THE INTEGRATION OF CORPORA AND LANGUAGE LEARNING

Anvarova Sarvinoz Jumanazar qizi.
sarvinozanvarova97@gmail.com
Mirzo Ulug‘bek nomidagi O‘zbekiston milliy universiteti magistranti

Annotation: In this article I wish to examine the relationship between corpus linguistics andlanguage learning and provide an overview of the most important pedagogical applications of corpora. *A key area to highlight in this context is that of language teaching, where the latest findings from corpus research have led to real innovations in material design and classroom practice.*

Key words: *syllabuses, dictionaries, concordance, authenticity, corpora, Web as a corpus, context, spelling.*

While corpus linguistics has enabled better descriptions of language in use, its real impact lies in the enhancement of applications based on those descriptions. A key area to highlight in this context is that of language teaching, where the latest findings from corpus research have led to real innovations in material design and classroom practice. There are two main areas in which corpora can benefit language teaching and learning: first, by incorporating the latest corpus-based findings into language syllabuses, teaching materials and dictionaries; second, by encouraging teachers and learners to examine language patterns in corpus as part of their (independent) learning activities in and outside classrooms (see Gavioli and Aston 2001).

Corpus linguists and language teaching researchers are often found collaborating in these two areas and there are now publications on the subject. Some of these (e.g. Meunier and Granger 2008) provide further corpus-based descriptions of aspects of language which target the needs of specific groups of language learners, e.g. ESP/EAP learners or learners of the same L1 background. Others (e.g. Hunston 2002; Sinclair 2003) aim to equip teachers and learners with the skills of

concordancing and extracting useful information from concordance lines. Other publications (e.g. Tribble and Jones 1997; O‘Keeffe et al. 2007) include practical suggestions on the various ways in which corpus research can be introduced into the language classroom to enrich the experience of language learners.

Despite the growing interest in the pedagogical applications of corpus linguistics, there have been a number of debates relating to the place of corpus linguistics in language teaching (see Sinclair 1991b; Widdowson 1991; Seidlhofer 2003). Widdowson (1991), for instance, argues that the fact that a language pattern is particularly frequent in a corpus does not necessarily mean that it should take priority in the language teaching syllabus. Further discussion centres around the issue of authenticity and whether it is beneficial to present learners with authentic, real language in use (see McCarthy and Carter 1995; Carter and McCarthy 1996; Prodromou 1996a, 1996b, 1998). According to Prodromou (1996b), it is a ‘fallacy’ to assume that real language is spontaneously interesting and useful to foreign language learners. He argues that train timetables, advertisements, letters published in British newspapers and consumer leaflets are only real to members of the speech community that these texts target. When such data are used as teaching material in a foreign language classroom, they mean very little to the language learners because they lack the same reality for this specific audience. Prodromou (1996a) suggests that an ‘authentic’ discourse has its ‘here and nowness’, and when the discourse is presented in a context that is detached from the ‘here and now’ it automatically loses its authenticity. Similarly, Widdowson (2000) argues that the language presented in a corpus is decontextualised and only partially real. If the decontextualised language in a corpus is to be presented to learners as language in use, it has to be recontextualised. Yet, the reconstituted context is not always the same as the original context of the texts (see Prodromou 2008).

Despite these arguments, corpus data are increasingly becoming an accepted and desirable basis for the development of English language teaching materials, and most major dictionaries and grammars now advertise the fact that they are based on ‘real’ language from a corpus. **The Web as corpus** Today, corpus size has long exceeded the one million word standard set by the Brown Corpus in the 1960s. The Cambridge International Corpus (CIC), which collects spoken and written texts of American English, British English and learner English, is currently one of the biggest corpora of English, with over a billion words. However, with the advent of the world-wide Web we now have access to language data which far exceeds even the most substantial corpus.

Kilgarriff and Grefenstette (2003) suggest that checking spelling and usage of a word by typing it into an Internet search engine is a practical example of how the World Wide Web is already being used as a language corpus on a daily basis by a large number of people. They give the example of ‘speculater’ and ‘speculator’. A search engine reveals that these two spellings generate, respectively, 67 hits and 82,000 hits on the Web. Therefore, based on the higher frequency of occurrence of ‘speculator’, one may conclude that this is the preferred spelling.

However, for the Web to provide more than free, instant suggestions on spellings, corpus linguists have developed Web-based interfaces that allow researchers to use the Web as a compatible resource for linguistic research. WebCorp, for example, allows users greater control over the type of texts to be searched. They can specify the register, textual domain, topic range, date of modification and so on. These facilities support investigations into both synchronic and diachronic changes in language (see Renouf 2003; Renouf et al. 2007). Another advantage of using WebCorp over general Internet search engines in lexical research is that the former offers basic statistical information, including the collocational profile of search items and the option to disambiguate polysemous items (Renouf

1993). The WebCorp interface can also be used to generate frequency lists of Websites specified by the user. It is clearly a valuable resource to use in its own right, but it can also be used to complement research on finite corpora in terms of the up-to-date evidence of language in use that it offers.

One of the main impacts of new technology on the area of corpus linguistics is no doubt the use of the Web as a corpus. In addition, there have been significant advances in spoken corpus linguistics which have been afforded by the alignment of different modalities with a transcript. This development started with the alignment of audio recordings with transcripts, and has recently been extended to include video data as well. It has long been pointed out by corpus linguists working with spoken data that the lack of audio and video leads to problems in the analysis of this kind of corpus data. De Cock (1998), for example, in a discussion of the sequence ‘you know’, argues that it is virtually impossible to decide whether ‘you know’ has a literal or a formulaic meaning on the basis of the orthographic transcript alone. Similarly, Lin and Adolphs (2009) observe that it is not possible to determine the functions of some instances of ‘I don’t know why’ in context unless one can refer to their prosody. Similar concerns arise from a corpus-based analysis of multimodal written texts, i.e. those containing images and graphics.

While the World Wide Web is a very large repository of naturally occurring language, further research is needed as to the type of language that is being used on the Web, what it represents, and how balanced it is in the context of a particular research question. Given the ubiquity of Internet-based and Internet-stored discourse, this endeavour becomes particularly urgent.

References:

1.Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004) ‘Applying corpus linguistics in a health care context’, Journal of Applied Linguistics 1(1): 9–28.

2. Biber, D. and Conrad, S. (1999) ‘Lexical bundles in conversation and academic prose’, in H. Hasselgard and S. Oksefjell (eds) Out of Corpora: Studies in Honour of Stig Johansson, Amsterdam: Rodopi.

3. Danielsson, P. (2003) ‘Automatic extraction of meaningful units from corpora: a corpus-driven approach using the word stroke’, International Journal of Corpus Linguistics 8(1): 109–27.

4. Kilgarriff, A. and Grefenstette, G. (2003) ‘Introduction to the special issue on the Web as corpus’, Computational Linguistics 29(3): 333–48.

5. Sinclair, J. McH. (1991a) Corpus, Concordance and Collocation, Oxford: Oxford University Press.

6. Widdowson, H. G. (1991) ‘The description and prescription of language’, in J. E. Alatis (ed.) Linguistics and Language Pedagogy: the State of the Art, Washington, DC: Georgetown University Press.