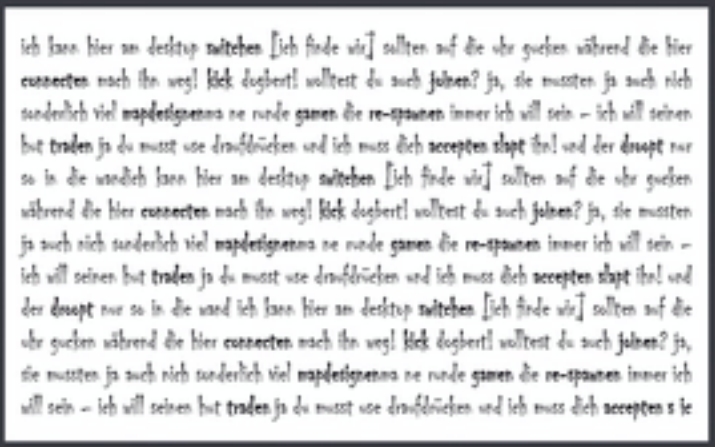
LANGUAGE COMPETENCE AND LANGUAGE AWARENESS
IN EUROPE 4

Markus Bieswanger / Amei Koll-Stobbe (eds.)

New Approaches to the Study of Linguistic Variability



PETER LANG
EDITION

New approaches to the study of linguistic variability

Markus Bieswanger

Scientific interest in the complex and multidimensional issue of linguistic variability is rather young, but has been one of the most rapidly growing research areas in linguistics in recent decades. Despite the undisputed fact that "[v]ariability is everywhere in language" (Wolfram 2006:333), the analysis and description of language variation had for a long time essentially been limited to dialectology in the traditional sense, i.e. to variation related to the geographical background of the speaker. It was only about half a decade ago that the pioneers of sociolinguistics founded a new and internally diverse but now well-established subfield of linguistics devoted to the study, description and explanation of language variation (cf. Bayley, Cameron & Lucas 2013:1). For such a task, largely based on empirical evidence in the form of observable language behaviour, a number of linguistic and social factors have to be considered individually as well as in combination and new methods of gathering and working with this kind of linguistic data have to be developed. In the introduction to *The Handbook of Sociolinguistics*, Coulmas (1997:6) states that "[m]ethodological questions concerning the delimitation, collection, and processing of empirical data have therefore been much more in the foreground than theory construction." Even today, there is an ongoing quest for innovative approaches to the analysis of linguistic variability, which is, for example, impressively illustrated by the fact that the *New Ways of Analyzing Variation* conference is already in its 42nd year (2013) and still going strong. The contributions to this volume entitled *New Approaches to the Study of Linguistic Variability* reflect the continuing need to develop and apply new ways of analyzing language variation and linguistic variability.

The analysis of linguistic variability is based on empirical data and the contributions to this volume are no exception. The first two articles share that they both address methodological issues related to the increasingly important but heterogeneous field of corpus linguistics (cf. McEnery & Hardie 2012:1-3). In "Embracing Bayes Factors for key item analysis in corpus linguistics", Andrew Wilson discusses the suitability of Bayes Factors for the widely used *key item* methodology in contemporary corpus linguistics and suggests that corpus linguists should replace the traditional frequentist p-values by Bayes Factors and make the "Bayesian turn." Martin Schweinberger presents "A sociolinguistic analysis of the discourse marker LIKE in Northern Irish English. A look behind the scenes of quantitative reasoning" and shows that it is necessary to choose and apply different statistical methods carefully when analyzing complex linguistic data sets, in order to unearth hidden patterns.

The following two papers have in common that the contact of English with another language plays a major role in each article. Lillian Kaviti addresses the development and current status of Sheng ('Swahili-English') and Engsh ('English-Swahili') in Kenya in "The evolution of urban hybrid languages in Kenya: The Case of Sheng and Engsh." She also presents an analysis of the role of Sheng and Engsh for urban youth identity in Kenya. Language contact is also one of several aspects included in Melanie Burmeister's article "Variability in death notices from Scotland, Wales and the Republic of Ireland: A comparative perspective," in which she compares English-language death notices published in different newspapers, looking for structural differences.

Mental aspects feature prominently in the next two articles. Florian Dolberg is concerned with "Subject to change: Decay and mutation of linguistic memory" and investigates the relation of verbatim and gist memory and the reliability of mental storage of linguistic information overall. In the following article "New contexts and new concepts: The use of German 'alt'", Judith Rossow turns to cognitive linguistic aspects of meaning construction, paying special attention to the conceptualization and contextualization of the German adjective *alt* ('old') in different situations.

The last two contributions share the fact that the data analyzed was produced by using means of so-called computer-mediated communication (CMC) or electronically-mediated communication (EMC), i.e. both articles fall into the realms of EMC language research, which is "a new and dynamically evolving field" (Herring, Stein & Virtanen 2013:3). In "Apologies and excuses in academic e-mail communication: Differentiation and characterization from a pragmatic perspective," Jana Kiesendahl analyzes language data from one register of e-mail communication, to find out whether apologies and excuses are two different types of speech acts. Caroline Schilling applies the established macro-sociolinguistic factor "age" to text messaging in her paper "Assessing texting behaviour and language use in SMS communication: A survey on age differences."

This brief overview illustrates the wide range of topics and approaches featured in this volume. The collection thus contributes to the growing body of research on linguistic variability, including methodological issues as well as innovative analyses of various text types. The papers presented here showcase a small number of new approaches to the analysis of linguistic variability and at the same time demonstrate that the search for the best ways to analyze language variation and variability is far from over.

References

Bayley, Robert, Richard Cameron and Ceil Lucas (2013): "Introduction: The Study of Language and Society." In: Bayley, Robert, Richard Cameron and Ceil Lucas (ed.) *The Oxford Handbook of Sociolinguistics*. Oxford et al.: Oxford University Press. 1-7.

Coulmas, Florian (1997): "Introduction." In: Coulmas, Florian (ed.): *The Handbook of Sociolinguistics.* Oxford & Malden: Blackwell. 1-11.

Herring, Susan C., Dieter Stein & Tuija Virtanen (2013): "Introduction to the pragmatics of computer-mediated communication." In: Herring, Susan C., Dieter Stein & Tuija Virtanen (eds.): *Pragmatics of Computer-Mediated Communication*. *Handbooks of Pragmatics*, Vol. 9. Berlin & New York: de Gruyter. 3-32.

McEnery, Tony & Andrew Hardie (2012): *Corpus Linguistics: Method, Theory, Practice.* Cambridge et al.: Cambridge University Press.

Wolfram, Walt (2006): "Variation and Language: Overview." In: Brown, Keith (chief ed.): *Encyclopedia of Language and Linguistics.* Volume 13. 2nd edition. Amsterdam et al.: Elsevier. 333-341.

I. Methodological issues in corpus linguistic analyses of variability

Embracing Bayes Factors for key item analysis in corpus linguistics

Andrew Wilson

Introduction

The key item methodology is one of the most widely used tools in modern corpus linguistics. Its goal is to highlight those lexical items – or other linguistic constructs such as part-of-speech categories or semantic fields – which are most distinctive of one text or corpus when compared against another. In other words, it sets out to identify the main elements of variability between two (or sometimes more) varieties, authors, texts, etc. When used in relation to lexical items, it is more commonly known as the *keywords methodology*.

However, this methodology is not without its difficulties. For instance, the uneven dispersions of items across parts of a text or corpus (Leech, Rayson & Wilson 2001; Gries 2008) and the actual magnitudes of any frequency differences that are discovered (Gries 2005) have both been highlighted as complicating factors in interpreting the results of key item analyses. In this short paper, I should like to focus on a more basic misunderstanding in relation to the key item methodology and on one possible solution to it.

Misunderstanding key items

Although it has been given some aura of novelty by the use of terms such as "keyness" in certain software implementations, the key item methodology is actually nothing more than an ordinary null hypothesis significance test applied to the frequencies of words or other items in two texts or corpora. Most commonly, the underlying test is based on a 2 × 2 contingency table of count data along the following lines:

Frequency of word *x* in text A	Frequency of word *x* in text B
Frequency of other words in text A	Frequency of other words in text B

Given such a table, the most frequently used test amongst corpus linguists today is the log-likelihood test (Dunning 1993). This is the default option in most standard programs, such as Wordsmith Tools and Antconc, and is the only test available in the Wmatrix environment for corpus processing (Rayson 2003). However, Pearson's chi-squared test and Fisher's exact test are also sometimes

recommended and used (e.g. Pedersen 1996; Oakes & Farrow 2007). To help correct for problems with dispersion, other authors have suggested instead dividing the two texts or corpora into parts (either natural or artificial) and employing two-sample tests such as the t-test (Paquot & Bestgen 2009) or the Wilcoxon-Mann-Whitney test (Kilgarriff 1996). Unfortunately, these latter options are rarely provided in corpus analysis software: the only implementation that I know of is in the PROTAN software, which allows the computation of t-tests (Hogenraad, Daubies, Bestgen & Mahau 2003).

All of these tests have in common the fact that they produce a test statistic (such as a log-likelihood or t-value) and a corresponding p-value. The p-value tells us the probability of obtaining an equal or more extreme result, given the null hypothesis (Jefferys 1995; Goodman 1999a). If the p-value is very small, then one conventionally infers that either (a) a very rare event has occurred or (b) the null hypothesis is unlikely to be true (Macdonald 2004). However, this p-value is often misinterpreted by corpus linguists as being the actual probability that an observed difference in proportional frequencies between two texts or corpora has occurred by chance (e.g. Culpeper 2009; Gabrielatos, Torgersen, Hoffmann & Fox 2010). This is an extremely widespread, but normally false, belief in all branches of the sciences (Jefferys 1995; Goodman 1999a; Goodman 2008): Carver (1978) calls it the "odds against chance fallacy".[1] A traditional p-value cannot, in general, be the same as the probability of the null hypothesis, because it is a special case of a conditional probability: it is conditional on the null hypothesis being true in the first place (Carver 1978).

Nevertheless, it is hardly surprising that such misunderstandings arise. For one thing, Cohen (1994) has noted that null hypothesis significance testing does not tell scientists what they typically want to know, which is the probability of the null (or some alternative) hypothesis in the light of their data. Interpretations along these lines are thus particularly liable to replicate as "mind viruses" or "memes" (cf. Koch 1986), probably according to a principle of subconsciously perceived utility (Heylighen 1997). More to the point, these misinterpretations of p-values have been taught by numerous textbooks for nearly a century, misleading successive generations of teachers into providing the wrong definitions to their students and replicating the same errors in their own texts and notes (Cohen 1994; Macdonald 2002; Gigerenzer 2004). Statements highlighting the correct interpretation of p-values, contrasted carefully and explicitly with the common misapprehensions, are still a rare event, especially in the sorts of textbooks read by practising scientists and students, where they are most needed.

1 I say normally false, because Altham (1969) - for instance - has shown that the one-tailed p-value of Fisher's exact test is the same as the posterior Bayesian probability of the null hypothesis that the odds ratio is equal to 1, albeit under very conservative prior assumptions. For other exceptions, see also Lindley (1965) and Goodman (1999b).

Nevertheless, whilst it would be possible merely to teach the correct interpretation more forcefully and clearly, this does not really address the issue that Cohen (1994) highlighted, namely that the traditional p-value of a null hypothesis significance test does not tell us what we actually want to know anyway. For this, a different solution is required.

The Bayesian solution

The brand of statistics that is usually applied in null hypothesis significance testing is commonly termed "frequentist". This has been the most dominant approach to statistical analysis during the last century, and still remains so. However, there is another, increasingly popular, approach to statistics, which is known as Bayesian statistics (Berry 1996; Lee 1997; O'Hagan & Luce 2003). It is named after the Reverend Thomas Bayes, an English Presbyterian clergyman who lived from 1701 to 1761 and authored a posthumously published paper on probability theory (Bayes 1763; Bellhouse 2004). In contrast to frequentist statistics, Bayesian statistics focuses on the probability of hypotheses in the light of observed data, rather than on the probability of observed (and more extreme) data in the light of hypotheses. It is thus able to provide answers to the questions we actually want to ask in a key item analysis in corpus linguistics – i.e., what is the probability that a particular difference in frequency has occurred by chance. Bayesian statistics is sometimes criticized on the grounds of subjectivity, because it involves the initial determination of a prior (or hypothetical) probability which is then updated, using Bayes' Theorem, by a posterior probability based on the actual data that are observed (O'Hagan & Luce 2003). However, in an *objective* Bayesian framework (Lindley 1965; Berger 2006; Robert, Chopin & Rousseau 2009), these prior probabilities are drawn from a very limited default set, thus obviating any such charges of subjectivism.

Within this Bayesian framework, the so-called Bayes Factor is a measure of the amount of evidence provided by a test against the null hypothesis. Although by no means as widespread as frequentist p-values, Bayes Factors are now becoming more commonly encountered in several fields, especially genetics (cf. Sawcer 2010). The Bayes Factor is related to the Bayesian prior and posterior probabilities as follows (Goodman 1999b: 1005), where the 'odds' are given by the probability divided by one minus the probability ($P/(1 - P)$):

$$\text{prior odds of null hypothesis} \times \text{Bayes Factor} = \text{posterior odds of null hypothesis}$$

Because they involve integrals, true Bayes Factors can be difficult to calculate, but Raftery (1986) and Kass and Raftery (1995) have provided very simple approximations – using the so-called Bayesian Information Criterion (or BIC) as

an approximate Bayes Factor – for some of the most commonly used null hypothesis significance tests.

For a log-likelihood statistic with one degree of freedom, the approximate Bayes Factor (BIC) is given by:

$$BIC \approx LL - \log(N)$$

where LL = the log-likelihood statistic and N = the size, in running words (or tokens), of the two corpora combined. Where there is more than one degree of freedom (df), the approximate Bayes Factor is instead given by:

$$BIC \approx LL - (df \times \log(N))$$

For a t-test, the approximate Bayes Factor is:

$$BIC \approx t^2 - \log(N)$$

where t = the t-value given by the test and N = the total number of texts or segments into which the two corpora have been divided for testing purposes.

These approximate Bayes Factors can then be converted into degrees of evidence against the null hypothesis (H0), as shown in Table 1 (based on Kass & Raftery 1995: 777).

Table 1: Degrees of evidence against the null hypothesis (H0)

Approximate Bayes Factor (BIC)	Degree of evidence against H0
0-2	not worth more than a bare mention
2-6	positive evidence against H0
6-10	strong evidence against H0
>10	very strong evidence against H0

Approximations for other tests (including Pearson's chi-squared and the Wilcoxon-Mann-Whitney tests) can be found in Johnson (2005) and Yuan and Johnson (2008).

Worked example

In this section, I demonstrate the effect of using Bayes Factors rather than p-values in a keywords analysis. It is not my aim here to list and discuss the actual keywords derived from the data, but merely to give an impression of how using the Bayes Factor affects their number and their subdivision into different levels of evidence strength.