Methodological Considerations in the Determination of Corpus Size for the Study of Frequent Multi-Word Units (MWUs) in Spoken Language

Dahlmann Irina[1]

Abstract

The question of the 'optimum corpus size' in corpus linguistics is not straightforward. It is mainly 'determined by the research question the corpus is intended to address as well as practical considerations' (McEnery et al. 2006:73), in other words the 'right size' has to be assessed time and again for different research questions and circumstances.

Especially for spoken data, practicality is one of the overriding criteria when considering corpus size, as the collection and transcription of spoken data requires substantial time and effort. The question of *How big is big enough?* is therefore of major importance.

In this study we will test a methodological approach to the question of 'sufficient' corpus size for the purpose of investigating frequent MWUs in two spoken corpora, a corpus of native speaker English language (400,000 words) and a corpus of English learner language (200,000 words).

The main issue which will be addressed is the size of the corpus and the effects on the stability of results when using different extraction methods of MWUs. This issue will be discussed in terms of a wider research project in which the use of MWUs by native speakers and by language learners is being compared. The two kinds of automatic extraction methods for MWUs that will be considered are n-grams and Wmatrix (Piao et al. 2003), both of which have been applied to successively increased portions of each of the two corpora. Preliminary results suggest that there is not only a marked difference in stability between the results generated with the two different methods but also between the results from the native speaker corpus and the learner corpus.

These findings will be discussed in relation to methodological issues of analysing native versus non-native corpora, and corpora of different contextual composition.

References

McEnery, T., R. Xiao and Y. Tono (2006) Corpus-Based Language Studies. London: Routledge.

Piao, S.L., P. Rayson, D. Archer, A. Wilson and T. McEnery (2003) Extracting Multiword Expressions with a Semantic Tagger. In: Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 12, 2003, pp. 49–56.

[1] *e-mail*: aexid@nottingham.ac.uk