ON CORPORA

CĂTĂLIN DEHELEAN

Abstract

This article is meant to provide some degree of insight into a certain idea much discussed within the realm of Computational Linguistics. It is the matter of finding and organising the parts of language of interest.

Structural Linguistics has been and still is using a traditional but proven invention, namely the dictionary, which may occur under a number of guises. While this is reasonable, there are limitations to any such lexicographical achievement. This is to say that a dictionary might not be quite the ideal tool for Computational Linguistic Research. Indeed, Computational Linguists prefer a different tool called a Corpus.

To circumscribe the term corpus is by no means a trifling matter. For the idea of corpus is not there as such. It is expressed only by means of certain phenomena. This is to say that corpora may be constituted from all acts of language. Thus corpora may provide vivid examples of use of any given language sample.

Keywords: *Computational Linguistics, dictionary, corpus, text, speech*

Introduction

It is in the interest of linguists to adapt their approach to language study in order to achieve new insights in the way language works. It has become necessary to move away from the idea of immutability in the study of the relationship between words and their meanings, and therefore their use, to the idea of studying language units in context and their associated gradual contextual change.

Dictionary and corpus

This first part of this article is based on a comparison between two linguistic terms. The relationship established in this case is between the terms dictionary and corpus. (See Figure 1.)

dictionary	corpus

Figure 1: A graphic representation of the two elements of this comparison

Dictionaries and corpora are two terms mainly used by two different branches of Linguistics. These two branches are Structural Linguistics and Computational Linguistics. Dictionaries are, for the most part, associated with Structural Linguistics, while corpora are often seen as a tool of Computational Linguistics. (See Figure 2.)

dictionary	corpus
Structural Linguistics	Computational Linguistics

Figure 2: A graphic representation of the branches of the Linguistics which use dictionaries and corpora

These preferences are relevant in the light of the dichotomy between two views on the collection of units of language. When it comes down to the essence of this dichotomy, it is to be found in the different views of two types of Linguistics. There is the view of Structural Linguistics which follows the theoretical tradition of definition and exemplification of units of language noticeable in the dictionary, and there is the empirical way of searching for examples, a view postulated by Computational Linguistics in the corpus. (See Figure 3.)

dictionary	corpus
theoretical view	empirical view

Figure 3: A graphic representation of the two views

Both of these views can be noticed in their respective approaches to units of language. There are two approaches to language units adopted by Structural Linguistics and by Computational Linguistics, respectively. While structural linguistics has adopted an extensive

approach to this problem in the form of the dictionary, computational linguistics has preferred the intensive approach of the corpus. (See Figure 4.)

dictionary	corpus
extensive approach	intensive approach

Figure 4: A graphic representation of the two approaches

The two approaches become rather evident when the principles behind the collection of language units are investigated. For purposes of ease of understanding, a minimal approach to the process of understanding principles is required. As such, dictionary can be seen as a large collection of language units. A corpus, on the other hand, might be understood as a collection of instances of use of a single language unit. (See Figure 5.)

dictionary	corpus
collection of language units	collection of instances of use of a single language unit

Figure 5: A graphic representation of the principles of the terms *dictionary* and *corpus*

Text and Speech

This second part of this article is also constructed on a comparison between two terms. These two terms are just as semantically close to each other as the previous set. They are the text and the speech. (See Figure 6.)

text	speech

Figure 6: A graphic representation of the two terms, text and speech.

There is a long honoured tradition of studying both the text and the speech. It is mainly focused on the analysis of their structure. For

the sake of simplicity, it may be postulated that the text is based on a set of words, while the speech is the result of the coming together of a string of sounds. (See Figure 7.)

text	speech
words	sounds

Figure 7: A graphic representation of a simplified structure of the two terms

When it comes to the preservation of language units, texts and speech exhibit quite a difference. Texts are not as problematic as speech. Texts are, by their very definition, found only in writing. Speech, as a different phenomenon, is bound to be recorded on any number of media for safe keeping. It could thus be said that texts are written while speech is recorded for the purposes of storing the language samples need for study. (See Figure 8.)

text	speech
written	recorded

Figure 8: A graphic representation of the two elements of the current comparison

The specificity of text and speech is another problem which must, at least, be taken into account. Both words and utterances are context-bound, that is to say they cannot exist outside a text or a speech. However while it may not be evident to which author a text sample pertains, any speech sample is intrinsically connected to a certain speaker. So, when it comes to specificity one may say that texts are general, while speech is specific. (See Figure 9.)

text	speech
general	specific

Figure 9: A graphic representation of the two elements of the current comparison

In the end, there have to be two different types of corpora to collect the two types of acts of language. In order to collect written samples one is bound to create a text corpus, while for the actual utterances, a speech corpus is required. (See Figure 10)

text	speech
text corpus	speech corpus

Figure 10: A graphic representation of the text and speech and their relationship with the respective corpora

Conclusions

In the end, any discussion about corpora may be a matter of perspective. Dictionaries might be perceived to be old and corpora are perceived to be new. This perception is most likely related to the types of linguistics that employ tem in the study of language: Structural Linguistics is old therefore dictionaries must be an older, while Computational Linguistics being new, Corpora must be newer as well. The inaccuracy of such a perception is easily disproved twofold.

Firstly, dictionaries and corpora are not just products, they phenomena based on ideas. The idea behind the dictionary is the typically attempt of structuralism to analyse every single element and determine its substance. The corpora are meant to show the language as it is without splitting it from its immediate context and turning it into an abstract entry. This type of approach is not necessarily bound to a type of linguistics.

Secondly, the process of writing a dictionary is complicated. One needs a list of words with their functions, then their definitions followed by a few of their exemplified uses. This is a typical scholarly endeavour. Therefore one can easily postulate that it could not have been the first way of collecting and ordering language in history. Corpora on the other hand are easier to make. Its basic requirement is to record language samples in their particular contexts. And indeed historical evidence shows just that.

This only comes to show that one cannot write the perfect lexicographical tool. One cannot have a book with all words of a language written down and thoroughly explained. Nor can one have all the instances of a sample recorded. But one may have enough of them in order to understand that sample in context.

Bibliography

Biber, D., Conrad, S., Reppen R. (1998). Corpus Linguistics, Investigating Language Structure and Use, Cambridge: Cambridge University Press.

Edwards, J., Lampert, M. (1992). Talking Data – Transcription and Coding in Discourse Research. Hillsdale: Erlbaum.

Facchinetti, R. (2007). Theoretical Description and Practical Applications of Linguistic Corpora. Verona: QuiEdit.

Facchinetti, R., Rissanen M. (2006). Corpus-based Studies of Diachronic English. Bern: Peter Lang.

McCarthy, D., Sampson G. (2005). Corpus Linguistics: Readings in a Widening Discipline, Continuum.

Rezumat

Scopul acestui articol este acela de a da posibilitatea cititorului de a se informa cu privire la o idee mult discutată din domeniul Lingvisticii Computaționale. Este vorba despre colectarea şi organizarea materialului lingual studiat.

Lingvistica Structuralistă a folosit şi încă mai foloseşte un instrument tradițional dar care şi-a dovedit utilitatea, şi anume dicționarul, care poate lua diverse forme. Deşi acesta este extrem de util, orice asemenea lucrare lexicografică îşi are limitările sale. Prin aceasta se doreşte a se sublinia faptul că dicționarul nu este instrumentul ideal pentru cercetare în Lingvistica Computațională. Aceasta din urmă foloseşte un alt instrument numit corpus.

A descrie acest termen nu este de loc uşor, fiindcă ideea de corpus nu se manifestă într-o singură ipostază. Fenomenele care o întruchipează sunt multiple. Lexicografii orientați spre Lingvistica Computațională pot alcătui corpusuri din orice fel de act limbă. Astfel corpusurile pot oferi exemple clare de utilizare în context al oricărui eşantion dintr-o limbă.