

IMPORTANCE OF CORPORA IN TEACHING FOREIGN LANGUAGE

Lapasova Obida Abduraimovna

Samarkand city 48th comprehensive school

Supervisor: Makhmudova Manzura Adashovna

Annotation: *A simple yet important role of corpora in language education is to provide more realistic examples of language usage that reflect the complexities and nuances of natural language. In addition, however, corpora may provide data, especially frequency data, which may further alter what is taught.*

Keywords: *corpora, advantages, English, teaching.*

Applications of corpus linguistics to language teaching began in the late eighties and early nineties. Examples of early work are Higgins and Johns (1984), Higgins (1988), Johns (1988, 1991), Tribble and Jones (1990), Stevens (1991), and J. Flowerdew (1993a). Most work in this area, as in other areas of applied linguistics and language teaching, has focused on English. However, some examples on other languages are Wichmann (1995), Ahmed and Davies (1997), Dodd (1997), King (1997), Kennedy and Miceli (2001), Rule et al. (2003), Belz (2004), Rule (2004), and Bolly (2005). Previous overviews of the field are Leech (1997), Aston (2001), Biber and Conrad (2001), Bernardini (2004), and Stubbs (2004).

One of the reasons for the relatively slow rate of classroom application has been the limitations of the technology. However, as Leech stated already in 1997, “computers have grown smaller, cheaper, and massively more powerful”.

A corpus is a large database of language. Although the first corpora were relatively small - the Brown corpus (developed at Brown University, USA in the early 1960s) consisted of one million words - there now exist corpora consisting of hundreds of millions of words (e.g., the British National Corpus (BNC), 100 million words; and the Bank of English (COBUILD at Birmingham University, UK), over 500 million words). At the same time, however, much smaller corpora with as few as 100,000 words or less are being created all the time for specialist applications. It should be borne in mind, however, that, as pointed out by Gavioli and Aston (2001, p. 238), even the very large corpora contain less language than the average user will have experienced in their daily life.¹ In addition, the linguistic content of corpora is different from what is experienced by individuals in real life, many of them consisting largely of written language. Furthermore, while each text is given equal weighting in a corpus, in real life some texts will hold more value and be experienced more times than others (poetry and religious texts, for example, might be highly valued and heard or read many times). While some corpora are kept in a “raw” state (e.g., Bank of English), many are “tagged” (i.e., coded, according to parts of speech) and “parsed” (i.e., analyzed for grammatical structure) (e.g., BNC).

As already mentioned, corpus techniques have created new knowledge about the behaviour of lexis, grammar, semantics, pragmatics, and textual features. Because corpus

linguistics is based on the theory that language varies according to context – across space and time – the potential for finding out new facts about language is infinite. If this theoretical insight is applied to pedagogy, then the case for the use of corpora in teaching becomes very powerful. Because no dictionary or grammar is able to fully describe the language, the educationist, whether materials designer or classroom practitioner – or indeed learners themselves – may play an important role in identifying regularities in the language which are not to be found in such texts.

At the most basic level, the corpus can provide word lists organized either according to frequency or alphabetically. Used in conjunction with the concordancer, frequency is not limited to the word forms, but may extend to the different meanings of a given word or phrase; the editing function of the concordancer can be used to group the items according to the different meanings. Frequency data can also be obtained for recurrent sequences (variously referred to as n-grams, pre-fabs, and lexical bundles) e.g., *I don't know, all of a sudden, all over the place, don't have a clue*. Furthermore, relative frequencies between two or more corpora can be calculated, those words occurring significantly more frequently in one corpus than another being referred to as keywords.

(Semantic field theory, which can be seen as an introspective precursor of semantic preference, has been applied (mostly intuitively) in language teaching for a very long time (Corder, 1973, p. 316). Indeed it can be seen as closely related to situational (“at the post office,” “at the airport,” “in the supermarket,” “in the office,” etc.) and topical (“travel,” “shopping,” “family,” etc.) syllabuses. It is also implicitly applied in notional syllabuses (Wilkins, 1972). The assumption here is that certain lexical (and grammatical) items belonging typically in given fields are likely to co-occur and can be learned together in semantic sets. However, a corpus approach takes us beyond introspection to identify empirically established relationships. The choice of corpus here is crucial, larger corpora being more reliable, because smaller corpora will not be likely to provide enough data to determine general preferences. On the other hand, specialist corpora consisting of specific genres or registers have great potential for application to LSP.

One of the problems with applying corpora to language teaching is deciding which the most appropriate corpora are. As Leech (1997, p. 18) has pointed out, “the corpora which are easiest to compile are not necessarily those which are most useful for language learning purposes.” Not all corpora will be suitable for all learners.

Until recently, the most pressing problem in this area was the dearth of spoken corpora, most corpora being wholly or primarily made up of written language. The reason for this is simple. It is difficult and expensive to collect spoken language, which then has to be recorded and transcribed.

The foregoing suggests that corpora made up of different language varieties might be needed. Hong Kong learners or Filipino learners, it might be argued, should have as their target educated Hong Kong or educated Filipino English, not British English. Similarly, it would seem sensible that learners of French in Canada might want a standard and hence a corpus based on Canadian French rather than the metropolitan variety. The problem is being

addressed to a degree, for English, with the ICE corpora, referred to earlier, a suite of corpora of 15 different national/regional varieties, such as Australian English, British English, East African English, Filipino English, Indian English, etc. Given the complexity of coordinating and collecting such a range of corpora, however, it is perhaps understandable that these corpora are relatively small, at one million words each. Of course, the question of what standard to adopt is itself controversial.

Further confusing the picture as regards suitable corpora, there are other learner differences that need to be taken into account. For example, a model for young learners might be child language, teenagers may want a teenage model, women and men might want models of the speech of either gender; then again, learners may want specific academic or professional language (see below on corpora and LSP). It is true that there are different types of corpora or sub-corpora (for example, the CHILDES corpus of children's speech (MacWhinney & Snow, 1991) and the British National Corpus has a section on young people's spoken language, referred to separately as the COLT corpus (Stenström et al., 2002)). However, these corpora or sub-corpora have not been designed with language teaching specifically in mind and their suitability, certainly in terms of their size and representativeness, might be questioned.

REFERENCES:

1. Ahmed, K. & Davies, A. (2017). The role of corpora in studying and promoting Welsh. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 157-72). London: Longman.
2. Bernardini, S. (2014). Corpora in the classroom. An overview and some reflections on future developments. In J. M. Sinclair (ed.), *How to use corpora in language teaching* (pp. 15-36). Amsterdam/Philadelphia: John Benjamins.
3. Bianchi, F. & Pazzaglia, R. (2017). Student writing of research articles in a foreign language: Metacognition and corpora. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years on* (pp. 259-87). Amsterdam/New York: Rodopi.
4. Gavioli, L. (2011). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (ed.), *Learning with corpora* (pp. 108-37). Houston: Athelstan.
5. Gavioli, L. (2015). *Exploring corpora for ESP learning*. Amsterdam/Philadelphia: John Benjamins.
6. Leech, G. (2017). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 1- 23). London: Longman.