ISLE Workshop:

The "quantitative crisis", cumulative science, and English linguistics

Abstracts

Bernd Kortmann

Reflecting on the quantitative turn in linguistics

Linguistics, English linguistics in particular, has witnessed a quite remarkable quantitative turn over the last two decades. Major drivers of this development have been the design of ever more and ever larger electronic corpora, the increasing importance of psycho- and neuro-linguistic experiments in exploring language processing, and the availability of ever more sophisticated statistical tools for handling large, complex data sets. Has this quantitative turn been to the detriment of qualitative methods, or even of linguistic theorizing in general? Has linguistics reached the point of a "quantitative crisis" yet, or is it still a discipline characterized by a healthy equilibrium, if not mutual reinforcement, of quantitative and qualitative approaches? What are major repercussions of the strong quantitative turn for the publication system of (English) linguistics?

These are the overarching questions underlying the reflections offered in this talk. On the way towards answering them, the following issues will be addressed: the advances made in largescale aggregation and metricization of language structures (e.g. McMahon & Maguire 2013; Szmrecsanyi 2013; Szmrecsanyi & Wälchli 2014), the (sometimes more, sometimes less explicit) cognitive claim underlying probabilistic linguistics (e.g. Bod 2010) and cognitive corpus linguistics (e.g. Arppe et al. 2010), and major caveats for (especially the future generation of) practitioners of quantitative methods. One general conclusion will be that, apart from the need for an increased level of sophistication and problem-awareness in choosing, applying and interpreting statistical methods in linguistic research (cf. in this respect e.g. Gries 2015), the natural next step for a strongly quantitatively oriented English linguistics needs to be the increasing adoption of multi-method approaches (cf. as one source of inspiration the contributions to the special issue of *Cognitive Linguistics* edited by Divjak, Levshina & Klavan in 2016). This will be exemplified by recent studies on entrenchment (Blumenthal-Dramé 2012) and the competition between syntheticity and analyticity in English (Kunter 2017).

References

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1): 1–27.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do not Reveal about the Mind*. Berlin: Mouton De Gruyter.

Bod, Rens. 2010. Probabilistic linguistics. In Bernd Heine & Heiko Narrog, eds. *The Oxford Handbook of Linguistic Analysis.* Oxford: Oxford University Press. 633–662.

Divjak, Dagmar, Natalia Levshina & Jane Klavan. 2016. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27(4): 447–463.

Gries, Stefan Th. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language & Linguistics* 16: 93–117.

Kunter, Gero. 2017. Processing complexity and the alternation between analytic and synthetic forms in English. Postdoctoral thesis, University of Düsseldorf, Germany.
McMahon, April & Warren Maguire. 2013. Computing linguistic distances between varieties. In Manfred Krug & Julia Schlüter, eds. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press. 421–432.
Szmrecsanyi, Benedikt. 2013. Analyzing aggregated linguistic data. In Manfred Krug & Julia Schlüter, eds. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press. 433–455.
Szmrecsanyi, Benedikt & Bernhard Wälchli, eds. 2014. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*. Berlin: Mouton de Gruyter.

Sean Wallis

What are the tasks of statistics in linguistics?

Linguistics, like many research domains, is becoming more statistical. Students on linguistics courses are enjoined to study statistics. Academic journals are tightening up on statistical claims. But what kind of statistics should linguists study? "Statistics" is usually discussed as if it were a single discipline, but it is a branch of applied mathematics driven by particular research problems. Population eugenicists in the early 20th century laid the foundations of experimental design and statistics (see Stigler 2012), building on the work of Laplace and Gauss in the 19th century. Mid-20th century actuarial mathematicians introduced risk models and effect sizes (e.g. Cramér 1944). In the late 20th century, medical statisticians refocused attention on confidence intervals (Newcombe 1998). Likewise, a recent emphasis on multi-variate model-fitting and statistical machine learning has both recognised the complexity of nature and exploited the availability of computation.

This diverse set of goals raises the question as to the role statistics has to play in linguistics. Not all methods are equally relevant, even if they are available to us. (It is difficult to see what role an actuarial risk model might have for linguistic theory, for instance.) Linguistics does not need an independent branch of statistics, but linguists should be trained in the most relevant methods.

Enter the 'crisis'. What has been called the 'quantitative crisis' concerns a debate in psychology triggered by the observation that a number of superficially statistically valid academic papers have failed to replicate (see corplingstats.wordpress.com/2017/02/16/the-replication-crisis). Correlations found in one data set disappear when new data is collected. There are multiple potential reasons for this, but particular attention is paid by Gelman & Loken (2013) to "p-hacking", which can be understood as *a crisis of over-fitting*, which might be due to computational overfitting or a selective citation of significant results. This crisis has triggered a critical reappraisal of experimental procedures termed "the New Statistics" (Cumming 2014). We contend that this problem is both statistical and domain-theoretical. Fitting data to a model based on a set of prescribed variables requires both an understanding of the statistical assumptions of the model *and* a linguistic justification for the experimental design.

In this paper we will sketch out what the New Statistics in Linguistics might look like. We contend that the following goals should be part of this programme

1. a deeper understanding of relevant basic statistical models among linguists;
2. clear instructions for experimental design, including controls, baselines, sampling, abstraction/operationalisation and replication planning;
3. optimal methods for visualising data, plotting confidence intervals and distributions;
4. differentiation between atheoretical exploratory methods and theory evaluation;
5. distinction and commensurability between experimental and corpus results.

References

Cramér, Harald. 1944. *Mathematical Methods of Statistics*. Princeton: Princeton University Press. Available at garfield.library.upenn.edu/classics1983/A1983QW37600001.pdf

Cumming, Geoff. 2014. The New Statistics: Why and How. *Psychological Science* 25(1): 7–29.
Gelman, Andrew & Eric Loken. 2013. The garden of forking paths. Columbia University. Available at www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
Newcombe, Robert G. 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 17: 857–872.
Stigler, Stephen M. 2012. Studies in the history of probability and statistics, L: Karl Pearson and the Rule of Three. *Biometrika* 99(1), 1–14.

Stephanie Hackert

(Re-)Defining the envelope of variation, or what goes in must come out

The past decades have seen a fundamental transformation of the field of (English) linguistics through quantitative approaches such as variationist sociolinguistics or corpus linguistics. Particular progress has been made in the application of statistical methods to linguistic data, with the use of sophisticated tools such as phylogenetic networks, conditional inference trees, or random forests now apparently firmly entrenched among even junior practitioners in the field.

Less attention, it seems, has been paid to the nature of the data itself. Obviously, though, what goes into any statistical analysis crucially affects its outcome. This is not new wisdom; accurate circumscription of the envelope of variation has always been at the heart of the quantitative sociolinguistic enterprise. As early as 1968, in fact, Labov himself pointed out that "even the simplest type of counting raises a number of subtle and difficult problems. The final decision as to what to count is actually the solution to the problem in hand" (1968: 14, quoted in Wolfram 1969: 47). And while variationist sociolinguistics has seen at least one heated debate about "count" vs. "don't count" cases (cf. Rickford et al. 1991; Blake 1997 on the copula in African American Vernacular English), discussion of the input to statistical analyses has not been equally explicit and controversial in other subfields of (English) linguistics.

In the paper, I will present and discuss a number of examples illustrating the difficulties in defining what data exactly should be subjected to statistical analysis, how such data should be categorized and coded, and how different definitions of the variable context affect the outcome of any analysis and the conclusions based on it. The examples include "classic" sociolinguistic variables like past inflection as well as phenomena such as genitive variation or the mandative subjunctive, which have most frequently been treated within a corpus-linguistic framework. I also cover problematic aspects of large-scale typological comparisons of varieties, such as they have become popular in creole studies of late. I conclude with some general suggestions as to how to improve data quality in quantitative linguistics.

References

Blake, Renée. 1997. Defining the envelope of linguistic variation: The case of "don't count" forms in the copula analysis of African American Vernacular English. *Language Variation and Change* 9: 57–79.

Labov, William. 1968. Contraction, deletion and inherent variability of the English copula. Paper given at the Linguistic Society of America annual meeting.

Rickford, John R., Arnetha Ball, Renée Blake & Raina Jackson. 1991. Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Language Variation and Change* 3: 103–132.

Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington: Center for Applied Linguistics.

David Tizón-Couto & David Lorenz

Everything matters, or what to do with all those variables...

The "quantitative turn" in linguistics has brought with it an increasing sophistication of methods of statistical analysis. Yet, the field appears to be in a methodological 'wild west' state where much is possible and new frontiers are being explored, but there is relatively little guidance in terms of firm rules or conventions. This can be exemplified in one of the most common inferential statistical tools, namely regression analysis, and the issue of variable selection.

Relevant text books (e.g. Baayen 2008; Johnson 2008; Gries 2009, 2013; Levshina 2015) typically suggest an Occam's razor approach to regression modelling, such that variables that do not make a sufficient contribution to explaining the variation should be left out of the model – this has been called a "minimal adequate model" (Gries 2009: 296). Most research papers on linguistic variation seem to implicitly or explicitly follow this approach. However, there seems to be little awareness of the other possible approach, which is to pre-select a number of variables (based on hypotheses or previous findings) and test their effects in a model without further selection – this has been called "deductive modelling" for effect estimation (Harrell 2015: 98).

In this paper, we compare the ramifications of both the minimal and deductive approaches to regression modelling by considering their applications to two different datasets: one comes from a corpus study and the other from an experimental study by the authors. The first part illustrates how a model containing a pre-defined set of variables might be useful when these variables are tested on different dependent variables and then compared (e.g. different loci of phonetic reduction in the same morpho-phonological unit). The second part illustrates the differences between the deductive and minimal approaches for a large dataset where response times are the dependent variable. The two methods are replicated on subsets of the experimental data in order to check on the statistical reliability of the models and their differences. Both methods produce valid results, but deductive models are more reliably comparable across replications.

By means of this reflective exercise, the present paper aligns against a strong version of the "bias towards the significant" by making a case for a maximal approach to regression modelling. While a minimal adequate model has the advantage of parsimony, it involves the non-trivial issue of how to select the relevant variables. Deductive models, on the other hand, have higher transparency (all variables get reported) and a greater accuracy of the reported effects (Harrell 2015: 98). We hope to show that they are also, in principle, more appropriate for the scientific endeavor of cumulative knowledge construction: They rely explicitly on prior knowledge, they are replicable on new data sets and they allow for a more precise comparison of effects.

References

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

Baayen, R. Harald. 2013. Multivariate statistics. In Robert J. Podesva & Devyani Sharma, eds. *Research Methods in Linguistics*. Cambridge: Cambridge University Press. 337–372

Gries, Stefan Th. [2009] 2013. *Statistics for Linguistics with R*. Berlin: Mouton de Gruyter.
Harrell, Frank E. 2015. *Regression Modeling Strategies*. Cham: Springer.
Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden: Blackwell
Levshina, Natalia. 2015. *How to Do Linguistics with R*. Amsterdam: Benjamins.

Jorge Aguilar-Sánchez

Power, sample size, and the quantitative crisis

Researchers sometimes, unknowingly, exclude the existence of the phenomenon in the population under study (effect size) to develop investigations that try to explain discrepancies they find in the constructs in previous research. Explanations justifying each study are based on theories; or conjecture related to specific independent variables and how researchers disagree with the construct proposed originally. These disagreements create variants of the original study, but not proper replications of it. This practice produces what Ottenbacher called "a contradictory research literature, apparently dynamic, but failing to resolve uncertainties; therefore, failing to establish statistical consensus" (Ottenbacher 1996: 275).

To resolve these uncertainties, and to produce proper replication, this paper addresses the concerns that arise by not reporting two statistics: power and effect size. Power is the probability of correctly rejecting the null hypothesis when it is false in the population. Power calculations were conducted for studies on linguistic phenomena with dichotomous dependent variables on one phenomenon in the context of multilevel logistic regression. Results show that these studies have a low power due to decisions that were made during their design (i.e. low scientific rigor). Others show low power due to the aggregation of naturally independent data or using the wrong methods to conduct research (i.e. poor research design decisions). The present paper adds to body or work done by Aguilar-Sánchez (2014, 2017, forthcoming a, forthcoming b), Arppe et al. (2010), Gries (2015a; 2015b), Ioannidis (2007) and their search for sound research practices in the study of language.

References

Aguilar-Sánchez, Jorge. 2014. Replicability of (socio)linguistics studies. *Journal of Research Design and Statistics in Linguistics and Communication Science* 1(1): 5-25.

Aguilar-Sánchez, Jorge. 2017. *Research Design Issues and Syntactic Variation: Spanish Copula Choice in Limón, Costa Rica*. Balti: Lambert Academic Publishing.

Aguilar-Sánchez, Jorge. Forthcoming a. Copula + adjective: An a-posteriori power analysis for the generalizability of results. *Journal of Research Design and Statistics in Linguistics and Communication Science*.

Aguilar-Sánchez, Jorge. Forthcoming b. Looking back to move forward: Design issues and best practices in the study of language.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1): 1–27.

Gries, Stefan Th. 2015a. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language & Linguistics* 16: 93–117.

Gries, Stefan Th. 2015b. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1): 95–125.

Ioannidis, John P. A. 2007. Why most published research findings are false: Author's reply to Goodman and Greenland. *PLoS Medicine* 4(6): e215.

Ottenbacher, Kenneth J. 1996. The power of replications and replications of power. *The American Statistician* 50(3): 271–275.

Terttu Nevalainen, Tanja Säily, Turo Vartiainen & Joonas Kesäniemi

Meta-analysis for historical corpus linguistics using the *Language Change Database*

English corpus linguistics provides an early example of open science in the field of linguistic research. The idea behind the creation of the first English digital corpora was for researchers to be able to share data with each other by distributing these resources free of charge. The history of English corpus linguistics extends back to 1964, when the first structured electronic corpus of the English language, the *Brown Corpus*, was published (Francis & Kučera 1964). The first diachronic corpus of English, the *Helsinki Corpus of English Texts*, was published in 1991, and it covered the entire history of the English language from Old English to the Late Modern period. Since then, many other historical corpora have been published, and thousands of articles that make use of quantitative corpus data have been written about the variation and change of the English language.

In this paper, we argue that the variety of research questions studied in the past twenty-five years, and the large amounts of data on which the research is based, can now open new avenues of research, just like the introduction of corpora did decades ago. However, the fact that the information included in research articles is typically not available in a form that would facilitate its reuse is a serious problem both from the perspective of meta-analysis and the accumulation of knowledge. Moreover, many of the early corpus-based studies of language change that are still relevant today were published in edited volumes and festschrifts that can be hard to find even in a well-stocked university library. We therefore suggest that a research database which not only includes detailed information about published articles on the history of English, but also the published data in annotated form, can be used as a basis for the replication of earlier research and for meta-analysis: by making use of existing data, we can explore questions that would otherwise be too labour-intensive to study.

Our paper introduces the *Language Change Database* (LCD), a new linguistic resource currently under compilation, which is designed to facilitate the dissemination, verification and reuse of linguistic research and research data (Nevalainen et al. 2016). The LCD draws together information about hundreds of articles on the history of English, summarizing their results and providing a rich annotation scheme to ensure accurate data retrieval. The LCD also includes numerical data extracted from the articles in an annotated form, which the end users of the database can download and use in their own research (see Figure 1).

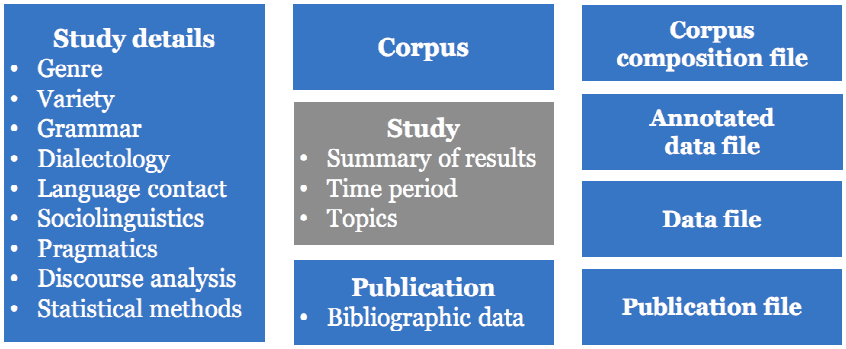

Figure 1. Information included in the LCD.

We will also introduce a tool called LADA (LCD Aggregated Data Analysis workbench), which provides corpus linguists with a systematic workflow to perform exploratory meta-analyses based on earlier research results (Kesäniemi et al. 2018). The LADA workflow re-uses data from the LCD and provides the user with the tools to filter, review and normalize existing data to create a new aggregated dataset, which can then be visualized and exported for further analysis.

References

Francis, W. Nelson & Henry Kučera. 1979 [1964]. *Manual to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers*. Providence: Department of Linguistics, Brown University.

HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Helsinki: Department of Modern Languages, University of Helsinki. www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

Kesäniemi, Joonas, Turo Vartiainen, Tanja Säily & Terttu Nevalainen. 2018. Open science for English historical corpus linguistics: Introducing the Language Change Database. *Digital Humanities in the Nordic Countries, 3rd Conference, 7–9 March 2018, Helsinki.*

Nevalainen, Terttu, Turo Vartiainen, Tanja Säily, Joonas Kesäniemi, Agata Dominowska & Emily Öhman. 2016. Language Change Database: A new online resource. *ICAME Journal* 40: 77–94.