fixed word orders tend to have less morphological marking of grammatical relations than languages with less rigid word order restrictions. The morphological marking of grammatical relations is also in many languages restricted to NP arguments which are non-prototypical or marked in terms of semantic or referential properties, given their functions (overt case marking of objects is, e.g., restricted to personal pronouns in English and Swedish). I first assess how these factors affect constituent order (i.e. the order of grammatical relations) in a corpus of Swedish and then test whether comprehenders use the statistical information contained in these cues.

Corpus study The distribution of SVO and OVS orders *conditional* on semantic / referential (e.g., animacy and givenness), morphosyntactic (e.g., case) and verb semantic (e.g. volitionality) information was calculated on the basis of 16552 transitive sentences, extracted from a syntactically annotated corpus of Swedish. Three separate mixed logistic regression models were fit to derive the incremental predictions that a simulated comprehender with experience in Swedish would have after seeing the sentence up to and including the first NP (model 1), the verb (model 2), or the second NP (model 3). The regression models provide separate estimates of the *objective* probability of SVO vs. OVS word order at each point in the sentence. This information was used to design stimuli for a self-paced reading experiment

to test whether comprehenders draw on this objectively present information in the input.

Self-paced reading experiment 45 participants read transitive sentences that varied with respect to word order (SVO vs. OVS), NP1 animacy (animate vs. inanimate) and verb class (volitional vs. experiencer). By-region reading times were well-described by the region-by-region shifts in the probability of SVO vs. OVS word order, calculated as the relative entropy. For example, reading times in the NP2 region observed in locally ambiguous, object-initial sentences were mitigated when the animacy of NP1 and its interaction with the verb class bias towards an object-initial word order, as predicted by the constraint-based and probabilistic theories.

A Quantitative Study of the Japanese Particle *-ga*

Masako Hoye

University of Rhode Island

It has been widely assumed that the Japanese particle *-ga* is a "subject marker" in the literature. Particularly representative is Masayoshi Shibatani who defines the Japanese particle-ga as follows: "The particle ga marks the subject of both independent and dependent clauses in Modern Japanese. In this regard it is comparable to the nominative case in European languages" (1990: 347). Shibatani further writes that "the subjects of both transitive and intransitive clauses are marked by the particle *-ga*" (1990: 258). The definition of 'subject', according to Shibatani, is "a syntactic category resulting from the generalization of an agent over other semantic roles" (1991: 103). Further, the archetypical subject, Shibatani states, is an agentive participant {A} of a transitive clause, from which one of the traditional definitions of the subject as an agent/actor obtains (1991: 101). Thus, Shibatani clearly defines the Japanese particle ga as follows: 1) its primary function is to mark the subject of a clause; 2) it marks the subjects of both transitive and intransitive clauses; 3) the 'subject' is semantically an "agent/actor"; and 4) the most "archetypical subject" represents a transitive clause whose subject is semantically an "agent". Shibatani's definition of the particle *-ga* described and listed above is the most dominant and most widely accepted view by a majority of Japanese linguists. The purpose of this paper is to investigate to what extent this so-called Japanese subject marker ga fits its definition in discourse Japanese. Through the quantitative analysis of 6255 predicates that appear in natural discourse data, the following statements can be made: 1) the occurrence of ga is actually infrequent (11%); 2) 85% of ga appears in the S role, instead of the {A} role; 3) the appearance of ga is strongly associated with certain intransitive, stative predicates, most notably "intransitive pairs" (20%); 4) 82% of ga-marked NPs are semantically "non-agentive"; 5) "intransitive pairs", especially, never allow an "agentive" interpretation for their NP-*ga* (0%); 6) and even among the "agentive NP- ga", 78% of them appear inside embedded clauses or relative clauses. Among present day Japanese, in conversation, however, these tokens, which show ga as a subject marker inside either an embedded clause or a relative clause, represent merely 1.5% of the total number of predicates in the data set examined in this study (94/6255). Further, the fact that ga functions as a subject marker in the independent clauses is even rarer. Only 27 tokens out of 6255 predicates in such sentences can be found in the data. This indicates that agentive NP-*ga* appearing in the independent clause, which supposedly represents the "prototypical subject" accounts for merely **0.4%**. What this analysis demonstrates is that ga as a subject marker is at most only one of the minor functions of the Japanese particle-*ga* in present day Japanese in conversation.

Heimat versus *kotimaa* – a cross-linguistic corpus-based pilot study of written German and Finnish

Claudia Jeltsch

University of Helsinki

When comparing languages it is especially interesting to compare those ones that are not related to each other as in the case of Finnish and German. And it is even more interesting to see how languages deal with untranslatable words, such as in the case of German *Heimat*.

Heimat is impossible to translate, it is considered a "hotword" (Heringer 2007).

The Finnish sentence *Hänellä ei ole kotimaata* can be translated *Er/Sie hat keine Heimat*, but also *Er/Sie hat kein Heimatland* – referring to slightly different concepts (the closest equivalent in English being: *homeland*).

Other possible uses of Heimat include: *Die Heimat der Menschheit liegt in Afrika* or *meine sprachliche Heimat. . . /Essheimat, Wohnheimat. . .* (the "*Dornseiff-Bedeutungsgruppen*" show the whole variety of how *Heimat* can be used in German).

In the following paper I present the first results of a corpus-related pilot study how Heimat is used in the after-war German language and how in comparison to that kotimaa is used in contemporary Finnish. The corpora used are DeReKo, the German Reference Corpus, the Leipzig Corpora Collection and the Korp-corpus of the Language Bank of Finland. Both DeReKo and Korp include similar source material, e.g. newspapers and literature, the Leipzig Corpora Collection only internet-based material — in both languages. Using both traditional and modern sources reflects the interest of the study: how contemporary users of German and Finnish utilize these words and what kind of place they have in their lexicon (this point is especially important since the research in question is part of a dissertation project that includes interviews with speakers of both Finnish and German language).

I will present the most prominent collocations of *Heimat* respective *kotimaa*. The comparison will also show how the different language types influence the collocations but also how different collocations are connected with different connotations and contexts. Here, I'm especially interested if the words are used in special semantic fields. Thus at a later point it can be compared if individuals with a Finnish-German background show the same approach to *Heimat* or *kotimaa* as the corpora show. The following table shows the results from DeReKo and the Language Bank of Finland:

The prominence of country names can be explained by the corpus of Korp: it includes a lot of speeches from the European parliament. The collocation in connection with *Verein* is

Collocations in semantic fields	DeReKo	Language Bank of Finland
1.	compound words with *Verein* (= association): Verkehrsverein(s) *Museumsverein* *Verschönerungsverein* *Kulturverein*	(*country names, e.g. Ruotsi (=* Sweden)
2.	*zurückkehren / zurückgekehrt* *verlassen* *finden / gefunden*	*kärsiä* (= to suffer) *palauttaminen* (= retrieval)
3.	alten/ neue	*oma* (= own)

particular for German and reflects that in German *Heimat* is connected with smaller local units (e.g. the village, city or region, but not the country in the first place). The above overview can also be seen as a reflection of both post-war German and Finnish history as I will elaborate in my presentation.

The TOST as a method of equivalence testing in linguistics

Tom Juzek[1] and Johannes Kizach[2]

[1]University of Oxford
[2]University of Aarhus

Introduction Classical analyses typically test for differences and their null hypotheses state that the compared samples come from the same population. If negative, the outcome is insufficient evidence to assume a difference between the samples; which is not, though, sufficient to assume equivalence (Altman and Bland, 1995). Linguistics heavily relies on classical tests (e.g. all 16 experimental talks at the LSA 2013 used classical tests). However, they are insufficient for many linguistic questions. Consider $RQ_{1\text{-}3}$ (p.2). Negative results for $RQ_{1\text{-}3}$ would probably go unreported. This disincentivises such research (Bakker, van Dijk, and Wikkerts, 2012) and the field might miss out. An equivalence test would be more suitable.

The TOST The TOST, attributed to Westlake (1976), is one of the most common equivalence tests (Richter and Richter, 2002). It performs two one-sided t-tests and the null hypotheses are ($H0_1$): the difference in means of the two samples is bigger than a pre-set boundary δ and ($H0_2$): the difference is smaller than -δ.

$$H0_1\text{: } \mu_1 - \mu_2 > \delta \qquad H0_2\text{: } \mu_1 - \mu_2 > -\delta$$

A positive outcome (rejecting both nulls) denotes *equivalence within the range* δ. The researcher sets δ based on her knowledge of previous research. However, this leaves room for subjectiveness (Clark, 2009). Hence, our goal is to find an objective way to set δ.

Data simulation The "right" δ value is the value that gives a positive test outcome (indicating equivalence) with statistical power at $1 - \alpha = 95\%$ and $1 - \beta = 80\%$. To observe how the desired δ-values behave for different data, we simulate a "two-samples-one-position" setting for various datasets (24 in total) over various Ns (3 to 50,000). In the simulations, we "TOSTed" random pairs of subsets from a dataset, over and over again. In total, we simulated 2.1×10^{12} data points.

Predicting and validating δ We found a relationship between observed δ (δ_{obs}; from our simulations) and the subsets' pooled standard deviation (s_p). This relationship is near-constant for N_p (pooled from each pair of subsamples) and we call its quotient τ (the *Tübingen Quotient*; τ comes from δ_{obs}, thus τ_{obs}; see f_1).

$$f_1\colon \tau_{obs} = s_p \div \delta_{obs} \qquad f_2\colon \tau_{pred} = (\sqrt{N_p}) \div 4.581 \qquad f_3\colon \delta_{pred} = s_p \div \tau_{pred}$$

Fig 1. shows τ_{obs} over increasing Ns_p. Curve-fitting τ_{obs} led to f_2, which *predicts* τ (τ_{pred}). f_2 and the 4.581 are our critical findings, because: *by reversing f_1 to f_2 can be used to objectively set δ* (δ_{pred}). In a validation phase, we then compared τ_{obs} to τ_{pred}. For large parts, they match within $\pm 0.1\%$ (Fig. 2). Further simulations indicate that our results also apply to non-linguistic data.

Conclusion In our view, the TOST equivalence test is a useful tool in a linguist's repertoire, allowing to investigate research questions that ask for equivalence. So far, the lack of instructions to objectively set δ might have been a barrier to use this test. The present work outlined such guidelines and we hope that they will help boost equivalence testing in linguistics.

References

Altman, D. G., Bland, J. M. (1995). "Absence of Evidence is Not Evidence of Absence". *British Medical Journal* 311, 485.

Bakker, M., van Dijk, A. & Wichterts, J. M. (2012) "The rules of the game called psychological science". *Perspectives on Psychological Science* 7, 543–554.

Clark, M. (2009). "Equivalence testing" [PowerPoint slides]. Retrieved 16 Dec 2013 from: www.http://www.unt.edu/rss/class/mike/5700/Equivalence%20testing.ppt

Richter, S. J., Richter, C. (2002). "A Method For Determining Equivalence In Industrial Applications". *Quality Engineering* 14 (3), 375–380.

Westlake, W. J. (1976). "Symmetric Confidence Intervals for Bioequivalence Trials". *Biometrics* 32, 741–744

Additional materials

$RQ_{1\text{-}3}$

RQ_1: Can highly experienced L2 learners attain a native-like level of language production?
RQ_2: At which age do teenagers typically reach adult-like reading times?
RQ_3: Are resumptive pronouns perceived as equally bad across modalities?

The datasets

Source: authors or colleagues (all 24 datasets). *Areas*: syntax (13), phonetics (8), psycholinguistics (3). *Units*: Likert-Scale data (13), normalised Likert-Scale data (4), Hz (4), ms (3). *Aggregation*: aggregated (18), non-aggregated (6). *Size of Datasets*: 42 to 152, mean = 85.79.

Graphs

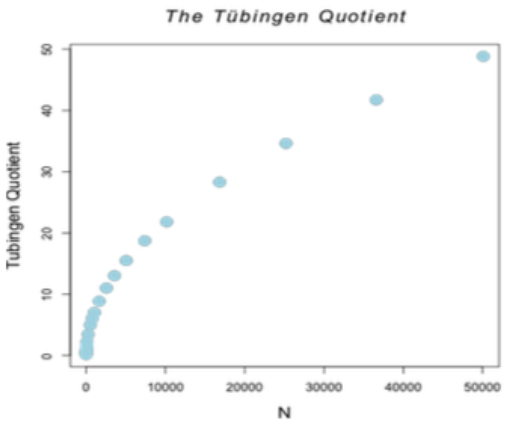


Fig. 1: τ_{obs} (y-axis) over increasing N_p (x-axis)

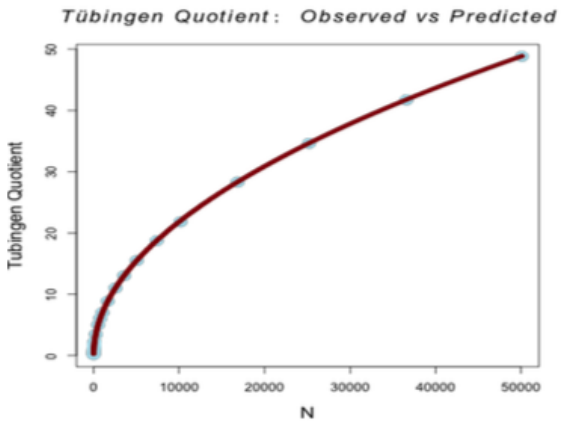


Fig. 2: τ_{obs} (y: blue) vs τ_{pred} (y: red) over increasing N_p (x)

Latent profile analysis (LPA) in L2 motivation research

Teija Kangasvieri

University of Jyväskylä

The aim of this paper is to show how latent profile analysis (LPA) can be used in L2 motivation research. LPA can be considered as a novel person-oriented statistical method in the field of L2 motivation research. In L2 motivation research, in the study of language learners' motivational profiles or types, cluster analysis has been used in a few studies (e.g. Csizér & Dörnyei 2005; Papi & Teimouri 2014). Cluster analysis resembles LPA, but according to statisticians LPA outperforms cluster analysis: LPA is model-based and thus allows comparison of different models based on the fit indexes it provides (see e.g. Pastor, Barron, Miller & Davis 2007). Therefore, it is of interest to explore how well LPA works as a statistical method in L2 motivation research.

More specifically, the target of this study was to find out if different kinds of L2 motivational profiles can be found among learners of different foreign languages (FLs) in Finnish comprehensive schools, and if these profiles differ depending on whether the FL is compulsory or optional. The target compulsory language in the study was English, and the optional languages were French, German, Russian, and Spanish. The data was gathered with a large-scale e-questionnaire, which included altogether thirteen different motivational scales on the language level, the learner level, and the learning situation level. A total of 1,206 answers were received from ninth-graders from altogether 33 Finnish schools. The data has been analyzed statistically with latent profile analysis (LPA).

The results of the LPA show that overall Finnish students appear to be quite motivated language learners, but they are clearly more motivated to study the compulsory language than the optional languages. Five different kind of motivational profiles can be found among the students: the most motivated, the average motivated with low anxiety, the average motivated, the least motivated, and students with high anxiety. Thus, LPA proved to work well as an analysis method in L2 motivation research. The pros and cons of the method (LPA), and the results of the analysis will be discussed in greater detail in the presentation.

References

Csizér, K. & Dörnyei, Z. 2005. "Language Learners' Motivational Profiles and Their Motivated Learning Behavior". *Language Learning* 55:4, December 2005, 613–659.

Papi, M. & Teimouri, Y. 2014. “Language Learner Motivational Types: A Cluster Analysis Study”. *Language Learning* 64 (3), 493–525.

Pastor, D. A., Barron, K. E., Miller, B.J. & Davis, S. L. 2007. “A latent profile analysis of college students’ achievement goal orientation”. *Contemporary Educational Psychology* 32 (2007), 8–47.

Complex networks-based approach to transcategoriality in the Bashkir language

Denis Kirjanov and Boris Orekhov

National Research University Higher School of Economics, Moscow

This study introduces a complex networks-based approach to quantifying transcategoriality. This approach is one of the most powerful ways of model description but it has been rarely used for linguistic needs (see [Sole et al. 2010], [Biemann et al. 2012]) and there are very few papers (e.g, [Brown, Hippisley 2012]) where it is applied to morphology.

The Bashkir language belongs to the Turkic languages which are considered to be agglutinative. Although the notion of agglutination was introduced in the 19th century, there is no generally accepted definition of an agglutinative language. Different features were supposed to be necessarily present in an agglutinative language (see, inter alia, [Haspelmath 2009]), however, there seems to be no correlation between them. Transcategoriality is sometimes considered as such a feature: "In linguistic typology it is accepted to associate the number of transcategorial morphemes with degree of language agglutination or analyticity (cf. Plungjan 2001)" [Plungjan 2011: 70]. In this study we discuss the data provided by our network and relevant for the notion of transcategoriality.

We conducted our study on Bashkir newspaper texts containing 5.8 mln tokens overall. They were annotated with the program "Bashmorph" [Orekhov 2014]. We built a network where nodes are affixes while edges represent cooccurrence of an affix pair. The network was built as weighted (based on the frequency of cooccurrences) and undirected. The network consists of 294 nodes and 3446 edges.

It turns out that several standard coefficients characterizing such a network help to quantify and describe certain characteristics of a language. In our case, most parameters correspond to transcategoriality. Namely, we discuss the meaning of assortativity coefficient, cliques number, maximal k-core, cluster coefficient and network density as well as some other data.

Thus the complex networks-based approach provides new data for describing transcategoriality and allows to formalize the the notion.

References

Biemann Ch., Roos S., Weihe K. (2012), *Quantifying semantics using complex network analysis*. Manuscript.

Brown D., Hippisley A. (2012), *Network morphology: A defaults-based theory of word structure*. CUP.

Haspelmath M. (2009), "An empirical test of the Agglutination Hypothesis", *Universals of language today*. (Studies in Natural Language and Linguistic Theory, 76.) Dordrecht, Springer, pp. 13–29.

Orekhov B. (2014), "Problems of morphologic annotation of Bashkir texts" [*Problemy morfologicheskoj razmetki bashkirskih tekstov*], *Proceedings of Kazan school on computational and cognitive linguistics TEL-2014* [Trudy Kazanskoj shkoly po komp'juternoj i kognitivnoj lingvistike TEL-2014], Kazan, Fen, pp. 135-140.

Plungjan V. (2001), "Agglutination and flection". M. Haspelmath et al. (eds.). *Language typology and language universals: An international handbook*. Berlin, Mouton de Gruyter, 2001, vol. 1, pp. 669-678.

Plungjan V. (2011), *Introduction to grammatical semantics: grammatical meanings and grammatical systems of the world's languages* [Vvedenie v grammaticheskuju semantiku. Grammaticheskie znachenija i grammaticheskie sistemy jazykov mira] Moscow, RSHU.

Sole R.V., Murtra B.C., Valverde S., Steels L. (2010), "Language networks: their functions, structure and evolution", *Complexity*, 15-6, pp. 20-26.

Latent profile analysis (LPA) in L2 motivation research

Jane Klavan[1] and Dagmar Divjak[2]

[1]University of Tartu
[2]University of Sheffield

Linguistic data is often described as "messy data" – it is complex and multivariate in nature with rampant intercorrelation among the explanatory variables. From a methodological perspective, this poses considerable challenges for the analyst. Statistical modelling is therefore an essential tool for a linguist working in the usage-based tradition. Reliance on data and statistics certainly gives us more confidence in our conclusions, but does it guarantee that our models are cognitively real(istic)?

Given that a multitude of phonological, morphological, syntactic, semantic, discourse-pragmatic, lectal and other parameters can influence the choice for one morpheme, word or construction over another, we need statistical modelling to determine the relative strength and importance of the various predictors. Until now, the most popular method for modelling the multivariate and seemingly probabilistic nature of linguistic knowledge has been logistic regression. But if we want our linguistics to be cognitively realistic, should we not consider using modelling techniques that are directly based on principles of human learning? Moreover, if interest is in modelling *human* knowledge, should we not compare our model's performance to that of native speakers of the language?

In our paper we will take up these and other pertinent questions regarding statistical modelling. One of the datasets we work with comes from present-day written Estonian. 900 occurrences of the adessive case and the adposition *peal* "on" were coded for 20 variables with 47 distinct variable categories. In our initial analysis we used binary logistic regression to predict the choice between the two alternative constructions. The regression model fitted to the data has a classification accuracy of 70%. In order to assess its performance, we compare the logistic regression model to a model arrived at using naive discriminative learning (Baayen 2010). Previous studies (Baayen 2011, Baayen et al. 2013, and Theijssen et al. 2013) have shown that, in general, logistic regression performs on par with other modelling techniques. Similarly to Divjak et al. (under review) we propose that in order to assess whether a statistical modelling technique yields a model that is cognitively more (or less) real(istic) we need to compare corpus-based models to native speakers. To this end, a series of experiments with native speakers was conducted.

In one of the experiments, the task of the native speakers was similar to that of the corpus-based classification model. 96 participants were presented with 30 attested sentences in which

the original construction was replaced with a blank. They were asked to choose which of the two constructions fits the context best. The mean number of "correct" choices for the participants was 22.6 (accuracy 75%, median 23, SD 2.5). Similarly to what Divjak et al. (under review) saw in their behavioral data, there was also considerable individual variation among the Estonian speakers (the scores ranged from 14 to 28). We analyse the errors made by the different models and compare those to errors made by subjects to establish which of the models shows the performance that is most similar to that of the subjects (cf Divjak et al. under review). Implications for methodology and theory will be discussed.

References

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova and Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. Russian Linguistics 37, 253-291.

Baayen, R. Harald. 2010. "Demythologizing the word frequency effect: A discriminative learning perspective". *The Mental Lexicon* 5, 436-461.

Baayen, R. Harald. 2011. "Corpus linguistics and naive discriminative learning". *Revista Brasileira de Linguística Aplicada* 11 (2): 295-328.

Divjak, Dagmar, Antti Arppe and Ewa Dabrowska. Under review. "Machine Meets Man: evaluating the psychological reality of corpus-based probabilistic models". *Cognitive Linguistics*.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen and Hans van Halteren. 2013. "Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation". *Corpus Linguistics and Linguistic Theory* 9(2): 227-262.

The use of multivariate statistical classification models for predicting constructional choice in Estonian dialectal data

Jane Klavan, Maarja-Liisa Pilvik, and Kristel Uiboaed

University of Tartu

A common presumption in usage-based linguistics is that the speakers' linguistic knowledge is probabilistic in nature. It has been shown that speakers have a richer knowledge of linguistic constructions than the knowledge captured by categorical judgements leads us to believe (Divjak & Arppe 2013, Bresnan 2007, Bresnan et al. 2007, Bresnan & Ford 2010, Szmrecsanyi 2013) In addition to the probabilistic nature of linguistic data, language use is also driven by multitude of factors. Speaker's choice between alternative forms is often influenced by semantic, syntactic, morphological, phonological, discourse-related, lectal, and other factors. The practical and methodological question is how can we capture this knowledge quantitatively. At the moment, multivariate statistical classification modeling seems to be the best tool available. The present paper continues this line of research and discusses the results of a multivariate corpus analysis of two near-synonymous constructions in Estonian. We take a usage-based and variationist perspective and focus on non-standardized, spoken spontaneous language. We look at the parallel use of the adessive case construction and the adposition *peal* 'on' construction in Estonian dialects.

The aims of the paper are twofold. We first evaluate how the model fitted to the dialect data performs in comparison to the model fitted to written language data. To this end a multivariate corpus analysis was carried out with 2,131 occurrences of the adessive case and the adposition peal 'on' in the Corpus of Estonian Dialects (CED 2015). The data were analysed using mixed-effects logistic regression. The minimal adequate model fitted to the written language includes four morphosyntactic and two semantic explanatory predictors and has a classification accuracy of 70% (Klavan 2012). We are interested in testing whether the same morphosyntactic and semantic predictors are also significant for predicting the choice in non-standard spoken language. We are furthermore interested to see whether the fit of the model can be significantly improved by including the geographical dimension in the model. It has been suggested that the use of analytic constructions (i.e. the adposition *peal* construction) is more characteristic of Southern Estonia, while the use of synthetic constructions (i.e. the adessive case construction) is more frequent in Northern Estonia.

The second goal of the paper is a methodological one – to discuss one of the ways how the performance of logistic regression models can be evaluated. In addition to the conventional model diagnostics, the goodness of fit can further be assessed by comparing it to models which are based on the same dataset, but arrived at using alternative techniques, such as, for example, the 'tree & forest' method, naive discriminative learning, Bayesian networks and memory-based learning. Similarly to Baayen et al. (2013) and Theijssen et al. (2013) we conclude that the different models generally provide converging results. The added bonus is that the methods come with complementary advantages. It is therefore concluded that for a best possible result, methodological pluralism is called for, i.e. applying different methodological tools to one and the same linguistic data.

References

.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova and Tore Nesset. 2013. "Making choices in Russian: Pros and cons of statistical methods for rival forms". *Russian Linguistics* 37, 253–291.

Bresnan, Joan. 2007. "Is syntactic knowledge probabilistic? Experiments with the English dative alternation". In Sam Featherston and Wolfgang Sternefeld (eds). *Roots: Linguistics in Search of Its Evidential Base*, 77–96. Berlin: Mouton de Gruyter.

Bresnan, Joan and Marilyn Ford. 2010. "Predicting syntax: processing dative constructions in American and Australian varieties of English". *Language* 86 (1), 186–213.

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

CED 2015. *Corpus of Estonian Dialects*, http://www.murre.ut.ee/mkweb/

Divjak, Dagmar and Antti Arppe. 2013. "Extracting prototypes from exemplars. What can corpus data tell us about concept representation?" *Cognitive Linguistics* 24 (2),221-274.

Klavan, Jane 2012. *Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy*. Tartu: University of Tartu Press.

Szmrecsanyi, Benedikt. 2013. "Diachronic Probabilistic Grammar". *English Language and Linguistics* 1(3): 41–68.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen and Hans van Halteren. 2013. "Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation". *Corpus Linguistics and Linguistic Theory* 9(2), 227–262.

Treebanks and historical linguistics: a quantitative study of morphosyntactic realignment in early medieval Italian Latin

Timo Korkiakangas

University of Helsinki

A researcher of ancient languages finds it difficult to speak of 'big data'. Treebanking has made it possible to speak of 'rich data', instead. This paper studies quantitatively the semantic and syntactic factors that influence the case form of subject (nominative or accusative) in early medieval documentary Latin. The study is based on Late Latin Charter Treebank (LLCT), a 200,000-word corpus of Tuscan private documents from between AD 714–869 (Korkiakangas & Passarotti 2011). LLCT is provided with lemmatic, morphological, and syntactic annotation (syntactic function and dependency relation) plus a light semantic annotation layer. T wo layers of diplomatic and sociolinguistic annotation have been merged to the linguistic annotation layers.

Cennamo (2009) and Rovai (2012) suggest that, in Late Latin, one can identify traces of a transitory change from nominative/accusative to active/inactive alignment (and back to nominative/accusative system in Romance languages). The six-case system of Classical Latin was reduced, through a two- case stage, to the neutral declension of the Romance languages. The nominative/accusative contrast was (re)semanticized so that the nominative came to encode all the Agent-like arguments and the accusative all the Patient-like arguments. Consequently, the accusative encroached on the traditional nominative domains. These 'extended accusatives' are found in substandard texts, such as charters:

(1) medieta-te de ipsa terrola possede-at ipsa sancta De-i
half-ACC(OBJ) of the plot possess-3SG the holy God-GEN
uertu-te
church-ACC(SBJ)
'this holy church of God possesses one half of the plot' (CDL 90, AD 747, Lucca)

As the first treebank of Late Latin, LLCT enables systematic empirical analysis of case marking system, which has been thus far studied based on about 150 haphazard sentences that happen to have accusative-form subjects. By applying quantitative methods, Latin linguistics is confronted with completely new questions: which kind of variable distributions represent an

on-going morphosyntactic realignment in the conservative and formulaic charter Latin? How are the variation patterns supposed to change in diachrony? In this paper, I seek to answer these methodological questions.

Although semantics was the driving force of the realignment, certain syntactic factors may have interfered in it. I assess the dependencies between the following variables by way of cross tabulation and chi-squared decision trees (CHAID) (Eddington 2010, Priiki 2014).

Dependent variable	Independent variable	Levels
case form of subject (nominative or accusative)	animacy/referentiality of subject	animate proper name, animate common name, inanimate
	verb type	transitive, unergative, unaccusative, passive
	subject position with respect to verb	linear word position (negative or positive integer)
case form of attribute (nominative or accusative)	attribute position with respect to its head within subject NP	distance in dependency levels (positive integer)

The above independent variables seem to correlate significantly with the dependent variables. The percentage distributions of the levels of each independent variable imply the following:

- The accusative subjects prefer low-animacy nouns and often occur with unaccusative verbs.
- The attributes located at the end of attribute chains have slightly higher accusative rates than the attributes closer to the head of the subject NP.
- The immediate preverbal clausal position of subjects correlates with high retention of nominative.

References

Adams, J. N. 2013. *Social variation and the Latin language*. Cambridge: CUP.

Cennamo, M. 2001. "L'extended accusative e le nozioni di voce e relazione grammatica nel latino tardo e medievale", Viparelli, V. (ed.). *Ricerche linguistiche tra antico e moderno*. Napoli: Liguori, 3–27.

Cennamo, M. 2009. "Argument structure and alignment variations and changes in Late Latin" *The role of semantic, pragmatic, and discourse factors in the development of case*. Ed. by J. Barðdal, S. L. Chelliah. Studies in language companion series 108, 307–346.

CDL = *Codice Diplomatico Longobardo* 1–2. A cura di Luigi Schiaparelli. Roma 1929–1933.

Eddington, D. 2010. "A comparison of two tools for analyzing linguistic data: logistic regression and decision trees", *Italian Journal of Linguistics* 22:2, 265–286.

Korkiakangas, T. & Passarotti, M. 2011. "Challenges in Annotating Medieval Latin Charters", *Proceedings of the ACRH Workshop*, Heidelberg, January 5 , 2012. *Journal of Language Technology and Computational Linguistics* (JLCL) 26:2, 2011, 103–114.

Korkiakangas, T. & Lassila, M. 2013. 'Abbreviations, fragmentary words, formulaic language: treebanking mediaeval charter material', in Proceedings of The 3 Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3), Sofia, 2013, 61–72.

La Fauci, N. 1997. *Per una teoria grammaticale del mutamento morfosintattico. Dal latino verso il romanzo*. Pisa: ETS.

Ledgeway, A. 2012. *From Latin to Romance. Morphosyntactic typology and change*. Oxford: OUP.

Priiki, K. 2014. 'Kaakkois-Satakunnan henkilöviitteiset *hän*, *se*, *tää* ja *toi* subjekteina', *Sananjalka* 56, 86–107.

Rovai, F. 2005. "L'estensione dell'accusativo in latino tardo e medievale", *Archivio Glottologico Italiano* 90, 54–89.

Rovai, F. 2012. *Sistemi di codifica argomentale. Tipologia ed evoluzione*. Pisa: Pacini.

Sabatini, F. 1965. "Esigenze di realismo e dislocazione morfologica in testi preromanzi", *Rivista di Cultura Classica e Medievale* 7, 972–998.

Sornicola, R. 2008. "Syntactic conditioning of case marking loss: a long term factor between Latin and Romance?", M. van Acker, R. van Deyck, M. van Uytfanghe (eds.). *Latin écrit – roman oral?: de la dichotomisation à la continuité*. Corpus Christianorum 5, Brepols: Turnhout.

Generalization about automatically extracted Russian collocations

Daria Kormacheva, Mikhail Kopotev, and Lidia Pivovarova

University of Helsinki

Our project aims to implement the model able to process multiword expressions of different nature on an equal basis. It has been systematically evaluated against Russian data and is applicable to various languages. The model is corpus-driven; it compares the strength of various possible relations between the tokens in a given n-gram and searches for the "underlying cause" that binds the words together: whether it is lexical, grammatical, or a combination of both. Taking syntactic, semantic and lexical properties equally into account, we follow the ideas that were first formulated by J. Sinclair, A. Goldberg, and Ch. Fillmore and developed recently by S. Gries and A Stefanowitsch (2004), Huston (2007) to mention just a few.

In order to define the most stable features of the given query, rather than apply a single multiword-extraction technique, we propose a cascade of procedures that lean on and deepen the results of the previous steps. The system takes as an input any 2-4-gram, where one position is a variable that is looked for, with possible grammatical constraints. The aim is to find the most stable lexical and/or grammatical features of the variables that appear in this query. The normalized Kullback-Leibler divergence is used to obtain a ranked list, where grammatical categories, tokens, and lemmas are equally treated. Then, having specified the most highly ranked categories, we define the particular values for them. At this step grammatical categories are processed separately from tokens and lemmas, because of the significant difference in their distributional properties; grammatical categories can take quite limited number of values — e.g., four for gender, three for number, dozen for case — while tokens and lemmas may have thousands variations each. For grammatical categories standard *frequency ratio* is used, while collocations are extracted using a more sophisticated version of this measure, that is the *refined weighted frequency ratio*, which has been chosen after the comparison of six statistical measures that our algorithm can calculate so far.

As the result, our model provides a multi-level description of a query pattern. For example, the following results are predicted for the Russian query [*bez* 'without'+ Noun].

1. This pattern exemplifies the grammatically restricted colligation [*bez* 'without'+ Noun.**GEN**];
2. it represents the semantic preferences of a stable construction [*bez* 'without' + Noun.GEN '**part of clothes**'], where lexical variables are interchangeable but belong to the same se-

mantic class (Cf. Eng. *sleight of* [*hand*/*mouth*/*mind*]). In this case, even if collocations as such may be rare, prediction of the whole semantic class is possible.

3. One collocation — *bez galstuka* ‘without a neck-tie’ — is frequently used being a fixed expression. It can be used not only literally, but also idiomatically meaning ‘informal’ (Cf. *vstreča bez galstuka* ‘shirtsleeve meeting’). This is the ultimate case of lexically stable multiword expressions -– such as Eng. *lo and behold* — where no generalization is possible at all. We assume that formally there is no border between the last two types and an idiomatic collocation is nothing but construction with one lexical variable.

Pupillometry as a window to real time processing of morphologically complex verbs

Aki-Juhani Kyröläinen,[1] Vincent Porretta,[2] and Juhani Järvikivi[2]

[1]University of Turku
[2]University of Alberta

In recent years, eye-tracking has been used to investigate the real time processing of morphologically complex words. This method offers a rich source of information, specifically numerous durational measures through time (e.g.. Kuperman et al., 2009; Pollatsek & Hyönä, 2006). In addition, eye-tracking opens the possibility to record changes in pupil dilation in real time (Laeng et al., 2012 for an overview). Pupillometry has been used to investigate, for example, the intensity of mental activity (Beatty, 1982), retrieval of memories (e.g., Papesh et al., 2012; Goldinger & Papesh, 2012), emotions and frequency-effects (Kuchinke et al., 2007). In this study, we examine the possible contribution of pupil dilation to the investigation of morpho-semantic processing, contrasting it with fixation durations. Specifically, we investigate the processing of Russian reflexive verbs (*-sja*) which represent a salient category associated with changes in argument structure (*serdit*' 'anger' versus *serdit'*sja 'become angry').

26 native Russian participants performed a lexical decision task with 160 tetramorphemic reflexive verbs <*serd-i-t'-sja*>, while their eye movements were recorded. In addition, each participant provided a semantic similarity estimation between the reflexive and the base verb on a five-point scale. To inspect the effects of morpho-semantic information of these verbs, mean semantic similarity was calculated and nine frequency- and dispersion-based measures were extracted from the Russian National Corpus. The distributional measures were submitted to principle component analysis to remove collinearity resulting in three components. PC1 relates to the changes in the overall distribution of the morphological construction <*serd-i-t'-sja*>. PC2 contrasts the distributional difference between the base <*serd-i-t*'> and the reflexive verb <*serd-i-t'-sja*> whereas the difference between the root <-*serd*-> and the reflexive verb are captured by PC3. Finally, participant age (M = 28.8 and SD = 5.5) was included in the analysis as a proxy for accumulation of experience across the lifespan (see Bybee, 2010; Ramscar et al., 2014).

Previous pupillometric studies have primarily relied on comparing differences in peak dilation. Here, the pupil response was modeled as a time series beginning at the onset of the

stimulus and continuing for 2000 ms. The analysis utilized generalized additive mixed-effects modeling (Wood, 2014) which allowed us to model the inherent non-linearity and account for any autocorrelation present in these data. In this manner, we were able to compare the time course of the pupil dilation to fixation durations.

The model indicated that the processing of these verbs was driven by the morphological construction frequency (PC1) and the relative distributional differences between the morphological constituents (PC2 and PC3). Furthermore, semantic similarity influenced pupil dilations early in time, whereas it did not influence first fixation duration. Finally, there was no effect of age in any of the fixation durations, even though it significantly influenced pupil dilation throughout the time course. This effect, along with capturing early effects not seen using fixation- related measures, suggests that pupillometry uniquely contributes to our understanding of morpho- semantic processing. The results are discussed in terms of probabilistic approaches to morphology.

References

Beatty, J. (1982). "Task-evoked pupillary responses, processing load, and the structure of processing resources". *Psychological Bulletin*, 2(91), 276–292.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Goldinger, S. D. & Papesh, M. H. (2012). "Pupil dilation reflects the creation and retrieval of memories". *Current Directions in Psychological Science*, 2(21), 90–95.

Kuchinke, L., Võ, M. L.-H.,Hofmann, M. & Jacobs, A. M. (2007). "Pupillary responses during lexical decisions vary with word frequency but not emotional valence". *International Journal of Psychophysiology*, 2(65), 132–140.

Kuperman, V., Schreuder, R., Bertram, R. & Baayen, R. H. (2009). "Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing". *Journal of Experimental Psychology: Human Perception and Performance*, 31(35), 876–895.

Laeng, B., Sirois, S. & Gredebäck, G. (2012). "Pupillometry: A window to the preconscious?" *Perspectives on Psychological Science*, 1(7), 18–27.

Papesh, M. H., Goldinger, S. D. & Hout, M. C. (2012). "Memory strength and specificity revealed by pupillometry". *International Journal of Psychophysiology*, 1(83), 56–64.

Pollatsek, A. & Hyönä, J. (2006). "Processing of morphemically complex words in context: What can be learned from eye movements". In Anders, S. (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 275–298). Hove: Psychology Press.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P. & Baayen, R. H. (2014). "The myth of cognitive decline: Non-linear dynamics of lifelong learning". *Topics in Cognitive Science*, 1(6), 5–42.

Wood, S. N. (2014). mgcv: *Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*. http://cran.r-project.org/web/packages/mgcv/index.html.

Lessons learned from compiling a cognate corpus

Antti Leino,[1] Kaj Syrjänen,[1] Terhi Honkola,[2] Jyri Lehtinen,[3] and Maija Luoma[1]

[1]University of Tampere
[2]University of Turku
[3]University of Helsinki

A series of research projects, starting in 2009, has resulted in a cognate corpus that covers 313 meanings and 26 languages across the Uralic language family, including a reconstruction of Proto-Uralic. The meanings in the data set include the 100 and 200 word Swadesh lists, as well as the Leipzig-Jakarta list of basic vocabulary. In addition to these, there are two basic word lists tailored for Uralic languages, as well as a list of less basic words derived from WOLD ranks 401–500.

In editing the data set for publication, one of the early decisions was to aim at compatibility with the Indo-European lexical cognacy database, IELex. Nevertheless, as the origins of the two projects were different, the database format has had to be extended slightly. The main reason for this is that the Uralic data contains not only strict cognates but also correlate relations, which include connections between words based on borrowing as well as based on common descent from a protolanguage. Currently the database format is being exended further, to allow storing typological data in addition to lexical cognates.

This presentation will give an overview of the design decisions and pilot studies that led to the current choice of word lists, as well as the process of editing the Uralic data set to be compatible with the Indo-European one.

Applying population genetic methodology to study linguistic variation among the Finnish dialects

Jenni Leppänen,[1] Terhi Honkola,[2] Jyri Lehtinen,[1] Perttu Seppä,[1] Kaj Syrjänen,[3] and Outi Vesakoski [2]

[1]University of Helsinki
[2]University of Turku
[3]University of Tampere

Both languages and biological species vary in time and space (Croft 2008). Genetic variation within species is commonly structured into populations which may further diverge to different species. Analogously linguistic variation is structured into geographical dialects which may later form closely related languages. Recently this analogy between species and languages has been utilized by growing number of studies that have analyzed linguistic data with quantitative methods and in a framework applied from biology, concentrating mostly on linguistic divergence among languages (i.e. linguistic macroevolution, e.g. Bouckaert *et al.* 2012; Dunn *et al.* 2013). We have initiated a new approach of paralleling populations and dialects in a microevolutionary framework and investigate linguistic variation within a language, among dialects where the process of diversification actually originates. We use the methods of population genetics that offer powerful tools to study variation also within languages. However, applicability of these tools has to be tested and demonstrated. Most population genetic analyses start with defining populations—a request that often puzzles also population geneticists. Here we concentrate to disentangle this first crucial step when applying population genetics to linguistics by studying the variation among the Finnish dialects. As our data we use the historical Dialect Atlas of Finnish collected in years 1920–1930 (Kettunen 1940a, b; Embleton & Wheeler 1997, 2000). We tested different clustering methods (such as the software Structure (Pritchard *et al.* 2000) and BAPS (Corander *et al.* 2003)) with the Dialect Atlas and compared the outcomes with each other and to traditional linguistic studies of Finnish Dialects. The clustering methods differ in their assumptions of the data (Excoffier & Heckel 2006; Guillot *et al.* 2009; Kalinowski 2011), which is why their comparison is fruitful with the language data. First, in the light of the theory of both population genetics and linguistics, we compare the special features of language data with the genetic data, and investigate which kind of genetic data type (e.g. microsatellite or amplified fragment length polymorphism data) is most suitable analogy for the dialect data. Second, given the differences and similarities between the language and genetic data, we evaluate the assumptions of different models and software on

the language data. Finally, we discuss the numerous applications that population genetics may offer to linguistics, such as measuring "flow of linguistic characteristics" and differentiation among dialects, and give future perspectives on the topic.

References

Corander J, Waldmann P, Sillanpaa MJ (2003) "Bayesian analysis of genetic differentiation between populations". *Genetics* 163, 367-374.

Croft W (2008) "Evolutionary Linguistics". *Annual Review of Anthropology* 37, 219-234.

Embleton S, Wheeler ES (1997) "Finnish dialect atlas for quantitative studies". *Journal of Quantitative Linguistics* 4, 99-102.

Embleton S, Wheeler ES (2000) "Computerized dialect atlas of Finnish: Dealing with ambiguity". *Journal of Quantitative Linguistics* 7, 227-231.

Excoffier L, Heckel G (2006) "Computer programs for population genetics data analysis: a survival guide". *Nature Reviews Genetics* 7, 745-758.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) "Statistical methods in spatial genetics". *Molecular Ecology* 18, 4734-4756.

Kalinowski ST (2011) "The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure". *Heredity* 106, 625-632.

Kettunen L (1940a) *Suomen Murteet III A. Murrekartasto*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Kettunen L (1940b) *Suomen Murteet III B. Selityksiä Murrekartastoon*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Pritchard JK, Stephens M, Donnelly P (2000) "Inference of population structure using multilocus genotype data". *Genetics* 155, 945-959.

Testing iconicity: A quantitative study of causative constructions based on a parallel corpus

Natalia Levshina

Université catholique de Louvain

Aims

Form-function isomorphism has been a prominent topic in functionally oriented typology. In this study we focus on iconicity of cohesion, i.e. correlation between the conceptual integration of events and their formal integration (e.g. Haiman 1983). The object of our study is causative constructions, such as *cause X to die*, *make X dead* and *kill* in English, which differ with regard to the degree of formal integration of cause and effect. To the best of our knowledge, the evidence in favour of such isomorphism has been based primarily on isolated, often self-constructed examples; quantitative empirical studies are still lacking. The present study aims to fill this gap. We use corpus data from a sample of ten languages that represent different language families (according to the Ethnologue classification): Finnish, French, Hebrew, Indonesian, Japanese, Korean, Mandarin Chinese, Thai, Turkish and Vietnamese, and employ cutting-edge statistical methods (namely, ordinal regression with mixed effects) in order to put the iconicity hypothesis to test.

Data

For this study we use a self-compiled parallel corpus of film subtitles in ten above-mentioned languages plus English. Subtitles are chosen because they represent informal language and contain highly diverse causative situations in comparison with other massively parallel corpora. First, we extract approximately 250 exemplars of different causative events (e.g. 'X causes Y to die' or 'X causes Y to break') from the English subtitles. Next, we check how these events are verbalized in each of the ten languages, and classify the language-specific causative expressions into several constructional types: analytic, resultative, morphological and lexical (cf. Comrie 1981), which are defined as comparative concepts (Haspelmath 2010). The English exemplars are also coded for more than a dozen semantic variables that have been mentioned in typological literature (intentionality of causation, control of the causee, etc.), among which Dixon's (2000) parameters of semantic variation between more and less compact causatives.

Statistical analyses and preliminary results

We use a mixed-effect ordinal logistic regression with the constructional types as the response, the semantic variables as fixed effects and the multilingual exemplars and individual languages as random intercepts and slopes. Since the semantic parameters are highly intercorrelated, we also use Multiple Correspondence Analysis as a dimensionality- reduction technique, which enables us to simplify the model. The preliminary results suggest that the iconicity hypothesis in general holds: the less cohesive the form, the less direct causation and the more autonomous the Causee. However, there is some language-specific variation in the effect of the semantic predictors, as can be seen from the random slopes in the mixed-effect model. We will also discuss the observed form-function correlation from the point of view of an alternative account based on the Principle of Economy and frequency effects, which has been recently developed by Haspelmath (e.g. 2008), and propose a unified model of form, frequency and function.

References

Comrie, B. (1981). *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Dixon, R. M. W. (2000). "A typology of causatives: form, syntax and meaning". In R. M. W. Dixon & A. Aikhenvald (Eds.), *Changing valency: Case studies in transitivity* (pp. 30–83). Cambridge: Cambridge University Press.

Haiman, J. (1983). "Iconic and economic motivation". *Language*, 59(4), 781–819.

Haspelmath, M. (2010). "Comparative concepts and descriptive categories in crosslinguistic studies". *Language*, 86(3), 663–687.

Haspelmath, M. (2008). "Frequency vs. iconicity in explaining grammatical asymmetries". *Cognitive Linguistics*, 19(1), 1–33.

Counting sheep and their tails: A quantitative approach to the interaction of the lexicon with grammatical number

Olga Lyashevskaya

National Research University Higher School of Economics

In this paper we use grammatical (inflectional) profiles which indicate the relative frequency distribution of the inflected forms of a word in a corpus (Newman 2008; Janda & Lyashevskaya 2011) to explore the lexical semantics of Russian nouns in its interaction with the grammatical category of number. Since the category of number is binary in Russian (singular forms VS plural forms), we define the grammatical profile for number (NumGP) as the ratio of plural forms to all forms.[1] Uncountable nouns are expected to have either 0% plural forms (singularia tantum) or 100% plural forms (pluralia tantum), while countable nouns are distributed normally between 0% and 100% with mean = 23%, cf. *lapa* 'paw', 57.8%, and *khvost* 'tail', 13.6% plural forms. Figure 1 shows the distribution of NumGPs for a sample of Russian nouns.

We argue that the ratio of plural forms can serve as a descriptive measure that helps to reveal certain functional and referential frames in lexical semantics. Firstly, we discuss a set of nouns with a very small ratio of plural forms which tend to be referentially unique, cf. *mama* 'mom', *batja* 'daddy', *njanja* 'nurse' among names of person,*vkhod*, *vykhod* 'entrance', *veranda* 'veranda', *zanaves* 'curtain', etc. among names that describe interior. These words have been mostly overlooked in the literature on number, cf. Lyashevskaya 2004:27.

Secondly, in a number of case studies we focus on rather compact taxonomic groups (e.g. body parts, kinship terms, transport, emotions, etc) and on differences in functional frames that reflect how objects/events a noun refers to are typically used and typically observed. For example, there is no surprise in the fact that the names of paired body parts have higher ratio of plural forms than the names of non-paired body parts (cf. *brov'* 'eyebrow', 87%, and *lob* 'forehead', 5%). However, the more passive the body part is, the lower the ratio of plural forms in the corresponding noun (cf. *podborodok* 'chin', 1%, and *lico* 'face', 13%). Table 1 shows uneven ratio of plural forms within the group of transport names. It occurs to be

[1] The data are drawn from the disambiguated version of the Russian National Corpus, 1950-present. Case distinction are not taken into account. The syntactically motivated uses of singular forms in paucal constructions (e.g. dva khvosta 'two tails', SyntSG) are calculated separately.

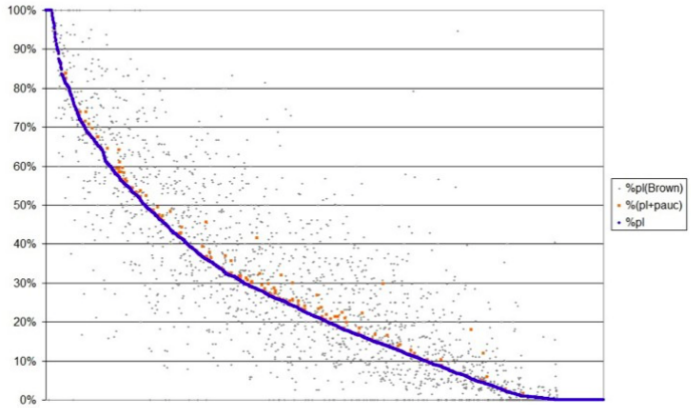


Fig. 1. NumGPs (%pl) of 2100 common nouns sorted in descending order. Supplementary points: %(pl+pauc) adds to counts the syntactically motivated uses of singular forms, %pl(Brown) are the data from Brown et al. 2013.

important, in addition to other functional properties, whether the observer is situated inside or outside a vehicle. If the observer is looking at the situation from outside, then it is possible to grasp a plural set of moving vehicles. Given that, the nouns that refer to planes, tractors, tanks have higher ration of plural forms than the names for taxi, tram, train, and boat.

sani	'sledges'	100 %
tank	'tank'	75 %
traktor	'tractor'	32 %
samolet	'plane'	29 %
poezd	'train'	20 %
lodka	'boat'	19 %
taksi	'taxi'	7 %
metro	'metro'	1 %

We provide top-10 constructions associated with either plural or singular forms in order to handle a gradual change in countability of concrete and abstract nouns (Lyashevskaya 2004; Pazelskaya 2006). Even though there is no clear borderline between plural and singular uses we can see a number of semantically transparent effects such as a correlation between attributive constructions and larger NumGPs. Furthermore, we discuss why previous attempts to induce semantically relevant hierarchies of nouns from the ratio of plural forms have been reported to fail (cf. Greenberg 1974; Brown et al. 2013). We conclude that since frequent nouns for the most part are semantically ambiguous, their profiles cannot reveal any straightforward effects for large lexical classes.

References

Brown, Dunstan Patrick, Greville Corbett, Sebastian Fedden, Andrew Hippisley, Paul Marriott. (2014). "Grammatical typology and frequency analysis: Number availability and

number use". *Journal of Language Modeling*, Vol. 1, No. 2, pp. 227-241.

Greenberg, Joseph H. (1974/1990). "The relation of frequency to semantic feature in a case language (Russian)", in Denning, K., S. Kemmer (eds.), *On Language, Selected Writings of Joseph H. Greenberg*, Stanford, pp. 207-226.

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

Lyashevskaya, Olga. (2004). *Semantika russkogo chisla* [*Semantics of number in Russian*]. Moscow.

Newman, John. (2008). "Aiming low in linguistics: Low-level generalizations in corpus-based research". *Proceedings of the 11th International Symposium on Chinese Languages and Linguistics*, National Chiao Tung University, Hsinchu, Taiwan, May 24, 2008. http://www.ualberta.ca/~johnnewm/Aiming%20Low.pdf.

Pazelskaya, Anna. (2006). *Nasledovanie glagol'nykh kategorij imenami situacij: na materiale russcogo jazyka* [*Verbal categories inherited in event nouns: a case of Russian*]. PhD thesis. Moscow, ms.

Expression of spatial relations in the Ngen language in typological perspective

Anna Maloletnyaya

Higher School of Economics, Moscow

The current work is related to the typological researches in the field of locative expressions. This study focuses on means, which are used to express spatial reference. I am going to discuss the basic semantic domains of these expressions. The basis of this research is special expression in Ngen language, Southeastern Mande language spoken in Côte-d'Ivoire. The system of locative expression in Ngen can be divided into three groups: postpositions, body part terms and cardinal points. The research into differences and similarities of the systems of locative terms in the languages of the world is going to form the typological constituency of the project.

The proposed research is going to follow the directions described in the existing publications concerning locative expressions. Some examples were taken from the grammar books. It is well known that space can be analyzed from two different points of view: deictic and non-deictic. Here, we will focus on the latter. This direction was determined by Charles Fillmore's Santa Cruz lectures on deixis [Fillmore 1971]. There are other possible ways to investigate this problem: analyze the form and the meaning, as it was realized in the article by Soteria Svorou "On the Evolutionary Paths of Locative Expressions" [Svorou 1986]. Also the core of this study is the data from the Ngen language, which I collected in the linguistic field trip. My study carried out in this trip lays the foundation of the first grammatical description of the Ngen language. The data for this research was collected in the linguistics expredition in Côte-d'Ivoire (Ivory Coast) in August-September 2014. To understand the basis of Ngen morphology, I used Dahl's questionnaire [Dahl 1985], as well as Kibrik's questionnaire [Кирик 1972]. I am going to make an attempt to combine and compare the ways of the expressing spatial relations across a substantial number of languages.

The major problems in this area are connected with complicated linguistic questions, namely, whether some lexical means (body parts as adpositions) can become grammaticalised. Actually, Soteria Svorou introduced the scale representing stages of graduations from nouns to bound affixes: from lexical means to grammatical [Svorou 1986]. Thus, I will try to analyse data regarding to some additional parameters such as the type of marking (head-marking or dependent-marking [Nichols 1986]), the position in genitive constructions, word order. I attempt to reconstruct the system of locative expressions, find some connections between dif-

ferent parameters.

References

Кибрик, А.Е. (1972). Методика полевых исследовании. Издательства Московского университета.

Dahl, Ö. (1985). *Tense and Aspect Systems*. Basil Blackwell.

Fillmore, C. J. (1971). *Santa Cruz lectures on deixis: space*. University of California, Berkeley.

Nichols, J. (1986). "Head-marking and Dependent-marking grammar". *Language* 62 no 1.

Nikitina, T.V. (2008). "Locative terms and spatial frames of reference in Wan."

Svorou, S. (1986). "On the Evolutionary Paths of Locative Expressions". *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*.

Quantifying the complexity of analogical paradigm changes in Murrinhpatha

John Mansfield and Rachel Nordlinger

University of Melbourne

Analogical extensions are generally assumed to bring greater regularity to morphological paradigms (e.g. McMahon, 1994). However this apparent truism has been established impressionistically, without any formal method for quantifying regularity versus irregularity in the paradigm. Recent investigations into morphological complexity provide such a method, in the quantification of *integrative complexity* (Ackerman & Malouf, 2013) – essentially, the preponderance of regular relationships between inflectional cells in a paradigm. In this paper we apply the integrative complexity measurement to analogical innovations in Murrinhpatha verb morphology, testing whether recently documented changes do indeed reduce integrative complexity, or in other words, bring greater regularity to the paradigm.

Murrinhpatha is a polysynthetic, non-Pama-Nyungan language of northern Australia, with a very large inflectional paradigm. In smaller inflectional paradigms, the quantification of regularity brought about analogical change might turn out to be a trivial formalisation. But in a large paradigm such as that of Murrinhpatha, the analogical extension of a pattern has greater potential to disrupt patterns that exist in other parts of the paradigm. Murrinhpatha verbs inflect according to 38 conjugation classes, each iterating over 46 cells marking person, number and tense. A change in conjugation Class 1 might improve the regularity of its patterning with Classes 2 and 3, but at the same time disrupt a pattern it shared with Class 4.

Table 1 shows an example of an innovation recorded in recent verb inflection elicitation. Sub-parts of three conjugation classes are shown here to illustrate some of the paradigmatic patterns at play.

The innovation appears in Class 34, 2sg Realis.nFut. This change makes the relationship between Class 34 1sg and 2sg cells match more closely the pattern found in Class 19. But on the other hand, Class 34 2sg and 3sg Realis.nFut previously matched the same pair of cells in Class 8, and this pattern has now been disrupted.

Our preliminary findings, using the conditional entropy measurements for integrative complexity (Ackerman & Blevins, 2009; Ackerman & Malouf, 2013), suggest that such changes do not clearly reduce complexity/irregularity, but instead result in improved integration in some parts of the paradigm, and weakened integration in others. We also find that the conditional entropy scores produced by the method are highly sensitive to coding choices made in

the segmentation of morphology, and the level of abstraction applied in analysing inflectional patterns.

		1sg	2sg	3sg
Class 34	Realis (non-future)	nga-m-am-	n-Ø-am- → **tha-m-am-**	Ø-m-am-
	Irrealis	nga-m-a-	tha-m-a-	ka-m-a-
Class 19	Realis (non-future)	nga-Ø-am-	tha-Ø-am-	da-Ø-am-
	Irrealis	nga-Ø-a-	tha-Ø-a-	ka-Ø-a-
Class 8	Realis (non-future)	Ø-m-am-	n-Ø-am-	Ø-m-am-
	Irrealis	Ø-m-a-	n-Ø-a-	Ø-m-a-

Table 1. Example of an analogical innovation

References

Ackerman, F., & Blevins, J. P. (2009). "Parts and wholes: Implicative patterns in inflectional paradigms". In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition*. Oxford: Oxford University Press.

Ackerman, F., & Malouf, R. (2013). "Morphological organization: The low conditional entropy conjecture". *Language*, 89(3), 429–464.

McMahon, A. M. S. (1994). *Understanding language change*. Cambridge: Cambridge University Press.

The effects of L3 motivation on L2 motivation – a moderated mediation analysis

Enikö Marton

University of Helsinki

This paper addresses the use of moderated mediational analysis (MacKinnon et al., 2013; Preacher et al., 2007) in research into L2 learning. After an overview of the statistical background of the method, results from an empirical study will be presented to illustrate the use of the method in practice. The focus of the empirical study is on how English as L3 affects the motivation to learn Swedish as L2 among Finnish-speaking secondary school students. The data was collected in Finnish language secondary schools in the Finnish capital region (N = 577). The conceptual model was analyzed by the means of the PROCESS macro (Hayes, 2013). The results indicated a paradoxical effect of English: among those who are more motivated to learn Swedish, greater interest in English increased even more the motivation to learn Swedish; however, among those who are less motivated to learn Swedish, interest in English decreased even more the motivation to learn Swedish. Both the use of the PROCESS macro and the visualization of the significant interaction effects with Excel will be demonstrated.

References

MacKinnon, D. P., Kisbu-Sakarya, Y., & Gottschall, A. C. (2013). "16 Developments in Mediation Analysis". *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*, 2, 338.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). "Addressing moderated mediation hypotheses: Theory, methods, and prescriptions". *Multivariate behavioral research*, 42(1), 185-227.

Quantitative language typology based on symmetry properties of syntactic structures

Gregory Martynenko and Yan Yadchenko

St. Petersburg State University

The paper relates to quantitative language typology studies which have their origin in works by J. H. Greenberg and L. Tèsniere. The authors propose an expansion of this approach based on the ideas of syntactic symmetry: the paper deals with symmetry properties of linearized verbal group trees. The research was made on the material of literary texts (fiction) in five European languages — Russian, Finnish, German, French and English.

Linearized verbal group tree is a syntactic structure having verb as a root (or the ancestor node of every node), while dependent nodes are located to the left or to the right from the root in particular structural and quantitative proportion. The goal of the given research is to reveal these proportions for each language in concern and to elaborate the system of diagnostic expansion of this approach based on the ideas of syntactic symmetry:

The research is based on counting occurrences of each type of language syntactic figures, their ranking according to their frequency and comparison of frequency distributions.

The table below shows symmetry proportions between left-branching and right-branching constructions.

Languages	Asymmetric		Symmetric	Left-/right-branching ratio
	Left-branching	Right-branching		
ENGLISH	22,8	22,6	54,6	1,01
GERMAN	28,6	20,6	50,8	1,39
FRENCH	29,6	18,4	52,0	1,61
FINNISH	14,6	35,0	50,4	0,42
RUSSIAN	45,8	23,8	30,4	1,92

The obtained data show that proportions of symmetric and asymmetric construction in English, German, French and Finnish are about the same value. Russian demonstrates here its syntactic originality.

What concerns ratio of two types of asymmetry, the maximal difference is observed between Russian and Finnish (the ratios of left-branching structures to right-branching structures are 1.92 and 0.477 correspondingly, i.e. a four-fold difference).

English takes here a central position. It is characterized by “symmetric asymmetry”. That means that the number of left-branching and right-branching structures are approximately equal. In general, English verbal group trees tend to mirror symmetry. From this point of view, French tends to the golden symmetry, as its ratio is not very different from the Phidias number (ϕ=1, 618).

Further, we consider rank distributions of these syntactic structures. We have got the following results:

1. For all the languages in question, the first rank has a symmetric structure with the unique left-node and the unique right-node. However, the percentage of this "high-end" structure is different between languages (Russian — 10,1%, German — 13,5%, French — 20,6%, English — 21,6%, and Finnish — 23,6%).

2. The number of diverse syntactic structures is minimal for German (10) and maximal for Russian (17).

3. The degree of rank distribution uniformity was measured according to the formula $R_{\text{mean}}{:}\frac{n}{2}$, where R_{mean} is the rank mean, n – the number of different syntactic structures. For the languages in question we have the following results: German — 0,544, Russian — 0,431, Finnish — 0,392, French — 0,362, English — 0,329.

The obtained results should be considered as preliminary. Further studies on the expanded range of text genres and on larger text samples are planned to be done.

Tracing Culture in Language Structures: Ecological Evidence for L1 Acquisition of Individualism

Matthias Meyer-Schwarzenberger

University of St. Gallen, Switzerland

This paper presents two scalar indices measuring the impact of cultural dispositions on the grammatical properties of natural languages.

Grammatical resources provided by different languages have occasionally been shown to correspond with different cultural priorities (e.g., Biber, 1995; Bickel, 1997; Bloom, 1981; Duranti, 1994; Fausey, Long, Inamori & Boroditsky, 2010; Mueller-Liu, 2009). Inspired by a descriptive method of data aggregation used in linguistic area research (cf. Haspelmath, 2001; van der Auwera, 1998), I argue that several such features, which may be uncorrelated or even conflicting in their linguistic nature, can accumulate in a given language to the extent that a cultural theme is predominant in the respective speech community.

The cultural focus of this study is the concept of individualism, which is recognized across academic disciplines as one of the most important dimensions of cultural variation. Previous research has associated a cultural emphasis on human individual agency with language features such as transitive constructions and obligatory subject pronouns (Fausey et al., 2010; Kashima & Kashima, 1998; Minkov, 2009). To attain a quantitative measure and a higher level of cross-linguistic generalization, I propose two indices measuring the degree to which individual agency is accentuated by the syntactical and morphological structures of different languages, respectively.

As for the syntax, I hypothesized that grammar features contributing to the distinction of a topical agent category—i.e., the 'subject' in European languages—emerge and persist in individualist societies more than in collectivist communities. The underlying assumption motivating this hypothesis is that L1 speakers acquiring a subject-prominent language are conditioned to decontextualize individuals more than the speakers of other languages. Consequently, they are comparatively more likely to adopt an individualistic mind-set and transmit the relevant grammar features to their own children in turn.

In light of this hypothesis, I searched the World Atlas of Language Structures (WALS) and Atlas of Pidgin and Creole Language Structures (APiCS) for indicators of subject prominence. I identified six constructions maintaining an explicit distinction between the subject and the predication to which the subject is subject even when one of these components is semantically empty or evident from the context of speech. Relevant features were dichotomized and

averaged for each language. The same procedure was applied to three WALS features on morphological agent marking.

The two resulting indices of Subject Prominence and Agent Morphology (see Table A1) are strongly and robustly correlated with survey-based measures of individualism at the ecological country level (e.g., Welzel's concept of Emancipative Values; cf. Figure 1). This finding indicates that contemporary individualism is largely due to cultural heritage, and cultural heritage is mirrored by language structures. In conjunction with recent experiments on linguistic priming (Fausey et al., 2010; Kühnen, Hannover, Pöhlmann & Roeder, 2013), my results suggest that L1 acquisition of grammar plays a key role in the collective transmission and preservation of cultural inclinations.

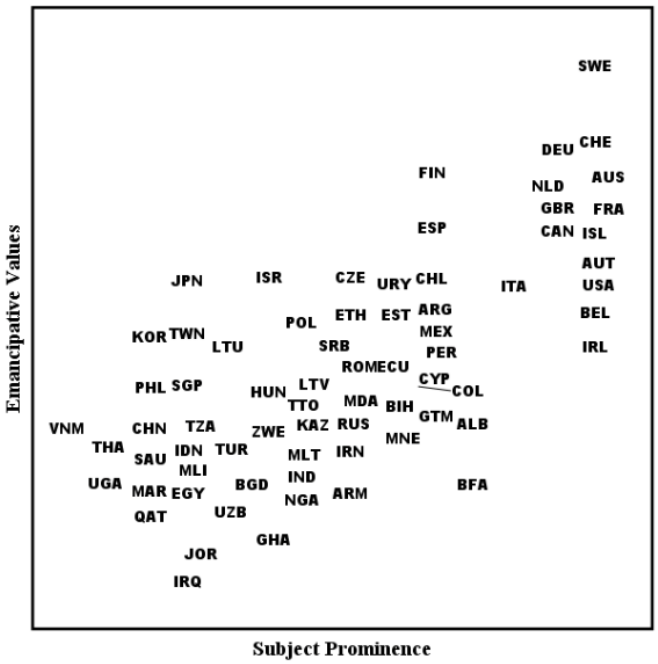


Figure 1 *The index of subject prominence predicts Emancipative Values, an operationalization of individualism proposed by Christian Welzel (2013).*

References

Biber, Douglas (1995). "Dimensions of Register Variation: A Cross-Linguistic Comparison". Cambridge: Cambridge University Press.

Bickel, Balthasar (1997). "Spatial operations in deixis, cognition, and culture: Where to orient

oneself in Belhare". In: Jan Nuyts and Eric Pederson (Eds.), *Language and Conceptualization* (pp. 46–83). Cambridge: Cambridge University Press.

Bloom, Alfred H. (1981). *The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West*. Hillsdale, NJ: Lawrence Erlbaum.

Duranti, Alessandro (1994). *From Grammar to Politics: Linguistic Anthropology in a Western Samoan Village*. Berkeley: University of California Press.

Fausey, Caitlin M., Bria L. Long, Aya Inamori, and Lera Boroditsky (2010). "Constructing agency: the role of language". *Frontiers in Cultural Psychology* 1, 1–11.

Haspelmath, Margin (2001). "The European linguistic area: Standard Average European". In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible (Eds.), *Sprachtypologie und sprachliche Universalien* (pp. 1492–1510). 2nd Vol. Berlin: Walter de Gruyter.

Kashima, Yoshihisa, and Emiko S. Kashima (1998). "Culture and Language: The Case of Cultural Dimensions and Personal Pronoun Use". *Journal of Cross-Cultural Psychology* 29(3), 461–

Kühnen, Ulrich, Bettina Hannover, Claudia Pöhlmann, and Ute-Regina Roeder (2013). "How Self-Construal Affects Dispositionalism in Attributions". *Social Cognition* 31(2), 237–259.

Minkov, Michael (2009). *Otrazhenia na culturata vyrhu srednovekovnia scandinavski i angliiski ezik i tehnite syvremenni variant* [Culture's impact on medieval Scandinavian languages and English and their modern variants]. Unpublished Dissertation, University St.Kliment Ohridski, Sofia.

Mueller-Liu, Patricia (2009). "Konzepte von Zeit, Distanz und Nähe—Kulturkontrastive Grammatik am Beispiel des Deutschen und Chinesischen". In: Lutz Götze, Patricia Mueller-Liu, and Salifou Traoré (Eds.), *Kulturkontrastive Grammatik — Konzepte und Methoden* (pp. 253– 307). Frankfurt: Peter Lang.

van der Auwera, Johan (1998). "Conclusion". In Johan van der Auwera (Ed.), *Adverbial constructions in the languages of Europe* (pp. 813–836). Berlin: Mouton de Gruyter.

Welzel, Christian (2013). *Freedom Rising: Human Empowerment and the Quest for Emancipation*. Cambridge: Cambridge University Press.

One million Hows, two million Wheres, and seven million Whys

Mikhail Mikhailov

University of Tampere

Researchers often complain that corpora do not yield enough examples of rare words and unusual grammatical constructions. And that conversely, for some words they get too much data: just a small corpus will contain thousands of occurrences of prepositions and conjunctions, and even common verbs and nouns can sometimes be extremely frequent. It is not very productive to manually check huge concordances. To study random examples is also problematic, because some important contexts may pass unnoticed.

A possible way out can be to study collocations. Collocations are understood here as words reoccurring within a specified word span before and after a given search word. The search application simply finds all such items, calculates the frequencies for each item in each position, and generates a summary table of those collocates with frequencies of co-occurrence higher than a lower limit specified by user. The collocator described here is a part of the program package for processing corpora that has been developed by me. Lists of collocations can be acquired with other applications as well, e.g. WordSmith Tools.

Collocations are not limited to types; they can be lexemes, parts of speech, or even semantic groups. The data that can be obtained will depend on the corpus annotation and the software used. The TACT package, for example, can generate collocation tables for a particular lemma (= a group of types represented by the same base form) and even collect lemmatized collocates (i.e. the collocates will be lemmas, not types).

Using collocation tables can help with the following:

- detecting syntactic patterns
- obtaining statistics on the use of words with several different meanings,
- disambiguation,
- improving annotation.

To give just one example, in the Multilingual Corpus of Legal Documents (MULCOLD, University of Tampere) the type *have* occurs 1845 times. But how many times is it used as an auxiliary verb? Although the corpus is lemmatized with the Connexor package, there are no special tags for auxiliary verbs, and so the tagging is of no help. A manual check, on the other hand, might take a few days. However, it is fairly easy to generate a collocation table for the

type *have* in a range of 1- 3 to the right (*have accepted*, *have * authority*, *have * * cancelled*, etc) and to calculate the sum of the frequencies for all the past participles (*created*, *decided*, etc). A figure of 1082 is produced in just a few minutes. This figure is not 100% accurate, but the margin of error should not be great. Using the table, it would not be difficult to draw up a list of patterns for the automated correction of the corpus annotation.

Here are a few more examples of using the same approach:

- disambiguating the English word *state*, which can be either a noun or a verb;
- disambiguating the Finnish noun *oikeus*, which has two meanings: 'justice, right' and 'court of law';
- finding common fixed phrases beginning with the Russian preposition *dlja* 'for'.

Using multivariate analysis to uncover evidence of cross-linguistic influence in learner corpora

Steve Pepper

University of Oslo

It is now 50 years since Seppo Mustonen of the University of Helsinki first demonstrated in print how the multivariate method of discriminant analysis could be applied to linguistic problems (Mustonen 1965), but the uptake of his ideas has been rather slow until recently.

This presentation describes the application of discriminant analysis to investigate cross-linguistic influence ('language transfer') in second language acquisition. Data from the Norwegian Second Language Corpus was analysed in order to address the following research questions:

1. Can the methods of data mining be used to identify the L1 background of learners of L2 Norwegian on the basis of their use of lexical features of the target language?
2. If so, what are the best predictors of L1 background?
3. Can those predictors be traced to cross-linguistic influence?

The source data consisted of Norwegian interlanguage texts written by 1,000 second language learners from ten different L1 backgrounds (German, English, Dutch, Spanish, Polish, Russian, Serbo-Croat, Albanian, Somali, Vietnamese). There was also a control corpus of 100 texts written by native speakers. Word frequencies computed from this data were analysed using ANOVA and discriminant analysis.

Discriminant analysis is defined by Klecka (1980) as "a statistical technique which allows the researcher to study the differences between two or more groups of objects with respect to several variables simultaneously." In the present study, the 'objects' under study were learner texts; the 'groups' were defined according to the authors' L1 backgrounds; and the 'variables' were the relative frequencies of the 50-60 most commonly occurring words in the texts. Based on this input, mathematical models were computed that proved capable of predicting the authors' L1 backgrounds with a statistically significant degree of accuracy. For example, in a test involving the five L1 groups German, English, Polish, Russian and Somali, the author's L1 was correctly predicted for 288 of the 500 texts, giving an overall success rate of 57.6%, compared to the 20% success rate of the null hypothesis. This shows that the discriminant

analysis had found a significant amount of L1 grouping structure in the data and the first research question was thus answered in the affirmative. Further analysis using the method of feature selection made it possible to determine exactly which lexical features contributed most to the model and thus constituted the best predictors of L1 background (research question 2).

Those predictors were subjected to additional tests in order to determine which L1 groups the various lexical features served to separate. The results both confirmed existing knowledge and revealed a number of hitherto unsuspected patterns. These were subjected to contrastive analysis (Gast 2012) in order to answer the third research question, and in most cases it proved possible to attribute the tendency in question to language transfer in one form or another. The presentation will briefly compare the results of this study with similar studies reported for English learner texts in Jarvis & Crossley (2012) and Jarvis et al. (2013).

References

Gast, Volker. 2012. "Contrastive Analysis". In Michael Byram & Adelheid Hu (eds.) *The Routledge Encyclopedia of Language Teaching and Learning*, 2nd Edition. London: Routledge.

Jarvis, Scott & Scott A. Crossley (eds.) 2012. "Approaching Language Transfer through Text Classification". *Explorations in the detection-based approach*. Bristol: Multilingual Matters.

Jarvis, Scott, Bestgen, Yves, & Pepper, Steve. 2013. "Maximizing classification accuracy in native language identification". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, pp. 111–118.

Klecka, William R. 1980. "Discriminant Analysis". *Quantitative Applications in the Social Sciences* 19. London: Sage Publications.

Mustonen, Seppo. 1965. "Multiple discriminant analysis in linguistic problems". *Statistical Methods in Linguistics* 4, 37-44.

Partitioning a closed set of meanings: How restrictive are the existing models?

Alexander Piperski

Russian State University for the Humanities / Russian Academy of National Economy, Moscow

Given a finite set of meanings, a language can employ various strategies to encode the members of this set. It may use a separate marker for each meaning, or it may prefer not to draw a distinction between some of the meanings, thus exhibiting a syncretism. Many models have been proposed to study syncretism patterns, the most popular of them being the semantic map approach (Haspelmath 2003). Implicational hierarchies can also serve a similar purpose; cf. the well-known color term hierarchy by Berlin and Kay (1969).

The most important question about these models is how restrictive they are. Potentially, we can partition a set of n elements in B_n various ways, where B_n is the n-th so-called Bell number (Bell 1938). B_n grows quite rapidly with the growth of n (B_1 = 1, B_2 = 2, B_3 = 5, B_4 =15,B_5 =52,B_6 =203, B_7 =877, B_8 =4140, B_9 =21147, B_{10} =115975,. . .).For instance, if there is a set of meanings M = {*a, b, c*}, it can be expressed in 5 ways using members of a set of words or morphemes W, the size of W ranging from 1 to 3:

1. $w_1 \leftrightarrow \{a, b, c\}$; 2. $w_1 \leftrightarrow \{a\}$, $w_2 \leftrightarrow \{b, c\}$; 3. $w_1 \leftrightarrow \{b\}$, $w_2 \leftrightarrow \{a, c\}$; 4. $w_1 \leftrightarrow \{c\}$, $w_2 \leftrightarrow \{a, b\}$; 5. $w_1 \leftrightarrow \{a\}$, $w_2 \leftrightarrow \{b\}$, $w_3 \leftrightarrow \{c\}$ $(w_1, w_2, w_3 \in \mathrm{W})$

However, for a larger n most of the theoretically possible $W \leftrightarrow M$ mappings are non-existing or at least quite improbable. This calls for constraining our models in a way that they would predict only existing or plausible $W \leftrightarrow M$ mappings. In other words, a semantic map where all vertices are connected to each other would describe any kind of syncretism, but would lack explanatory adequacy linguistics strives for.

Various restrictions have been imposed on semantic maps. For instance, one-dimensional (linear) maps are preferred over two-dimensional maps (planar graphs), which are in turn preferred over three-dimensional ones (Haspelmath 2003); cf. also Plank (1991) where one and two-dimensional semantic maps are proposed to describe case syncretism. The restrictions on implicational hierarchies have not been studied, but a hierarchy is by its nature more restrictive than a semantic map with the same number of vertices.

I claim that the restrictiveness of a model can be measured numerically. If a semantic map or an implicational hierarchy has n vertices, its restrictiveness R_n is equal to G_n / B_n, where G_n is the number of possible patterns under the conditions imposed on the model (such as one-dimensionality, two-dimensionality, etc.). In my talk, I am going to compare R_n for semantic maps with different geometric properties and implicational hierarchies.

References

Bell, Eric Temple. 1938. "The iterated exponential integers". *Annals of Mathematics* 39: 539–557.

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

Haspelmath, Martin. 2003. "The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison". In Michael Tomasello (ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Vol. 2. Mahwah NJ: Lawrence Erlbaum Associates. 211–242.

Plank, Frans. 1991. "Rasmus Rask's dilemma". In Frans Plank (ed.), *Paradigms: The economy of inflection*. Berlin: Mouton de Gruyter. 161–196.

A step forward in the analysis of visual world eye-tracking data

Vincent Porretta,[1] Aki-Juhani Kyröläinen,[2] Jacolien van Rij,[3] and Juhani Järvikivi[1]

[1]University of Alberta
[2]University of University of Turku
[3]Eberhard Karls Universität Tübingen

The Visual World Paradigm (Tanenhaus, et al., 1995) is an eye-tracking method that provides a means by which to examine online spoken language processing. Eye movements to objects in a visual scene reflect the direction of attention and are driven by both the visual input and the linguistic information that unfolds over time. This method has been used to investigate spoken word recognition (Magnuson, et al., 2007), the influence of visual context on language comprehension (Tanenhaus, et al., 1995), reference (Järvikivi, et al., 2014), and predictive processing (Kamide, et al., 2003). Because it does not rely on reading or overt behavioral response, it is particularly advantageous with populations that cannot read or manipulate response buttons, such as young children (e.g., Järvikivi, et al., 2014).

One of the strengths of this method is that it allows for real-time inspection of the time course of linguistic and visual effects on language comprehension in a fine-grained, millisecond- by-millisecond manner. Typical experiments have employed a factorial design and relied on analysis of variance (ANOVA) for comparing the proportion of looks to a particular object by taking a fixed window of time after a critical item is spoken. While still used, this approach limits the types of questions that can be asked as well as the conclusions that can be drawn from the rich data obtained from this method. The use of generalized linear mixed-effects modeling has attempted to address some of these concerns (see Barr, 2008); however, this still has notable limitations. Here we suggest generalized additive mixed modeling (GAMM) as a method for analyzing visual world data, a regression analysis which can easily account for the inherent non- linearity present in this type time-series data. GAMM has been successfully applied to a variety of linguistic time series data, including the visual world paradigm (Anonymous, accepted; Anonymous, submitted) and event-related potentials (Tremblay & Newman, 2015).

Using a single dataset, we examine the benefits, limitations, and trade-offs inherent to different types of analyses carried out on visual world data. We first turn to a traditional factorial approach with ANOVA which requires averaging over a window of time, subjects, and items. We then examine the benefits gained from a regression approach using linear mixed-effects modeling. Finally, making use of the mgcv package (Wood, 2014), we provide an example

application of GAMM. In this talk we will discuss how this new approach can address many of the limitations inherent to other methods of analysis, making use of the rich time-series data collected using the Visual World Paradigm. We will address four key issues: 1) data loss through averaging; 2) the selection of arbitrary time windows for analysis; 3) inherent autocorrelation present in time-series data; and 4) the interdependence of looks to multiple objects over time. We will demonstrate how GAMM is capable of addressing the aforementioned issues and how this statistical method allows researchers to ask new questions using the Visual World Paradigm.

Bibliography

Anonymous. (submitted). "The influence of gradient foreign accentedness and listener experience on word recognition". *Journal of Experimental Psychology: Human Perception and Performance*.

Anonymous. (accepted). "Using context to resolve object pronouns". In A. Holler, C. Goeb & K. Suckow (Eds.), *Experimental Perspectives on Anaphora Resolution. Information Structural Evidence in the Race for Salience* (pp. t.b.d.).

Barr, D. J. (2008). "Analyzing 'visual world' eyetracking data using multilevel logistic regression". *Journal of Memory and Language*, 59(4), 457–474.

Järvikivi, J., Pyykkönen-Klauck, P., Schimke, S., Colonna, S., & Hemforth, B. (2014). "Information structure cues for 4-year-olds and adults: tracking eye movements to visually presented anaphoric referents". *Language, Cognition and Neuroscience*, 29(7), 877–892.

Kamide, Y., Altmann, G. T. M., Haywood, S. L. (2003). "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements". *Journal of Memory and Language*, 49(1), 133–156.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., Aslin, R. N. (2007). "The Dynamics of Lexical Competition During Spoken Word Recognition". *Cognitive Science*, 31(1), 133-156.

Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). "Integration of visual and linguistic information in spoken language comprehension". *Science*, 268, 1632–1634.

Tremblay, A. & Newman, A. (2015). "Modeling non-linear relationships in ERP data using mixed-effects regression with R examples". *Psychophysiology*, 52(1), 124–139.

Wood, S. N. (2014). *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*". http://cran.r-project.org/web/packages/mgcv/index.html

A computational text analysis of the vapour intrusion corpus

Jeroen Provoost[1] and Karen Victor[2]

[1]Independent researcher, Finland
[2]Birmingham City University

Humans can be exposed to chemicals originating from contaminated land and therefore the related health risks are assessed to reduce it to acceptable levels if needed. One of the exposure routes for humans living on or near a contaminated site is vapour intrusion (VI). This is the process where volatile organic contaminants migrate to the indoor air of a building so humans can inhale the contaminated air. Over the last 25 year an increasing number of publications have become available on the topic of VI. The purpose of this study was to provide a content analysis of relevant publications and uncover emerging areas of research.

The corpus was compiled by searching the 'Web of Science' for articles containing the terms "vapour" or "vapour" and "intrusion" in the title or abstract. In addition relevant reports were collected from other sources. A total of 348 relevant publications were selected resulting in a corpus of around 3 million words. The corpus was analysed by using the computational text-mining tool Leximancer. The tool first divided the corpus in 330.000 text blocks of each 2 sentences long and identifies terms that frequently occur in the corpus, and used these as concept seed words. The next step builds a thesaurus of terms for each concept based on its co-occurrence with the seed words. Additional concepts were discovered during this analysis resulting in a total of around 300 concepts. Leximancer then calculated the co-occurrence of the concepts and creates concept maps indicating the relationship and connections between the concepts, as well as grouping of concepts into logical themes. The free open source Knime data-mining tool was used in addition for further aggregation and visualisation of the raw data from Leximancer. Several figures and graphs, like heat maps, were produces to investigate the relationship between concepts and publications. Leximancer allows for a sentiment analysis of the identified concepts by using an in-build thesaurus of sentiment terms (negative versus positive). In this study the sentiment analysis was used as an adjunct analysis only, indicating the relative frequency of sentiment associations for each term.

The results demonstrate that the major themes and associated concepts (between brackets) are: (1) soil (air concentration), (2) model (parameters), (3) diffusion (coefficient), (4) site, (5) (volatile) compounds, (6) (ground) water, (7) risk (assessment), (8) house (construction), (9) uncertainty, (10) monitoring and (11) (hazardous) waste. Emerging areas of research are (1)probabilistic risk assessment of software model parameters with Monte Carlo analysis (uncertainty), (2) how hazardous waste is related to the site and how sampling, monitoring and

the remediation of ground water is organised, (3) (ad)sorption and equilibrium phase distribution in the water phase and the need to clarify the mass transfer and transport of contaminants, including the diffusion through the boundary flux layer. More fundamentally, this study raises questions whether the way in which VI is modelled, for example the use of the Henry concept for calculating the soil air concentration, is representing the reality. The sentiment analysis shows that a set of publications and their year of publication are associated with negative terms, although the overall sentiment of the corpus is positive. The sentiment analysis should be treated cautiously, given the complexity and short- coming of computer driven sentiment analysis.

This study contributes to the insight in the direction of VI by examining the changes in the literature. The results from this study suggest that VI research is continually changing and will continue to evolve. It is thus possible to track the evolution of science by looking at semantic relationships and clusters of words.

The presentation will focus on the approach followed and show how a corpus can be analysed by using a set of different tools. Results will be used as examples.

The role of correlational studies in linguistics

Sean Roberts

Max Planck Institute for Psycholinguistics

Recent years have seen an increase in correlational studies within linguistics, though many linguists remain sceptical of their role in linguistic research (Ladd, Roberts & Dediu, 2015). This is partly due to the relatively weak tradition of quantitative training within linguistics, but much of the scepticism is also justified. I discuss three problems with correlational studies in linguistics, and some possible solutions.

The first problem is that aspects of language have complex, co-evolutionary relationships with other aspects of language, and with social structures (Roberts & Winters, 2012). This means that a correlation between two variables is difficult to understand without the broader context.

The second problem is the abundance of correlations in large-scale, cross-linguistics databases. Cultural traits are inherited and borrowed, meaning that languages and linguistic groups are not statistically independent. This can lead to apparent correlations between any pair of cultural traits (Roberts & Winters, 2013).

The first two problems lead to the third problem. Now that demonstrating simple correlations between cultural traits is easy and potentially misleading, can correlational studies be trusted at all? More concretely, what is the role of correlational studies? Traditionally, it was hypothesis testing, but increasingly it is being used for hypothesis generation. Is this approach valid in light of the first two problems?

I suggest that these problems can be overcome. The first solution is to adopt stricter standards when performing correlational studies. These include controlling for the shared history of languages and cultures, for example through phylogenetic techniques or mixed effects modelling, but also testing for the serendipity of a correlation. These standards will be demonstrated from a case study of a new analysis of the correlation between future tense and economic behaviour (Roberts, Winters & Chen, under review).

Finally, I will argue that correlational studies do have a role to play in linguistics, if they are used as part of a wider approach to research. In general, I will argue for an approach, where results from different sources are combined to provide a robust argument. Robustness can be achieved by demonstrating the same correlation with different data sources or with different statistical frameworks, or by demonstrating the same core causal components interacting in model systems such as computational models or experiments (Irvine, Roberts & Kirby, 2013). Used in this way, correlational studies have at least two valuable roles. First, to study the

history of language that can no longer be directly observed. Secondly, since collecting detailed linguistic data can be expensive and takes time, correlational studies can be used as a kind of feasibility study to guide future research.

References

Irvine, L., Roberts, S. G., & Kirby, S. (2013). "A robustness approach to theory building: A case study of language evolution". In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (CogSci 2013) (pp. 2614-2619)

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015). "Correlational studies in typological and historical linguistics". *Annual Review of Linguistics*, 1, 221-241. http://www.annualreviews.org/doi/abs/10.1146/annurev-linguist-030514-124819.

Roberts, S. G., & Winters, J. (2013). "Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits". *PLOS ONE*, 8(8): e70902

Roberts, S. G., & Winters, J. (2012). "Social structure and language structure: The new nomothetic approach". *Psychology of Language and Communication*, 16, 89-112. doi:10.2478/v10057-012-0008-6.

Roberts, S. G., Winters, J. & Chen, K. (under review) "Future tense and economic decisions: controlling for cultural evolution". *PLOS ONE*.

On the design, in practice, of typological microvariables

Erich Round and Jayden Macklin-Cordes

University of Queensland

Modern, large-scale typology, with its enormous datasets and batteries of algorithms rather than humans doing the comparisons, is more sensitive than ever to choices in how to code up language data. Thus we see increasing theoretical emphasis on the need for variables which robustly compare like with like (Haspelmath 2010, *et seq*.); which typologize language facts, not quirks of descriptive traditions (Hyman 2014); which decompose traditional variables into their finer-grained constituent notions (Bickel 2010, Corbett 2005, *et seq*.); and which attend closely to the logical relationships between those constituents (Round 2013). But how does this theory translate into the work of actually building such variables? We offer a view as we decompose one traditional typological macrovariable: the presence or absence in a language of phonemic pre-nasalised stops. Our findings help to nuance the theory microvariate analysis, and assess the practicality of some recent, theoretically- motivated proposals for the design of very large typological datasets.

The presence or absence of pre-nasalised stops, as a typological variable, has figured in several recent, large-scale typological studies (Dunn *et al*. 2005, Reesink *et al*. 2009, Donohue *et al*. 2013), however preliminary investigation (Round 2013) suggested that the variable performed poorly at comparing like languages with like. In response, we set about to develop a finer-grained set of microvariables, which should encode richer information and perform better at comparing like with like. We coded them for >200 Australian and Papua New Guinean doculects (Cysouw & Good 2013), and paid particular attention to the challenges we encountered.

A first finding is that, as a process which researchers undertake, the decomposition of typological variables is iterative. The aim of decomposing a macro-variable is to tease apart some of the linguistic properties which it conflates, and which lead to false comparisons of unlike with unlike. Our experience shows that after a given round of decomposition, there will likely be further conflations that emerge and require addressing. Accordingly, it would be mistaken to view the building of microvariables as a 'fell-swoop' process, or one which replaces 'imperfect variables' with 'perfect variables'. Rather, it is a process for improving dataset design in an iterative fashion.

A second finding is that the logical dependencies between variables, including those which can be problematic for statistical analysis, may also be iteratively related. Our method, however, tracks those relationships as microvariable are elaborated.

Finally, our preliminary suspicions about the pre-nasalised stop macro-variable appear to be valid. The macro-variable 'are there prenasalised stops' is primarily, and covertly, a variable about the size of consonant clusters, but one which (i) does poorly at grouping like languages with like, and (ii) is modulated by a second micro-variable which appears to us to act as a proxy for different schools of linguistic analysis, and not linguistic facts. Therefore, we strongly endorse a rapid shift towards microvariate typology (Bickel 2010; Round 2013), to reach better, clearer and deeper generalizations about human languages and human Language, from sound empirical data.

REFERENCES

Bickel, Balthasar. 2010. "Capturing particulars and universals in clause linkage". In Bril, I. (ed.) *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*. Amsterdam: John Benjamins, pp. 51-101.

Corbett, Greville G. 2005. "The canonical approach in typology". In Z. Frajzyngier, A. Hodges, & D.S. Rood (Eds.), *Linguistic diversity and language theories*. Amsterdam: John Benjamins, pp. 25-49.

Cysouw, Michael & Good, Jeff. 2013. "Languoid, doculect, and glossonym: Formalizing the notion 'language"'. *Language Documentation & Conservation*. 7:331–59.

Donohue, Mark, Rebecca Hetherington, James McElvenny and Virginia Dawson. 2013. *World phonotactics database*. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au.

Dunn, M., A. Terrill, G. Reesink, R.A. Foley & S.C. Levinson. 2005. "Structural phylogenetics and the reconstruction of ancient language history". *Science* 309: 2072-5.

Haspelmath, Martin. 2010. "Comparative concepts and descriptive categories in crosslinguistic studies". *Language* 86.3: 663-687.

Hyman, Larry. 2014. "What (else) depends on phonology?" *Dependencies Among Systems of Language workshop*, Château de la Poste, Ardennes.

Reesink, Ger, Michael Dunn & Ruth Singer. 2009. "Explaining the linguistic diversity of Sahul using population models". *PLoS Biology* 7: e1000241.

Round, Erich R., 2013. "Big data typology and linguistic phylogenetics: design principles for valid datasets". *21st Manchester Phonology Meeting*, Manchester.

Computational traces of semantic polysemy: the case of Finnish epäillä and its derivatives

Jutta Salminen and Antti Kanner

University of Helsinki

In our study, we seek to find ways to quantitatively map semantic and contextual similarities of derivationally closely related group of words and their meaning variants. The object of our investigation is the various constructions containing either the verb *epäillä* ('to doubt, suspect, suppose') or any of its derivatives, such as *epäily*(*s*) 'a doubt, suspicion', *epäilyttää* 'be in suspicion', *epäilyttävä* 'suspicious'. The verb *epäillä* and the nouns *epäily*(*s*) have a dual semantic nature: they can be used with either negative or affirmative implication towards the propositional content of their syntactic object, or other contextually available complement. This polar polysemy is indicative of the previous qualitative studies about the verb epäillä (Salminen 2012, forthc.).

In addition to earlier studies, our observations on the use of the noun *epäily* in legal language motivate the study: this noun refers to both suspicion of crime and reasonable doubt, the term applied in the judgment of evidence in court. The existence of these two terms may cause contradictory interpretations: the established nature of the 'suspicion of crime' term occasionally interferes with the comprehension of the more newly introduced 'reasonable doubt' term. The latter term is about to be officially implemented in Finnish law, which highlights the topicality of the subject (HE46/2014). Thus, in the current study, we examine the usage of these lexemes in a corpus composed of legal documents gathered from the FDB legislative database.

Disambiguation is one of the main topics of natural language processing (Bakx 2006: 1–2 and references therein). In our study, however, we do not see polysemy as a technical problem looking to be solved, but more as an inherent property of any natural language and as an object for interest in itself. Thus, we intend not to discuss the algorithms the methods we use are based on. Instead, our perspective is linguistic: we seek to evaluate the usefulness of these methods as tools of a linguistic study. We look to examine how the polysemy of the derivations of the verb *epäillä* shows up in different quantitative distributions. Moreover we will try to analyze which computational methods yield results which are most compatible with the view of semantics of the studied words described in previous qualitative studies.

The exact quantitative methods in this study range from crude distributional frequency data (described in, for example, Baayen 2001) to statistical analysis used in comparing similari-

ties in word contexts, such as term-document (LSA, cf. Landauer, Foltz & Laham 1998) and word- context (WordICA, cf. Honkela, Hyvärinen & Väyrynen 2010) matrices. The study is still in its early stages but we hope to present its preliminary results and discuss its methodological choices with the focus on evaluating the usefulness of the abovementioned methods in describing the semantics of inherently polysemous words.

References

Baayen R. H. 2001. *Word Frequency Distributions*. Springer.

Bakx G. E. 2006. "Machine Learning Techniques for Word Sense Disambiguation". Dissertation. Universitat Politécnica de Catalunya. Electronically available: http://www.cs.upc.edu/~escudero/wsd/06-tesi.pdf

FDB = Finlex Data Bank. Finnish Ministry of Justice. www.finlex.fi. HE46/2014 = Hallituksen esitys eduskunnalle oikeudenkäymiskaaren 17 luvun ja siihen liittyvän todistelua yleisissä tuomioistuimissa koskevan lainsäädännön uudistamiseksi. [*Government bill about renewing the legislation regulating legal proceedings (chapter 17) and presentation of evidence*.] Electronically available: https://www.finlex.fi/fi/esitykset/he/2014/20140046

Honkela T., Hyvärinen A. & Väyrynen J. J. 2010. "WordICA—emergence of linguistic representations for words by independent component analysis". *Natural Language Engineering* 16 (3): 277–308. Cambridge University Press.

Landauer T. K, Foltz P. W. & Laham D. 1998. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Salminen J. Forthcoming. From doubt to supposition: The construction-specific meaning change of the Finnish verb *epäillä*. In Juhani Rudanko, Jukka Havu, Mikko Höglund & Paul Rickman (eds.), Perspectives on Complementation. Proceedings of Tampere Linguistic Forum 2013, 157–176. London: Palgrave.

Salminen J. 2012. Epäilen, että tämä verbi kyseenalaistaa täydennyksensä proposition. *Epäillä*-verbin merkityksen kehityksestä ja monitulkintaisuudesta. [I doubt/suppose that this verb questions the proposition of its complement. On the development and ambiguity of the meaning of epäillä] Master's thesis. The department of Finnish, Finno-Ugrian, and Scandinavian Studies. University of Helsinki.

The Kruszewski–Kuryłowicz Rule: On Its Potential And How To Apply

Nezrin Samedova

Azerbaijan University of Languages

1. The goal of the paper is to draw attention to a quantitative method the core of which is a simple (but little known, to my knowledge) rule, namely the Kruszewski–Kuryłowicz rule. The rule is of a special interest, as it is a manifestation of the fundamental principle of economy. Cf. its most specific wording (among those available): "...The more generalized (poorer) the content of a sign, the wider its sphere of using by speakers; the more special (richer) the content, the narrower the sphere of its not only internal usage (=inside the system) but also external usage (=in the linguistic community)" [1: 19].

2. We have applied the rule to the testing of the solution to an old semantic puzzle.

2.1. It is long believed that, in Russian, there exists the construction *stat'+INF* the main component of which is a perfective stat' that has no imperfective counterpart. According to the literature, its meaning contains the seme 'beginning' and the construction is synonymous to the construction *nachat+INF*, cf., eg, [2]. Interestingly, some researches compare their usage frequency ([3: 31, 32; 4: 67, 68; 5: 5; 6: 163; 7]), however, the data across those studies vary greatly and are rarely interpreted. At the same time, researchers point out that the seme may be absent and, respectively, the constructions are not synonymous, cf., e.g., [2: 99].

On one hand, the verb is qualified as perfective, cf., eg, [2: 97]. On the other hand, it is common to state that phrases like *stanu+INF* can be synonyms of the periphrastic form for the future tense of imperfectives (*budu+INF*), cf., eg, [2: 99]. As for their usage, it is noted without any interpretations that the former are less frequent [8: 233; 9: 41].

2.2. Theoretical perspective I employ differentiates two homonymous constructions *stat'+INF*.

The one that is synonymous to the construction *nachat+INF* is used much more frequently than its synonym. For instance, in our random sample, the ratio is 2600:1017.[1] Having interpreted the fact with the rule, we have become able to prove once more that *stat'+INF* is characterized with the aspectual seme 'initial bound'. Then we compared the postulated homonyms and revealed that one of them has a wider sphere of usage than the other. The Kruszewski–Kuryłowicz rule has allowed us to conclude that the former is imperfective and that it possesses the same syncretic seme 'process' as its perfective homonym.

Having described both homonyms, we have got the possibility to explain why they differ

[1] As one of the requirements for applying the Kruszewski–Kuryłowicz rule is to ensure that compared entities are accurately distinguished, only phrases like *stal+INF* and *načal+INF* used in affirmative sentences were counted (it is the context within which the postulated homonym is identified exactly).

so significantly from the construction *budu+INF* regarding their usage frequency (cf., e.g., the ratio 10,183 : 135,232 in the Russian National Corpus). The meaning of the perfective homonym is richer than the meaning of *budu+INF*. As for the meaning of the imperfective homonym, though it contains the same number of semes as the meaning of *budu+INF*, the syncretic seme 'process' it possesses is less clear. In other words, the comparison exposes the limitation to the Kruszewski–Kuryłowicz rule, namely the principle of clarity.

References

1. Kuryłowicz, Je. (1962). "Lingvistika i teorija znaka". In: Kurilovič, Je. *Očerki po lingvistike*. Moscow. P. 9-20.

2. Korotkova, N.A., Saj, S.S. (2006). "Glagol *stat'* v russkom jazyke: semantika, sintaksis, grammatikalizacija". In: *Tret'ja konferencija po tipologii i grammatike dlja molodyx issledovatelej*. Materialy. S. Petersburg. P. 96-101.

3. Sukhotin, V.P. (1960). *Sintaksičeskaja sinonimika v sovremennom russkom literaturnom jazyke*. Glagol'nyje slovosočetanija. Moscow.

4. Tikhonov, A.N. (1959). "Sposoby vyrajenija načinatel'nogo značenija glagolov v russkom jazyke". In: *Trudy Uzbekskogo gos. un-ta im.* A. Navoi. Novaja serija, No 95, Samarkand. P. 43-75.

5. Moldovan A. M. (2010). "K istorii fazovogo glagola stat' v russkom jazyke". In: *Russkij jazyk v naučnom osveščenii*, No 1 (19), P. 5-17.

6. Khrakovskij, V.S. (1987). "Semantika fazovosti i sredstva jejo vyraženija". In: *Teorija funkcional'noj grammatiki. Vvedenije. Aspektual'nost'. Vremennaja lokalizovannost'. Taksis*. Leningrad. P. 153-178.

7. Divjak, D., Gries, S.Th. (2009). "Corpus-based cognitive semantics: a contrastive study of phasal verbs in English and Russian". In: *Studies in cognitive corpus linguistics*. Ed. K. Dziwirek, B. Lewandowska-Tomaszczyk. Frankfurt am Main. P. 273-296.

8. Bulakhovskij, L.A. (1952). *Kurs russkogo literaturnogo jazyka*. Vol. 1, Kijev.

9. Fuksman, A.A. (1959). "Sočetanije infinitiva s vspomogatel'nym glagolom stanu v sovremennom russkom jazyke". In: *Kratkije soobščenija Uzbekskogo gos. un-ta im. A. Navoi. Kafedra russkogo jazykoznanija*. Samarkand. P. 39-43.

Exploring distributional patterns in complementation systems

Karsten Schmidtke-Bode

Friedrich-Schiller-Universität Jena

Complementation ranks among the most intensively studied phenomena in many subdisciplines of linguistics. In linguistic typology, too, due attention has been paid to various syntactic, semantic and information-structural aspects of complementation (cf. Horie 2001 for an overview), and two landmark reference articles (Noonan 1985, Dixon 2006) have provided useful frameworks for investigating complement clauses in individual languages around the world. However, what is notably absent from the literature is a cross-linguistic study that puts these reference works onto a broad empirical basis, examining their conceptual apparatus and distributional hypotheses against a balanced representative sample of the world's languages, and using statistical techniques that reflect current developments in typological methodology. In this paper, I present selected findings from a recent project that sought to address this desideratum.

In the course of the project, 100 carefully sampled languages were scanned for their (major) complementation constructions, each of which was then submitted to multivariate typological coding and analysis (cf. Bickel 2010). In particular, the internal structure of the complement clause was dissect- ed with regard to parameters like its verb form, TAM expression, coding of the internal arguments and modifiers, the presence and type of nominal flagging on clause boundaries, among others. Similarly, the distributional potential of each construction over a wide array of syntactic and semantic environments in the matrix was recorded (e.g. how readily and idiomatically the clause can function as a complement of perception predicates, as the subject argument of a transitive predicate, etc., and which syntagmatic positions it can occupy in these functions). When operationalized appropriately, these variables can be taken to gauge the degree to which a given complementation pattern is 'desententialized' (Lehmann 1988) and turned into a nominal structure, and how this correlates with the semantic and syntactic relationships of the complement to the matrix. In this way, it becomes possible to develop an entirely data-driven, bottom-up account of distributional trends in comple- mentation systems. For example, cluster-analytical and scaling techniques allow for a robust statistical underpinning of patterns like Givón's (1980) 'binding hierarchy' or purely theoretical classifications of complement-taking predicates. By the same token, novel dimensions of analysis can be taken into consideration: for instance, it has never been investigated empirically which syntactic functions (S/A/P/E) tend to be covered by complements of different structural types, and which restrictions apply in such environments. Also, it can be shown that when complement

clauses are clustered according to their distributional profile in the data, the resulting groups often share a similar historical background. In other words, the available evidence suggests that the contexts in which complement clauses emerge give rise to similar patterns of lexical diffusion and, ultimately, to synchronic distributions.

Therefore, in keeping with the general goals of the symposium, my aim in the talk will be to illustrate the potential that multivariate approaches to clause linkage and exploratory methods like cluster analysis and *NeighborNet* harbour for the comparative study of complementation systems, and to reflect on the conceptual challenges that are involved even after decades of intensive research.

References

Bickel, Balthasar (2010). "Capturing particulars and universals in clause linkage: A multivariate analysis". In: *Clause Linking and Clause Hierarchy: Syntax and Pragmatics*. Ed. Isabelle Bril. Amsterdam, Philadelphia: John Benjamins. 51–102.

Dixon, R.M.W. (2006). "Complement clauses and complementation strategies in typological perspective". In: *Complementation: A Cross-Linguistic Typology*. Eds. R.M.W. Dixon and Alexandra Y. Aikhenvald. Oxford: Oxford University Press. 1–48.

Givón, Talmy (1980). "The binding hierarchy and the typology of complements". *Studies in Language* 4: 333–377.

Horie, Kaoru (2001). "Complement clauses". In: *Handbook of Language Typology and Language Universals*. Eds. Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible. Berlin, New York: Walter de Gruyter. 979–993.

Lehmann, Christian (1988). "Towards a typology of clause linkage". In: *Clause Combining in Grammar and Discourse*. Eds. John Haiman and Sandra A. Thompson. Amsterdam, Philadelphia: John Benjamins. 181–226.

Noonan, Michael (1985). "Complementation". In: *Language Typology and Syntactic Description. Vol. II: Complex Constructions*. Ed. Timothy Shopen. Cambridge: Cambridge University Press. 52–150.

Quantitative Study of Russian Spoken Speech based on the ORD Corpus

Tatiana Sherstinova

St. Petersburg State University

The ORD corpus of Russian Everyday communication is the largest collection of contemporary Russian spoken speech recorded in natural communicative situations. It contains more than 1000 hours of recordings made by 110 participants-volunteers who spend a whole day with switched-on dictaphones, hanging around their nects and recording all their audible communication. The ORD corpus provides rich material for phonetic, lexical and grammatical studies of Russian everyday speech, for research in sociolinguistics and psycholinguistics, for studies in pragmatics, discourse analysis, for behavioral and communication studies, anthropological linguistics, speech technology developments and for teaching Russian as a foreign language.

Sound data are being annotated on different levels — phonetic, lexical, grammatical, pragmatic — and each of these aspects becomes a subject for quantitative analysis. To describe speech material and annotation data we actively use descriptive statistics and statistical tools for testing hypothesis. Besides, to study particular linguistic aspects we use correspondent quantitative measures (indices of lexical richness, concentration index, syntactic complexity measures, etc.).

Automatic processing of speech transcripts allows us to calculate number of words and syllables for each phrase. For example, according to our data, the average utterance length in Russian spoken communication is 4.35 words (SD is 4.02 words); 25.26% of all spoken Russian utterances consist of a single word or word-like particles, two-word utterances make 15.58% of the whole data, and three- word utterances has the third rank making 12.45%.

Due to multimedia annotation of speech signal made in linguistic annotator ELAN, each annotated phenomena (sound, morpheme, word, phrase, turn, minidialogue, miniepisode, etc.) refers to particular segment in correspondent sound file, therefore it has duration. Thus, temporal study of elements is possible for all linguistic levels. For example, on phonetic level we study speech rate, rhythmic patterns, temporal registers of Russian everyday speech and other temporal phenomena.

Further, we use quantitative methods for studying functional activities of different elements and their classes (phonemes, words, speech patterns, syntactic models, etc.) for the corpus in the whole and for its particular samplings (e.g. for diverse communication situations). For this purpose we make and compare frequency lists. Thus, we obtained the "top list" of Russian everyday utterances, frequency word lists for different scenes of communication, speakers'

social roles and other parameters. Besides, we've got pilot data concerning distribution of grammatical forms and even speech act categories in Russian everyday speech.

Recently, we have started a large sociolinguistic project having aim to analyze special characteristics of everyday Russian used by different social groups, and revealing the actual functioning of Russian in society. For describing linguistic characteristics of everyday speech for each analyzed group (sociolect) and their comparison we will widely use quantitative methods, too. The study is supported by Russian Scientific Foundation, project # 14-18-02070 "Everyday Russian Language in Different Social Groups".

Register comparisons in the study of contrastive negation in English

Olli O. Silvennoinen

University of Helsinki

Contrastive negation refers to expressions in which one element is negated and another, parallel element is affirmed (Gates Jr. & Seright 1967; McCawley 1991). The central cases of contrastive negation are those that display a replacive contrast, such as (1). These can be considered contrastive negation proper. However, contrastive negation is also formally associated with forms that present a *restrictive* contrast, as in (2), as well as additive contrast, exemplified in (3):

(1) a. When, I wonder, did it become fashionable for politicians to talk **not about the world but about the planet**?
    b. We have to accept this is **showbiz** now, **not a sport**.

(2) Environmental protection and wealth creation are **in tension but not necessarily in conflict** [. . .].

(3) a. a. Wire purports to be about **America not just Belushi**, but this is trite stuff about decadence in the Tinseltown [. . .]
    b. THERE are times when life **not only kicks you in the teeth, but follows up with the knee in the groin and the rabbit punch to the back of the neck as well**.

Previous studies on contrastive negation have primarily been based on intuitive, anecdotal or experimental data (e.g., Konietzko & Winkler 2010). In addition, the various classes of contrastive negation are usually treated in isolation, without accounting for the links (both formal and semantic) between *not X but Y* and *not only X but also Y*, or, conversely, without drawing a clear line between them. In my paper, I shall consider constructions such as (1)–(3) in real language use, by using data from the broadsheet newspaper and conversation components of the British National Corpus (BNC). This allows me (i) to see the way in which register affects the distribution and use of negative-contrastive constructions, and (ii) to explore the differences and similarities between speech and writing in the expression of contrastive negation. In other words, I shall consider the kinds of research questions that can be answered by comparing apples and apples (i.e., two newspaper genres) as well as apples and oranges (i.e., newspaper articles and casual conversation) in studies on grammar. This has implications for the reliability and generalizability of claims on grammar.

My analysis of the newspaper data shows that the prevalence of contrastive negation is sensitive to register in ways that are meaningful and predictable. For example, replacive contrast

favours argumentative genres such as editorials and letters-to-the-editor while restrictive contrast is less prevalent in them. On the other hand, narrative texts such as sports articles have a lower prevalence of replacive contrast. I explain these quantitative findings by placing them in the context of previous studies on the register distributions of grammatical constructions (see Biber 2012) and by considering the functions of the text types being investigated.

References

Biber, Douglas. 2012. "Register as a predictor of linguistic variation". *Corpus Linguistics and Linguistic Theory* 8(1). 9–37.

Gates Jr., Dave L. & Orin Dale Seright. 1967. "Negative-contrastive constructions in standard modern English". *American Speech* 42(2). 136–141.

Konietzko, Andreas & Susanne Winkler. 2010. "Contrastive ellipsis: Mapping between syntax and information structure". *Lingua* 120(6). 1436–1457.

McCawley, James D. 1991. "Contrastive Negation and Metalinguistic Negation". *CLS* 27(2). 189–206.

A Synchronous Corpus in Chinese: Methodology and Rationale in Construction and Enhanced Application

Benjamin Tsou

City University of Hong Kong

The major rationale and methodology in the curation of language data and the construction of linguistic corpus have been motivated by the need to reliably provide objective and quantitative evidence for qualitative analysis of language variation beyond what may be readily obtained through personal introspection. In the last two decades this approach has represented an increasing realization that going beyond the ideal speaker could contribute to fresh and fruitful studies of the relevant language.

This paper makes use of LIVAC, a gigantic *synchronous corpus* of language data [http://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus] which have been rigorously and continuously drawn from the pan-Chinese media over the past 19 years. It has evolved into a monitoring corpus which can be conceptualized as consisting of serial time capsules containing not only linguistc artefacts but also providing a means to appreciate deeper aspects of the society and culture which have embodied them and which also contain the undercurrents and cross-currents in their developments. It attempts to show how useful it can be for exploring and tracking developmental trends in language and beyond language, among the major Chinese speech communities in mainland China, Hong Kong and Taiwan.

They include: (1) the societal significance of the differential neologistic developments and transfers amongst these major communities in the context of exocentric and endocentric influence, and of their broader social and cultural relevance (e.g. *see Fig.1*), (2) ethnicity and sometimes controversial national identity in the two recent decades when there have been colossal economic, social and political transformations amongst them, and (3) their separate reception and expectation of leaders, within the local community and extending to the much broader international contexts.

Relevant issues will be discussed in relation to various natural language processing efforts involving neologism mining and tracking, collocation analysis of conceptual metaphors, opinion mining based on sentiment analysis and the curation and analysis of parallel texts (e.g. *see Fig.2*). The usefulness of a monitoring corpus based on the synchronous corpus approach will be demonstrated through the application of familiar and innovative computational techniques

in corpus linguistics.

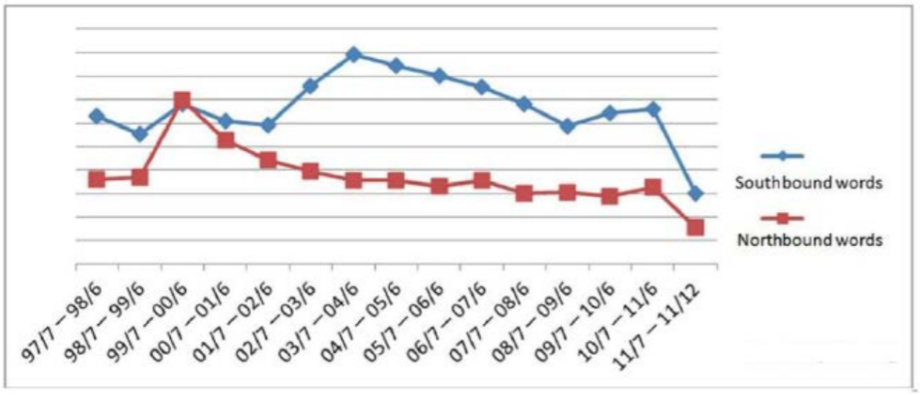


Figure 1. Comparative lexical transfer 1997-2012: Mainland China to Hong Kong (Southbound) and Hong Kong to mainland China (Northbound)

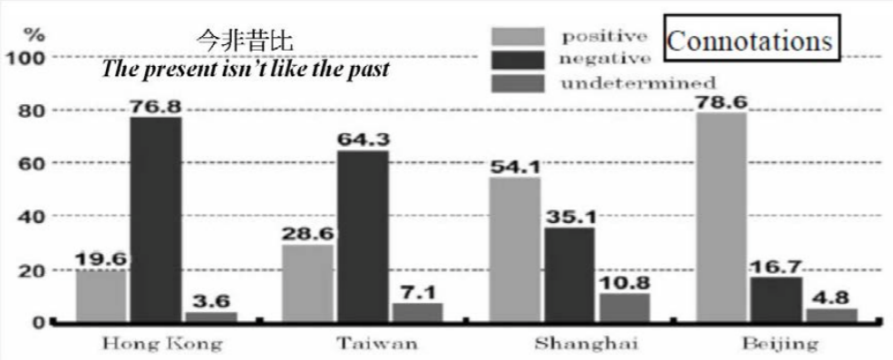


Figure 2. Variations in polarity of idiomatic expressions among Pan-Chinese communities

References

Tsou, Benjamin, and Kwong, Olivia. (in press). "Some Quantitative and Qualitative Characteristic Features of the Chinese Language". In Wang S.-Y. William and Sun C.-F. (eds.), *Oxford Handbook of Chinese Language anld Linguistics*. Oxford University Press.

Zōu jiāyàn, Lí bāngyáng. 2003. "Hànyǔ gòngshí yǔliàokù yǔ zīxùn kāifā" zài "zhōngwén zīxùn chǔlǐ ruògān zhòngyào wèntí" *Issues in Chinese Information Processing* ["973 j'huà guójiā yǔyán zìrán yǔyán lǐjiě yǔ zhāshì wājué" zǒngtǐ kānwù] Xú bō, Sūn màosōng, Jìn guāngjǐn zhǔbiān, kēxué chūbǎn shè, pp147-165.

Measuring pitch in learner speech

Ullakonoja, Riikka

University of Jyväskylä

Pitch of voice (F0) is commonly regarded as a correlate of intonation in speech. According to some previous studies languages are usually spoken with different pitch characteristics, e.g. mean pitch and pitch range. For example, Russian women speak with a higher pitch than Finnish women (Ullakonoja 2007) Based on recent findings on how Finns perceive Russian intonation (Skrelin, Volskaya, Evgrafova & Ullakonoja 2014), it seems possible that Russian accented Finnish can convey an unintended emotional message due to the pitch contour typical of Russian. Thus, language learning point of view, it would be important to learn to use pitch correctly in the target language. The present study aims at comparing pitch characteristics in different languages by learners of different first language backgrounds.

More precisely, I investigate pitch in Russian spoken by Finnish and Finland-Swedish speakers as well as Finnish spoken by Russian and Finland-Swedish speakers. The study is a part of a larger research project, **Fokus på uttalsinlärningen med svenska som mål- och källspråk** (http://www.jyu.fi/fokus), that focuses on learning and teaching pronunciation in these languages.

Pitch can be measured and calculated acoustically from speech signal using computer software, such as e.g. Praat (http://www.praat.org). The measuring can be done in absolute values (Hertz) or in a relative scale (such as e.g. semitones). Semitone scale is used in the present study as it can best represent the auditory perception of pitch (Nolan 2003). Different statistical measurements have been previously used in pitch comparisons of different speakers and languages, e.g. mean, median, 95% of the values around the mean or the range between lower 25 % and upper 75 % or lower 10 % and upper 90% of the pitch points (Bezooijen 1995; Carlson et al. 2004; Lennes 2009; Mennen et al. 2007). Instead of these statistical measures, Mennen et al. (2008) and Patterson (2000) have used more linguistically based measures and investigated linguistically significant pitch turning points instead of absolute pitch values. In the presentation I will compare different pitch measurement techniques to my data. Comparing pitch range measurements in different languages, or even in one language spoken by learners of different first language backgrounds is relatively rare in previous studies, probably due to the lack of consensus on the most suitable measurement technique.

References

Bezooijen, R. V. 1995. "Sociocultural aspects of pitch differences between Japanese and Dutch women". *Language and Speech* 38, 253–265.

Carlson, R., Elenius, K. & Swerts, M. 2004. “Perceptual judgments of pitch range”. In B. Bel & I. Marlin (eds.) *Proceedings of Speech Prosody* 2004, Nara, Japan, 689–692.

Lennes, M. 2009. “Segmental features in spontaneous and read-aloud Finnish”. In V. de Silva & R. Ullakonoja (eds.) *Phonetics of Russian and Finnish, general description of phonetic systems, experimental studies on spontaneous and read-aloud speech*. Frankfurt am Main: Peter Lang, 145–166.

Mennen, I., Schaeffler, F. & Docherty, G. 2007. “Pitching it differently: A comparison of the pitch ranges of German and English speakers”. In J. Trouvain & W. J. Barry (eds.) *Proceedings of the 16th ICPhS, 6–10 August 2007, Saarbrücken*. Saarbrücken: Universität des Saarlandes, 1769–1772.

Mennen, I., Schaeffler, F. & Docherty, G. 2008. “A methodological study into the linguistic dimensions of pitch range differences between German and English”. In P. A. Barbosa, S. Madureira & C. Reis (eds.) Proceedings of the Speech Prosody 2008, Campinas, Brazil: Editora RG/CNPq, 527–530.

Nolan, F. 2003. ‘´Intonational equivalence: An experimental evaluation of pitch scales”. In M. J. Solé, D. Recasens & J. Romero (eds.) Proceedings of the 15th ICPhS, Barcelona, 3–9 August 2003, 771–774.

Patterson, D. 2000. *A linguistic approach to pitch range modelling*. University of Edinburg. Doctoral Dissertation.

Skrelin, P.A., Volskaya, N.B., Evgrafova K.V. & Ullakonoja, R. 2014. “The development of new corpora for under-resourced languages using data available for well-resourced ones”. *Proc. of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages*. 243–246.

Ullakonoja, R. 2007. “Comparison of pitch range in Finnish (L1) and Russian (L2)”. In Trouvain J. & Barry, W. (eds.), *Proc. 16th ICPhS, Saarbrücken*, Germany. Saarbrücken: Universität des Saarlandes. 1701–1704.

Using Bayesian structural equation modeling in second language research

Laszlo Vincze

University of Helsinki

This paper will provide an insight into some of the benefits of using Bayesian structural equation modeling (BSEM; Muthén & Asparouhov, 2012) in second language research. A specific focus will be placed on the applicability of BSEM for data with small sample sizes and non-normal distributions (e.g. Schoot et al., 2014). The empirical material was collected in 2015 among (monolingual) Finnish speaking recruits, who do their military service in the Swedish language marine infantry unit located in Dragsvik, Raseborg/Raasepori (N = 43; total sample). The tested theoretical model fit the data well. The results indicated that negative experiences with learning Swedish as a second language in secondary school led to poor competencies in Swedish, which, in turn, inspired soldiers to do their military service in Swedish in order to improve their Swedish skills. In addition and in consistence with the predictions of self-determination theory (for a recent review, see Deci & Ryan, 2011) poor competencies in Swedish motivated participants to do their military service in Swedish mostly when they perceive that they are able to master Swedish and also when they are in a great need of competence in Swedish.

References

Deci, E. L. & Ryan, R. M. (2011). "Self-determination theory". In Paul A. M. Van Lange, Arie W. Kruglanski & E. Tory Higgins (eds), *Handbook of theories of social psychology* (pp. 416– 433). London, UK: Sage.

Muthén, B., & Asparouhov, T. (2012). "Bayesian structural equation modeling: a more flexible representation of substantive theory". *Psychological methods*, 17(3), 313–335.

Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). "A gentle introduction to Bayesian analysis: applications to developmental research". *Child development*, 85(3), 842–860.