A paper presented at the *Asia Pacific Corpus Linguistics Conference* in Auckland (NZ) in February 2012.

Significant or Random? A Critical Review of Statistical Analyses in Corpus-Based Sociolinguistic Studies

Vaclav Brezina[1] and Miriam Meyerhoff[1]

[1]*Department of Applied Language Studies and Linguistics*
The University of Auckland, New Zealand

Keywords**:** sociolinguistics, corpus linguistics, statistical procedures, social variation, meaningful variation

Introduction

This paper offers a critical review of a methodology often employed in corpus-based sociolinguistic studies (e.g. Macaulay, 2002a, 2002b; McEnery & Xiao, 2004; Xiao & Tao, 2007; Barbieri, 2008, 2009; Murphy, 2010), which make use of aggregate data. This methodology relies on a general comparison of frequencies of a target linguistic variable in socially defined sub-corpora (e.g. speech of all men vs. speech of all women in the corpus). An issue with this procedure lies in the fact that it emphasises the inter-group differences and ignores within group variation. The methodology thus often yields falsely positive results (with highly significant log likelihood scores). The paper presents evidence which shows that sociolinguistic studies based on aggregate data are in principle unreliable. We will demonstrate that random (and therefore sociolinguistically irrelevant) speaker groupings can often yield statistically significant results.

Data and methodology

The analyses are based on *BNC32*, one- million-word corpus of British informal conversation, which was extracted from the demographic part of the *British National Corpus* (*BNC*). It represents the speech of 32 British English speakers – 16 women and 16 men. The speakers form a balanced sample in terms of gender, age and socioeconomic status.

The following table (Table 1) summarises the main features of the corpus. The major advantage of *BNC32* is the fact that (unlike the majority of commonly used corpora) it enables us to search for language forms in the speech of individual speakers.

Table 1 Structure of BNC 32

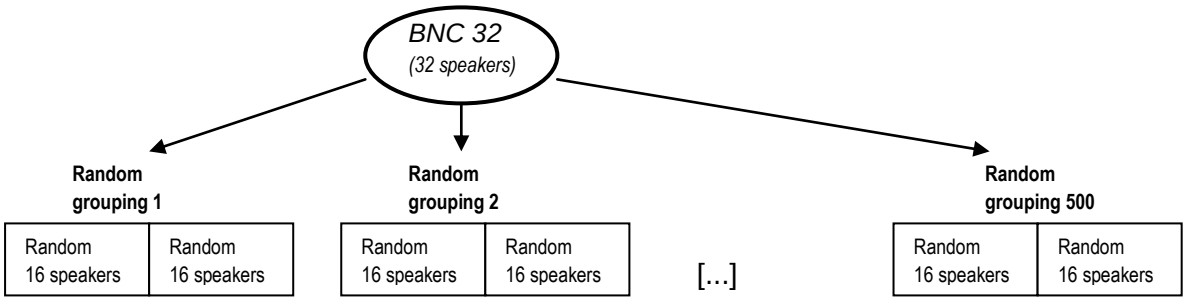
Corpora	Tokens (running words)	No. of speakers	Speaker's gender	Speaker's age	Speaker's socio-economic status	Genre	Discourse mode	Variety of English	Period
BNC 32	1.04 mil.	32	16 male 16 female	15-34: 10 35-54: 13 55+: 9	AB: 7 C1: 9 C2: 9 DE: 5 unknown: 2	informal conversation	highly interactive	UK	early 1990s

The first step of the analysis was to compare the occurrence of seven linguistic forms in *BNC 32* subcorpora based on gender, age and socio-economic status. This comparison was based on aggregate

data with the use of log-likelihood statistic, which is a common procedure in the corpus comparison studies (Rayson et al., 2004).

The second step was to apply the same methodology to random speaker groupings. *Random Integer Set Generator* (www.random.org) was used to randomly assign the 32 speakers into 500 group pairs (see Figure 1). This sampling procedure was repeated three times with the final number of 1,500 random assignments. For every pair of random speaker groups, the log-likelihood score was calculated comparing the occurrences of seven dependent variables. Finally, the percentage of statistically significant differences between the pairs of random groups was calculated for each of the dependent variables.

Figure1: BNC 32 - Random grouping



Findings

Consistent with many previous sociolinguistic studies, the results based on aggregate data show an effect for all social variables in question (speaker's age, gender and sociolinguistic status). At the same time, however, a large proportion of randomly created groups of speakers differ significantly from other random groups. Table 2 below presents an overview of percentages of statistically significant results in the three 500 random groupings for seven dependant variables. We can see that with frequent linguistic variables such as the classical hedge *sort of*, the epistemic phrase *you know* or the definite article, almost 80 per cent of comparisons of random speaker groupings show statistically significant differences.

Table 2. Random variation in BNC 32: per cent of statistically significant results (p<.05)

Grouping	*kind of*	*sort of*	*the*	*I @think you*	*I@think I*	*you know*	*ain't*
random 500 I	45.4%	81.4%	80.4%	34.8%	8%	76.2%	76%
random 500 II	44.4%	78.2%	79.2%	40.8%	5.6%	75.6%	74.8%
random 500 III	43.4%	77.8%	79.8%	37.6%	4.8%	82.2%	75.2%
MEAN	44.4%	79.1%	79.8%	37.73%	6.13%	78.00%	75.33%

References

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. Journal of sociolinguistics, 12(1), 58-88.

Barbieri, F. (2009). Quotative be like in American English Ephemeral or here to stay? English World-Wide, 30, 68-90.

Macaulay, R. (2002a). Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. Journal of sociolinguistics, 6(3), 398-417.

Macaulay, R. (2002b). You know, it depends. Journal of pragmatics, 34(6), 749-767.

McEnery, A., & Xiao, Z. (2004). Swearing in modern British English: The case of fuck in the BNC. Language and literature, 13(3), 235-.

Murphy, B. (2010). Corpus and Sociolinguistics: Investigating age and gender in female talk. Amsterdam: John Benjamins.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon & A. Dister (Eds.), Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004) (pp. 926-936). Louvain-la-Neuve: Presses universitaires de Louvain.

Xiao, R., & Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. Sociolinguistic studies, 1(2), 241-273.