

# **Chapter 1. Corpus linguistics and its connection to**

## **ELT**

This chapter acquaints the reader with the field of corpus linguistics by presenting a brief overview of its history, aims and areas of application. In the second part, mechanisms of language learning are explained from the perspective of recent linguistic research and different corpus-based methods language instruction are described.

### **1.1 A short history of corpus linguistics**

Since the invention of the computer was strictly aimed at mathematical purposes, the idea of processing text files came as a surprise and a revolution to the linguistic world. The most radical difference in comparison to paper work was in management. A computer can easily remember the contents of a whole library, enabling access and non-problematic changes in both the index and the works. Such possibilities must not have been overlooked and ignored. Therefore in the sixties a new branch of linguistic research emerged treating the computer as a basic tool. The primary concern of researchers was creating a well-structured corpus that could later be dwelled into. The created corpora were supposed to represent proportionally the qualities of the language as a whole ('general' or 'balanced' corpora – Aston and Burnard 1997: 5) or a more specific part of language, e.g. the spoken language, a specific dialect or a professional jargon. Corpora could be used for e.g. lexicography, stylistic, semantic, pragmatic or structural analysis of the language, contrastive studies, and, most importantly (in the context of this paper), teaching and learning methodology.

The most significant initiative of the eighties, which since its inception emphasised the importance and potential of computer corpora and concordancers, was the COBUILD Dictionary project set up by Birmingham University and Collins publishers and directed by John Sinclair (*Collins COBUILD English Language Dictionary* 1987). The product of the project was a dictionary entirely based on corpus evidence.

## **1.2 Methods of exploring a corpus**

As one of the most convenient ways of retrieving data from a corpus one can mention concordancing. “A concordance is a list of occurrences of a particular word, part of a word or a combination of words in its contexts drawn from a text corpus. The search word is sometimes also referred to as key word” (Kettemann 1997: 63).

The history of concordancing is much longer than the history of corpus linguistics itself and dates back to medieval times. The purpose of original concordancing, performed by monks in the Middle Ages, was to create large indexes of most commonly analysed texts (e.g. the Bible), containing sorted lists of all non-function words of the original work, together with contexts of appearance (Tribble and Jones 1997: 1-2). With the discovery of computers’ potential in the field of linguistics, computers were quickly taught to take over the arduous task of concordancing, which has become one of the most powerful tools of language research.

Methods other than concordancing rely mostly on statistical data, without analysing the text-driven hits one by one. Though with such processing vast amounts of linguistic information are lost and cannot be further analysed, statistical tools are the only reasonable ones when the research is based on large corpora (e.g. ‘monitor corpora’ with a continuous data flow – Sinclair 1991: 26). Moreover, statistical analysis of corpora allows for noticing linguistic patterns and phenomena by “embodying a view of the language which is beyond any one individual’s experience” and by “providing objective evidence of frequency” (Aston and Burnard 1997: 6). Therefore many linguists conduct their research in two steps, first analysing the statistical data and only then exploring a selected set of instances retrieved by a concordancer.

## **1.3 Lexicogrammar**

The introduction of concordancing as a common tool of linguistic research changed the traditional view of the internal structure of human language i.e. ‘the open-choice principle’ (Sinclair 1991: 109), which claimed that grammar provides empty skeletons of utterances, later filled with appropriate lexis in the course of discourse formation. This view implies the rigidity of the deep structure but allows for endless possibilities of component substitution, i.e. ‘slot-filling’. Recent research undermined this theory, providing evidence for an intuitive observation that certain collocations are rather fixed,

though they are not necessarily idioms (in the traditional meaning of the word), and that certain vocabulary and grammar items often co-occur in an idiomatic manner. “Words do not occur at random in a text and the open-choice principle does not provide for substantial enough restraints on consecutive choices” (Sinclair 1991: 109). The new, schematic approach introduces the notion of lexicogrammar and combines the previously separate fields of grammar and vocabulary. It sees utterance production “as exploiting ready-made memorized building blocks or ‘pre-fabs’, put together using simpler ‘jerrybuilding’ operations” (Aston 1995: 261). It implies that a speaker is equipped with a set of schemata, rather than two separate sets of grammar rules and lexis, and that the process of learning can be seen as approximating the observed patterns to form the schemata. A vital question arises – how to optimise students’ effort put into language learning with the use of these theoretical findings.

#### **1.4 How to use concordancing for ELT purposes**

Here emerges a not yet fully explored field of CALL research, which formed itself on the border of corpus linguistics and methodology. This fairly new, unnamed discipline tries to find new ELT methods that effectively and actively make use of corpora. There are two prevailing tendencies in this field of study, the first being the so-called ‘COBUILD approach’, which treats the teacher as the primary user of corpora and a filter of corpus data which reaches the student. In the second practical application of corpora – ‘the data-driven learning approach’ – “the corpus primarily constitutes a resource for the learner, either via printed concordances or hands-on concordancing of corpora” (Aston 1997: 51).

##### **1.4.1 The COBUILD approach – the corpus for the teacher**

Teachers may want to treat concordancers as reliable sources of relevant and up-to-date linguistic information, which can be found more quickly than in a dictionary and accompanied by numerous examples of use.

A concordancer’s output can be used as authentic teaching material and can be either selected and pre-processed by the teacher, or it can serve as ‘raw’ material, which can be evaluated and analysed by learners themselves during in-class concordancing.

Another option for the teacher is to prepare his own teaching material based upon corpus data. By editing and rearranging the concordancer’s output, she/he may thus design

several types of language exercises. Kamińska (2001: 9-10) mentions such exercises as gap-filling (where students guess a questioned word from context) or guessing and developing a context (students try to reconstruct full sentences from fragments). Furthermore, corpora can enable the construction of exercises which focus on word-formation or which improve students' writing skills by introducing, e.g., various stylistic figures or discourse markers. Yet other options are to create semantic exercises (students focus on the meaning of the words) or literary-oriented exercises (students may compare different literary works or genres in terms of style) (Kamińska 2001: 10).

Learner corpora created from collections of students' essays or transcribed utterances may well serve as a diagnostic tool. By exploring such corpora the teacher may objectively assess the students' correctness and find the most typical mistakes in the use of both grammar and vocabulary that can be dealt with during the lessons through additional explanations and carefully prepared drills.

#### **1.4.2 The DDL approach – the corpus for the learner**

This approach puts emphasis on enhancing students' creativity, cleverness and analytical skills, since each student has to find solutions on his own or notice a prevailing pattern in the provided data without the teacher's help. It can also be argued that such learning through exploration makes the language information easier to remember and accelerates students' progress.

The idea requires, however, all students' in-class access to computer hardware equipped with concordancing software, which may be a considerable obstacle. Moreover, concordancing sessions may be time consuming providing that every student is expected to reach the target conclusion or solve a problem entirely by himself, especially if the teacher does not actively participate and give hints.

The student can be allowed to browse freely through the corpus and by inventing his own queries for the concordancer he/she may infer some reoccurring patterns or check some of his/her earlier presumptions concerning grammar or usage ('serendipity learning' – Kamińska 2001: 12).

Students may be asked to analyse their mistakes with the use of the concordancer or to work with the corpus to solve linguistic problems. This second application can be further divided depending on who has the dominant role in the reasoning process, i.e. whether the

teacher provides students with numerous hints and poses questions to be answered, or allows for students' initiative in formulating problems.

CALL software is supposed to substitute for the teacher and serve as helping tools for the student. Due to technical problems, however, typical CALL programs still do not effectively fulfil their didactic purpose.

Most CALL programs tend to be based on pre-designed modules, in which the user (learner) may be permitted to key in custom entries, such as glossary items or exercises, but hardly to perform an independent examination of a linguistic feature on a larger context of authentic text, let alone on self-selected text. Reasons of copyright lend software producers justifiable excuse, but the restriction must come as a shame today, when so many other language learning materials (dictionaries, courses, vocabulary syllabi etc.) grow from corpus-based lexicogrammatical findings. (Kaszubski and Wojnowska 2002: 1)

Therefore CALL software producers should shift their emphasis from traditional methods onto more promising corpus-based techniques of program implementation. Using authentic examples from a built-in corpus in explanations and knowledge checking exercises could make this type of software non-deterministic, natural and thus more interesting. Among corpus-based CALL programs and educational packages, which go towards meeting these goals, one could mention Cobuild's *The English Collocations CD-ROM* – <http://www.athel.com/cobuild/collscd.html>, *WordPilot 2000* (a concordance-based aid for apprentice writers – <http://www.compulang.com>) or *ClozeMaker* (<http://www.edict.com.hk/clozemaker/>). There remain, however, still many problems with automatic corpus search because of the machine's inability to distinguish e.g., different meanings of the same set of characters. Some of them can be partially solved in the case of tagged corpora.

## 1.5 Conclusions

Corpus linguistics together with computer technology form a powerful methodological tool which must not be overlooked. Corpora give numerous opportunities to both teachers and learners, enrich the language instruction process and create a new didactic potential which has yet not been fully assessed and appreciated. The chapter presents shortly the existing corpus-based CALL software and introduces the most vital aspects of corpus linguistics which are the cornerstone of the TestBuilder project.