Abs-213

Gaëtanelle Gilquin (F.N.R.S. – Centre for English Corpus Linguistics, University of Louvain) and Sylviane Granger (Centre for English Corpus Linguistics, University of Louvain)

The use of discourse markers in corpora of native and learner speech: from aggregate to individual data

Traditionally, corpora have been treated by corpus linguists as 'one big text', in which "the data obtained from (...) different speakers or writers are pooled" (Rietveld et al. 2004: 350). This approach relies on the assumption that well-sampled corpora allow for reliable and valid generalisations about a population as a whole (see Kennedy 1998: 74). However, the fact that "corpora are inherently variable internally" (Gries 2006: 110) suggests that, while generalisations about populations may still be valid and useful, interesting findings are also likely to emerge if we investigate corpus data as a series of individual texts rather than as an aggregate. This is particularly true of learner corpora, because of the "highly heterogeneous nature of learner language" (Granger et al. 2009: 3; see also Durrant & Schmitt 2009: 168).

Following recent studies like Paquot (2010) which take account of the possible variance within a corpus, we will adopt both a global and individual approach to the study of discourse markers in native and non-native speech. Our data will come from the newly published Louvain International Database of Spoken English Interlanguage (LINDSEI, see Gilquin et al. 2010) and its native counterpart, the Louvain Corpus of Native English Conversation (LOCNEC, see De Cock 2004). Starting from the whole of LINDSEI and LOCNEC, we will show how the use and frequency of discourse markers such as 'you know' or 'I mean' differ in native and non-native English. This level of analysis reveals, for instance, an underuse of 'sort of' in non-native English as compared to native English. At the next level of analysis, we will make a distinction between the different learner populations represented in LINDSEI, that is, the groups of learners who share the same mother tongue. This will enable us to highlight features that seem to be transfer-related, such as for example the heavy overuse of 'in fact' in the French component of LINDSEI. Finally, we will consider individual speakers in LINDSEI and LOCNEC in an attempt to identify idiosyncratic features that are limited to just a few speakers. By adopting this threefold level of analysis, we hope to shed new light on the use of discourse markers by native speakers and learners of English and to distinguish between the characteristics that are typical of native or non-native speech in general, those that are limited to certain populations and those that are only found among certain speakers. More generally, we wish to advocate for a combined approach in (learner) corpus research which takes into consideration both the pooled data and the individual texts making up a corpus.

References

De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. Belgian Journal of English Language and Literatures (BELL), New Series 2: 225-246.

Durrant, P. & N. Schmitt. 2009. To what extent do native and non-native writers make use of collocations? IRAL 47: 157-177.

Gilquin, G., S. De Cock & S. Granger. 2010. The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (eds). 2009. International Corpus of Learner English. Handbook and CD-ROM. Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.

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

Kennedy, G. 1998. An Introduction to Corpus Linguistics. Longman: London & New York.

Paquot, M. 2010. Academic Vocabulary in Learner Writing: From Extraction to Analysis. London & New York: Continuum.

Rietveld, T., R. Van Hout & M. Ernestus. 2004. Pitfalls in corpus research. Computers and the Humanities 38: 343-362.