Panel: Corpus statistics – key issues and controversies

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

Gerold Schneider
University of Zürich
gschneid@es.uzh.ch

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Jefrey Lijffijt
University of Bristol
jefrey.lijffijt@bristol.ac.uk

Paul Rayson
Lancaster University
p.rayson@lancaster.ac.uk

Sean Wallis
University College London
s.wallis@ucl.ac.uk

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

1 Motivation

The application of sound statistical techniques in descriptive linguistics is increasingly seen as a vital methodological requirement. However, there are still many studies that fail to carry out a statistical analysis or, more commonly, apply significance tests and other well-established methods in an overly simplistic manner. Typical examples are significance testing of frequency differences with a chi-squared or Fisher exact test instead of multifactorial models; the exclusive use of p-values, disregarding effect size; and the visualization of keywords in the form of word clouds (which are particularly popular in the digital humanities community).

There are various reasons for this problem: researchers may not be aware of an appropriate statistical test, they may not have the tools to execute that test, or it may be an open scientific question which test would be most applicable. Accordingly, there is an urgent need for discussions about the appropriate use of statistics in quantitative linguistic studies, the development of new methods and appropriate software tools, and the dissemination of new methodological findings to the corpus linguistics community.

2 Speakers

The panel discussion brings together researchers who are well known for their research on statistical methodology, their teaching efforts in this area and/or the implementation of relevant software tools. Conference delegates will gain a deeper understanding of key problems and learn about the latest methodological developments.

3 Format and topics

We have defined a list of five key topics for the panel. Two panellists are invited to give position statements on the topic, sketching opposite points of view or suggesting alternative solutions. This is followed by a discussion among panellists. We then invite comments and questions from the audience.

3.1 Experimental design – which factors should we measure?

Recent work has shown that simple frequency comparisons and similar approaches are inappropriate in most cases (e.g. Evert 2006). Instead, multifactorial models could be used (Gries 2006) in order to account in full for the variability of frequency counts and other measures, or the data could be modelled differently (Lijffijt *et al.* 2014). Key questions to be discussed are (i) the unit of measurement and (ii) which predictive factors should be included in the analysis. Regarding the unit of measurement, should studies report and model per-word counts or per-text relative frequencies, or rather predict the outcome of a speaker decision? In the latter case, we base our investigation on an envelope of variation (Labov 1969), such as an alternation, and are potentially less affected by corpus sampling. When selecting a set of predictive factors, we need to strike a reasonable balance between too few, which runs the risk of excluding important factors and thus resulting in an unsatisfactory goodness-of-fit, and too many, which leads to sparse data problems, overadaptation of the model to the data set, and limited scientific insights.

3.2 Non-randomness, dispersion and violated assumptions

"Language is never, ever, ever random" (Kilgarriff 2005). In particular, words and other linguistic phenomena are not spread homogeneously across a text or corpus (Church 2000), their appearance depending on the style and topic of a text as well as previous occurrences in a discourse. As a result, the random sample assumption underlying most statistical techniques is very often violated. For example, the individual texts comprising a corpus have usually been sampled independently, but the word tokens within each text are correlated. Therefore, when using words as a unit of measurement, the independence assumption made by frequency comparison tests and many multifactorial models is violated. We discuss the precise assumptions of different statistical techniques, under what circumstances they are violated, which violations are most harmful, and how this problem can be solved or mitigated.

3.3 Teaching and curricula

Corpus linguistics employs quantitative methods that rely on correct use of different statistical procedures. It therefore necessarily presupposes a certain awareness of statistical assumptions and principles. The question, however, is to what extent corpus linguists (researchers and students) should be able to perform complex statistical procedures such as mixed effects modelling using R or similar software packages. This also raises a number of other questions:

How can we improve the understanding of basic statistics among researchers and in the linguistics curricula? Should statistics courses be compulsory at BA or MA level? And perhaps even an introduction to computer programming? We also report on our personal experiences of teaching the statistics language R to students with no previous programming experience.

3.4 Visualisation

In statistical textbooks, initial visualisation of the data (using scatter plots, box plots, etc.) is often recommended as an important stage of data exploration before statistical tests are applied. Indeed, good visualisation can provide us with a holistic picture of the main tendencies in the data, help to discover interesting patterns, and reveal outliers and other problematic aspects of a data set. In corpus linguistics, different visualisation techniques have been used: word clouds, word trees, collocation networks, bar charts, error bars, etc. (see, e.g., Siirtola *et al.* 2011). Which of these visualisation techniques are helpful for the researcher and the reader? Does visualisation really help the reader to understand a concept and the researcher to detect interesting patterns and crucial zones, on which to focus in further investigations? Is visualisation merely a form of presentation of the data or does it play a more fundamental role in the research process?

3.5 Which models can we use?

There is a large range of statistical models to choose from (e.g. Schneider 2014). In topics 3.1 and 3.2 we have already talked at length about regression models, but alternative, computationally more demanding techniques are also available, such as probabilistic models from natural language processing (taggers, parsers, machine translation, text mining tools, semantic classifiers, spell-checkers) and dimensionality reduction approaches. Both the possibilities and their complexities are vast, making this discussion topic open-ended.

References and select bibliography

Brezina, V. and Meyerhoff, M. 2014. "Significant or random? A critical review of sociolinguistic generalisations based on large corpora." *International Journal of Corpus Linguistics* 19 (1): 1–28.

Church, K. 2000. "Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p^2." In: *Proceedings of the 17th conference on Computational linguistics*, pp. 180–186.

Evert, S. 2006. "How random is a corpus? The library metaphor." *Zeitschrift für Anglistik und Amerikanistik* 54 (2): 177–190.

Evert, S.; Schneider, G.; Lehmann, H. M. 2013. "Statistical modelling of natural language for descriptive linguistics". Paper presentation at *Corpus Linguistics 2013*, Lancaster, UK.

Gries, S. Th. 2006. "Exploring variability within and between corpora: some methodological considerations." *Corpora* 1 (2): 109–151.

Gries, S. Th. to appear. "Quantitative designs and statistical techniques." In D. Biber and R. Reppen (eds.) *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, A. 2005. "Language is never ever ever random." *Corpus Linguistics and Linguistic Theory* 1 (2): 263–276

Labov, W. 1969. "Contraction, deletion, and inherent variability of the English copula." *Language* 45 (4):715– 762.

Lijffijt, J; Nevalainen, T.; Säily, T.; Papapetrou, P.; Puolamäki, K.; Mannila, H. 2014. "Significance testing of word frequencies in corpora." *Digital Scholarship in the Humanities*, online ahead of print.

Pipa, G. and Evert, S. 2010. "Statistical models of non-randomness in natural language." Presentation at *KogWis 2010*, Potsdam, Germany.

Schneider, G. 2014. *Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics*. Cumulative Habilitation, Faculty of Arts, University of Zürich.

Siirtola, H; Nevalainen, T.; Säily, T.; Räihä, K.-J. 2011. "Visualisation of text corpora: A case study of the PCEEC." In: T. Nevalainen and S. M. Fitzmaurice (eds.) *How to Deal with Data: Problems and Approaches to the Investigation of the English Language over Time and Space* (Studies in Variation, Contacts and Change in English 7). Helsinki: VARIENG.