A Practical Demonstration

Sorina POSTOLEA*

Key-words: corpus linguistics, corpus-based analysis, corpus-analysis tools, lexicography, terminology

Along with the advances of information and communication technologies (ICT), the means available to linguists and text researchers have grown exponentially. To begin with, the advent of digitalized text production, text editing, and text storage tools fostered the creation of very large collections of texts, also known as electronic corpora. In fact, in recent years, various digitalized collections of textual material and various computer programs specifically designed for their analysis – corpus tools – have been extensively used for various types of textual investigations and in a wide array of applied language studies. Small-sized to mega-sized digitalized collections of texts and corpus-analysis tools are used nowadays to support research in such fields as general linguistics, lexicography, grammar studies, terminology, translation studies, or literary studies. Corpus linguistics, the discipline that deals with corpora and corpus tools, has developed exponentially in the Western world, to the point that most language-related studies are nowadays based on its principles and tenets. Yet, because the development of corpus-analysis tools specifically designed to support the peculiarities of Romanian as a language would require insight from interdisciplinary teams of researchers, i.e. at least from the fields of linguistics and natural language processing, corpus linguistics is still a tentative branch of research in Romania.

Using a corpus of 140 English ICT news articles and press releases, this article aims to discuss some of the basic concepts and principles used nowadays in corpus linguistics as well as to provide a practical demonstration of how the main types of corpus-analysis tools may be used to investigate a collection of texts.

1. Word-lists and keyword tools

Creating word-lists is the most basic way of analysing a corpus. Unlike humans, computer programs are able to break up a text into all of its components (words) and then re-organise these elements according to various criteria in a matter of seconds. While calling it “a transformation”, Scott and Tribble emphasise that the process of creating word-lists “changes the object being considered radically from a

* “Al. I. Cuza” University of Iaşi, Romania; “Petre Andrei” University of Iaşi, Romania.

text which can be read linearly to some other form which will give rise to important insights, pattern recognitions, or teaching implications” (Scott and Tribble 2006: 12). Most corpus-analysis programs are able to sort the words in a corpus in alphabetical order, in order of frequency, or according to other criteria. For instance, *Figure 1* and *Figure 2* below show the first 19 words in the corpus of English ICT news articles compiled for this article, listed in order of frequency and in alphabetical order by a computer program called AntConc¹.

---

Figure 1. Word-list by frequency in AntConc

![Figure 1](image1)

Figure 2. Word-list in alphabetical order in AntConc

![Figure 2](image2)

¹ AntConc is a corpus-analysis program designed by Lawrence Anthony. For details, see http://www.antlab.sci.waseda.ac.jp/index.html
When dealing with word-lists, corpus linguists distinguish between two distinct kinds of units: (word) types and (word) tokens. In plain words, type refers to a particular word taken into account only once, while token refers to all of its occurrences or instantiations in that corpus. For instance, according to Figure 1 above, the word-type “ability” is instantiated in 25 tokens (occurrences) in the English ICT news articles corpus. As Michael Stubbs puts it, “each word-form which occurs in a text is a word-token”, while “when we are talking of the number of different words in a text, we are referring to word-types” (Stubbs 2002: 133). As shown in Figures 1 and 2 above, the English ICT news articles corpus taken as example comprises 7,424 word-types and 72,347 word-tokens.

Word-lists are an excellent way of analysing the general lexical structure of a corpus: “comparing the number of tokens in the text to the number of types of tokens (...) can tell us how large a range of vocabulary is used in the text” (McEnery and Hardie 2012: 50). This analysis implies computing the type-token ratio, which, “is a measure of the lexical diversity of a text. It depends on the size of an author's vocabulary and on the way in which the words in this vocabulary are used in the text” (Stubbs 2002: 133). This ratio is determined by a simple mathematical formula, which involves dividing the number of types by the number of tokens in a corpus. In the case of the English ICT news articles corpus taken as example, 7,424/72,347 gives a type-token ratio of ~0.1027 or, expressed as a percentage, of ~10.3%. As McEnery and Hardie explain, “this allows us to measure vocabulary variation between corpora – the closer the result is to 1 (or 100 if it’s a percentage), the greater the vocabulary variation; the further the result is from 100, the less the vocabulary variation” (McEnery and Hardie 2012: 50). Thus, the 10.3% type-token ratio computed above shows little lexical variation in our corpus.

However, the data show another picture if we take into account another important distinction, made between grammatical and lexical words or, otherwise put, between function and content words. Although they are the most frequent word types in any kind of text and thus they produce the highest number of tokens (see Figure 1 above), grammatical words do not carry meaning on their own. Their role “is mostly to glue texts together by supplying grammatical information to a lexical warp of nouns, verbs, adjectives, and adverbs” (Scott and Tribble 2006: 23-24). This is why they are also called minor, empty, form, or functional words. On the flip side, content words are very important in corpus analysis because “they carry most of the lexical content, in the sense of being able to make reference outside language” (Stubbs 2002: 40). Content words are also referred to as major, full, or lexical words.

The data shown in Figures 1 and 2 above refer to all the words in the corpus, including both content and function words, treated as lowercase words. Yet, nowadays, most corpus tools allow researchers to draw a particular kind of lists, called stop-lists, in which they are given the possibility to specify all the function words that they need to exclude from their analyses, such as, for instance, definite and indefinite articles, demonstrative and possessive adjectives, pronouns, particles, etc. With a stop-list that comprised over 500 grammatical words, AntConc showed a different composition for our corpus, comprising 6,892 types (content words) and 42,118 tokens. This resulted in a higher type-token ratio, of ~16.4%,
and, therefore, showed higher lexical variation (see Figure 3 below). Of course, based on the specific corpus design adopted and on the research goals pursued, this type of statistical data may be used for various qualitative analyses: to compare the lexical variation of different corpora, to analyse lexical density within a specific textual genre, etc.

**Figure 3.** Top 20 content words by frequency

**Figure 4.** Keywords (compared to three other corpora)

Word-lists may also serve as a basis for another type of analysis, referring to *keywords* (KW). Simply put, “this method identifies items of unusual frequency in comparison with a reference corpus of some suitable kind” (Scott and Tribble 2006: 55). It may also be used to check the keywords in a text with respect to a corpus, or, in other words, “to see which words occur significantly more frequently (according to standard statistical tests) in the text than in the corpus” (Stubbs 2002: 129). In relation to KWs, researchers talk about keyness, defined as “a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail” (Scott and Tribble 2006: 55-56). Figure 4 above shows the top 20 keywords in our English ICT news articles corpus in comparison with other three sub-corpora of English texts (*press releases, product descriptions, and user manuals*). A summary analysis of these KWs is enough to show that, with respect to the other corpora, this corpus focuses mainly on reported statements (*said*), on new products and technologies, on *users*, and several corporations active in the ICT field. Thus KWs provide a very useful quantitative insight into what ICT news articles are, in fact, supposed to be about.

2. **Concordancers**

Although researchers do not openly discuss this distinction, it seems that *word-lists* and *keyword tools* work mainly vertically, breaking down corpora and rearranging them on a vertical axis. On the contrary, *concordance tools* work mainly horizontally, breaking texts down into horizontal lines and contexts. The data thus
obtained is then rearranged vertically, according to various vertical criteria, such as frequency or alphabetical order. Computer programs purposely designed for this type of horizontal analysis of corpora are called concordancers\(^2\) or concordance programs\(^3\). In plain words, this particular corpus tool “allows us to search a corpus and retrieve from it a specific sequence of characters of any length – perhaps a word, part of a word, or a phrase” (McEnery and Hardie 2012: 35).

**Concordance, collocation, and cluster** are the three main concepts – and corresponding tools – used in this particular type of linguistic analyses. Central to all of them is the notion of context, because, unlike word-lists and keyword tools, these corpus aids are deliberately designed to allow for words (or groups of words) to be studied in their more or less immediate environment. Simply put, when researchers carry out a concordance analysis they use a concordancer in order to search a “specific sequence of characters”, called a node word or node in their corpus. Based on specific algorithms, the concordancer retrieves from the corpus all the occurrences of that node and then inventories and displays them along with their context (e.g., the sentences, lines, paragraphs, etc. in which the node was used). An option offered by all concordancers allows for the node to be displayed in the middle of the screen and thus be highlighted in its context; this particular way of displaying the search term is referred to as the keyword in context or KWIC format.

As Michael Stubbs explains, concordance is the “main tool of corpus linguistics” because concordancers are able to manipulate the lines thus obtained in

---

\(^2\) As McCarthy and O’Keeffe show, the origins of concordancing may be traced back to the 13\(^{th}\) century, when Bible scholars needed to “specify for other biblical scholars, in alphabetical arrangement, the words contained in the Bible, along with citations of where and in what passages they occurred”. The first concordance of the Bible, *Concordantiae Morales*, is said to have been produced during the first decades of the 13\(^{th}\) century (McCarthy and O’Keeffe 2010: 3).

\(^3\) In fact, present-day concordancers, such as AntConc, for instance, are complex tools, which include word-listing and keyword tools, as well as concordance, collocation, and cluster tools.

Figure 5. Concordance example in KWIC format
various ways: “sorting them alphabetically to left or right often makes it much easier for the human being to spot other patterns” (Stubbs 2002: 62). These patterns usually refer to co-occurrence. Since in discourse words are never used on their own, but are linked together by very wide and intricate grammatical, lexical, and semantic nets, spotting the surrounding elements with which they tend to associate is one of the fundamental goals of corpus-based research. But, as Scott and Tribble emphasize, “co-occurrence does not of itself tell us much about the relationship between the two items, any more than finding out that person X is usually to be found near person Y does not tell us whether they necessarily like each other” (Scott and Tribble 2006: 36).

The mere observation that particular words tend to appear more or less close to each other in particular types of texts does not reveal much about their meaning or the way in which they interact in those texts. This is why in corpus linguistics co-occurrence is deemed to be relevant only within the context of a limited number of words, to the left and to the right of the search term. The words which co-occur with the node are called collocates. The left and right collocates of a node-word, usually 3-4 on each side, form a collocation span. Collocation spans are noted by specifying the number of words taken into account on each side – a 3:3 collocation span would thus consider the 3 words to the left and the 3 words to the right of the search term. “Position in the span can be given as N-1 (one word to the left of the node), N+3 (three words to the right), and so on” (Stubbs 2002: 29). As “corpus linguistics is based on the assumption that events which are frequent are significant” (Stubbs 2002: 29), it is the most frequent collocations of a particular node-word that are considered to be relevant. This is where concordancers become crucial, as they are able to determine and range in order of frequency all the collocates of a particular node in a particular corpus in a matter of seconds.

The figures above show the top 20 left and right collocates of the word technology in the English ICT news articles corpus, determined on a 1:1 span, in order of frequency. The left-side span thus reveals collocations such as screen
technology, thermal technology, LED technology, transistor technology and so on, while the right-side analysis shows phrases like: technology solutions, technology officer, technology offers, enables, ensures, etc.

Finally, another way of using concordancers is to look for clusters. Corpus linguists distinguish between N-grams and clusters. The former, also known as n-word clusters or lexical bundles, are basically word-lists that do not refrain to a single word. As Scott and Tribble show, “the mechanism for listing words can be adapted to compute 2-, 3- or some other number of word-clusters” (Scott and Tribble 2006: 19). The n- in the name of this type of cluster (see Figure 8 below) thus stands for the number of words to be included in the search, a number which is established by the researcher. On the other hand, clusters refer to “repeated groups (...) found within the set of concordance lines, using the collocation horizons established by the user” (Scott and Tribble 2006: 41). While N-grams search the entire corpus for the most frequent groupings of an “n” number of words, clusters are a type of concordance because they necessarily include a specific search term (see Figure 9 below).

![Figure 8. Example of N-gram (3-word)](image1)

![Figure 9. Example of cluster (node-word: “technology”)](image2)

Both N-grams and clusters are particularly useful to linguists. Although not all the frequent groupings of words in a corpus may be lexically relevant, these two types of analyses could potentially reveal multi-word units or some other lexical “units which hang together in semi-fixed phrases” (Scott and Tribble 2006: 41). In Figures 8 and 9 examples of multi-word units would be: as well as, on the go, High Endurance Technology, chief technology officer, advanced graphics technology, or digital LED technology, while a variety of, a wide range, in addition to, or energy-efficient technology could be considered “semi-fixed phrases”.

Figure 8. Example of N-gram (3-word)  Figure 9. Example of cluster (node-word: “technology”)
3. Conclusions

Although limited in size, it is our hope that this practical demonstration was enough to show the very rich potential and the countless possibilities of inquiry provided by corpora, corpus-analysis tools, and the methodology put forth by corpus linguistics. Unfortunately, due to the peculiarities of Romanian as a highly inflectional language, the computer programs available nowadays to foreign linguists are rather difficult to use with Romanian corpora. For instance, most of the concordancers available on the market would count the various inflected forms of a singular Romanian noun – e.g. computer, computerul, computerului – as three different types whereas, in fact, they are tokens of the same word-type. Yet, as we have seen, corpus-analysis tools are able to provide precious insight into the “inner life” of texts, and it is our belief that such tools should be indispensable to modern textual analyses. However, the design of such computer applications, able to process Romanian as a language, would necessarily need to be carried out by large research teams, in which linguists and computer programmers would have to work side by side. But this would also mean appropriate funding and, at least so far, the Romanian government and authorities have shown neither interest nor much generosity in this respect.

Bibliography


Scott and Tribble 2006: Mike Scott and Christopher Tribble, Textual Patterns: Keywords and Corpus Analysis in Language Education, Amsterdam/Philadelphia, John Benjamins.


State-of-the-art Text Linguistics: Corpus-Analysis Tools. A Practical Demonstration

In recent years, electronic corpora and the computer programs specifically designed for their analysis have been extensively used for various types of text analyses and in a wide array of applied language-related studies. Small-sized to mega-sized digitalized collections of texts and corpus-analysis tools are used nowadays to support research in such fields as general linguistics, lexicography, grammar studies, terminology, translation studies, or literary studies. Corpus linguistics, the discipline that deals with corpora and corpus tools,
has developed exponentially in the Western world, to the point that most language-related studies are nowadays based on its principles and tenets. Yet, because the development of corpus-analysis tools specifically designed to support the peculiarities of Romanian as a language would require insight from interdisciplinary teams of researchers, i.e. at least from the fields of linguistics and natural language processing, corpus linguistics is still a tentative branch of research in Romania. Based on a corpus of English news articles that approach information and communication technology topics this contribution aims to provide a practical demonstration of how the main types of corpus-analysis tools that are now available to Western researchers may be used to explore a collection of texts.