CORPUS LINGUISTICS – THE BASIC FORM OF LINGUISTIC ANALYSIS

Nikola Dobrić
Alpen-Adria Universität Klagenfurt

Abstract: The corpus approach in its contemporary framework marks the return of linguistics within the boundaries of empirically founded sciences from the long predominance of introspection-based analysis. This first part of this paper elaborates on the advent of corpus-based research in linguistics. The second part describes the notion and types of corpora. The third part describes the advantages of corpus-based research and the basic characteristics of corpus linguistics. The last part of the paper explains certain limitations of the corpus-based analysis.

Key words: corpus, empiricism, emergent grammar, functional grammar

1. ARMCHAIR LINGUISTICS VS. CORPUS LINGUISTICS

In the late 1950’s and early 1960’s publications by Noam Chomsky shifted the focus of linguistic research. The shift was from empiricism to rationalism and from language performance to language competence. Chomsky and his followers proclaimed natural that language was a poor reflection of language competence and corpora a poor source for competence research. Intuition was seen as a far better way of language analysis than natural language corpora. From the 1980’s the idea that intuition alone is not sufficient a method to provide solid evidence for research results and that perhaps language performance and meaning should actually be the primary focus of linguistics instead of syntax. Questions were raised whether introspection can provide any answers at all regarding semantics and pragmatics. Can it tell us is “starting” always replaceable by “beginning”? Can it show if it is only “time” that is “immemorial”? Native speakers’ intuition is unreliable regarding such issues and provides no information on frequency of occurrence or usage.

Such questions cannot be answered without systematical investigation of natural language. Of course, the idea of working with natural language is not a novel one. It dates from the times of the old-world philosophers who were the first to understand the significance of an empirical base for any kind of solid language reasoning. That is how the notion of a corpus was conceived. The difficulty of capturing the natural language and its nuances was overcome with the advent of large collection of texts which were supposed to represent the whole discourse or a specific part of it.

Work with such collection of natural language is connected with researching semantic markings of a given language. It has to be understood that while the ability to produce communicate is innate to all people, the very language we produce and the meaning expressed by it is entirely socially conditioned. Language is the product of arbitrary idiosyncrasy and anomaly as much as of the innate grammatical laws permeating through it. Natural language is
a system internally governed by the negotiation of meaning within it. Corpus-based linguistics focuses on describing the content and the relationships connecting the natural language and the study of discourse as a mediator of knowledge as conditioned by fundamental pragmatic implications (Hopper, 156).

2. WHAT IS A CORPUS?

“A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.” (Sinclair, 12). It is important to grasp the difference of the term corpus in its usual use and in its referents regarding corpus linguistics. As seen from the previous definition, the term corpus as it will be considered in this paper stand, as well as in corpus linguistics, stands for a large collection of natural texts, both written and spoken, and they can be found in digital form stored in vast databases. The texts are usually collected by universities or similar institutions. Such a corpus consists of millions of words from various sources and registers: fiction and non-fiction writing, academic papers, newspaper articles, telephone conversations, commercials, lectures, public speeches, television interviews, written and spoken, formal and informal. There are precise rules for developing a corpus detailing the types and quantities of texts (Meyer, 30) and they are designed to insure that a corpus is representative and balanced (Dobrić, 2009, 7).

For every serious investigation it is important to clearly define the size and the variation of the research corpus, because that is a crucial matter in making sure that the feedback information will be of merit. According to what it represents one can determine the scope and the scale of the research to be done and hence what kind of results to expect. For example, if a corpus is made only of academic texts relating to physics one can hardly expect it to yield any good results in investigating language varieties. Hence, the representativeness of a corpus is a crucial issue and the most important characteristic of any corpora.

There are many types of corpora (general vs. specialized; static vs. monitor; native vs. learner; native vs. translated; monolingual vs. multilingual; parallel) all which serve different types of research and can provide invaluable feedback and results. The most interesting type is definitely the general monitor corpora because their size and synchronicity allows them to be regarded as representing the entirety of a given language and are as such interesting to a very wide array of scientists. Among the first and the most respected general corpora is the British National Corpus. It is a large online corpus compiled by the Oxford University Computing Services starting from the early 1980’s with texts dating back to 1964. It contains over 100 million words both from spoken and written discourse. Around 90% are the written texts from mainly informative writings, chosen from the fields of applied sciences, arts, belief and thought, commerce and finance, leisure, natural and pure science, social science, and world affairs with a section of literary and creative works. The oral part of some 10 million words containing the spoken corpus draws upon transcriptions of spontaneous natural conversations made by members of the public and a context-governed part, containing transcriptions of recordings made at specific types of meeting and event, making the corpus diverse in its register and scope. There are other similar corpora of both English and many other languages such as Corpus of Contemporary American English (http://www.americancorpus.org/); Russian National Corpus (http://www.ruscorpora.ru/en/index.html); Croatian National Corpus (http://www.hnk.ffzg.hr/); Czech National Corpus (http://ucnk.ff.cuni.cz); Hungarian National Corpus (http://corpus.nyud.hu/mnsz/index_eng.html); and many others. There is however a lack of such national corpora in Serbia, which is a basic prerogative for a successful corpus linguistics community, despite the fact that working with electronic
corpora has a long history within our linguistic community. A little known fact is that the first
corpus of Serbian (or rather Serbo-Croatian) was created only a year after the first larger
corpus of in the world came into being. It was the result of a large Yugoslavian linguistic
project which was conducted by Rudolf Filipović and Željko Bujas (Tadić 1997: 388), among
others. Yugoslavia continued to be involved in all of the major corpus linguistics projects up
until the 1990’s when the war severed all of the ties with the world. The cooperation still has
not been reestablished to a satisfactory degree.

There are currently only two general computer corpora of Serbian: Corpus of Contemporary
Serbian and Corpus of Serbian. Both of them are the direct outcomes of major national
projects aimed at creating a national corpus which would serve both as a source of research
but also as a repository for all of the linguistic richness of the Serbian language. It is therefore
surprising how both of them seem to function as private corpora, since access to them is
usually obtained with difficulty or not at all.

The Corpus of Contemporary Serbian is the result of the project Matematička i računarska
lingvistika completed in 1981 under the leadership of Duško Vitas, and was, with certain
additions and improvements, made available online several years ago. It comprises of 24
million words and is mostly not annotated. The Corpus of Serbian was truly a visionary work
instigated by Đorđe Kostić as a part of a large linguistic project started in 1957. It was
digitalized in 1996 by his son Aleksandar Kostić and it has about 11 million words. It is a
diachronic corpus with sources starting from the 12 century and is fully annotated. Both of
the represent a tremendous effort and truly large enterprises but unfortunately cannot serve as
good general corpora for several reasons. The first reason that makes them fall short of good
general corpora is the lack of representativeness. The Corpus of Contemporary Serbian has no
spoken language sources, while its sources of written language are not representative enough
since 22 million words come solely from the Politika newspapers. The Corpus of Serbian also
lacks any spoken language and its written language sources are not sufficiently contemporary
so it lacks a representative synchronic dimension. The second fault we can observe is the
annotation. While the Corpus of Serbian is superbly annotated, the Corpus of Contemporary
Serbian is only 80% annotated automatically using unreliable software. The third reason is of
course the fact that these corpora are for the most part closed for the general users, so most of
the linguistic public never even had a chance to see them or use them. None of the corpora
provide an option to subscribe to the corpus, either for free or for a fee so the permission
needs to be asked individually. The permission by the owners of the Corpus of Contemporary
Serbian is generously given while it is virtually impossible to gain access to the Corpus of
Serbian.

It is possible hence to draw a conclusion how both Serbia as a country and its linguistic
community very much need a national general monitor corpus which would be conducted as a
national corpus, which would be free to access and which would both store the diachronic
language and its synchronic material (Tadić 1997: 388).

4. CORPUS ANALYSIS
The reason why this approach to linguistic research is so significant is the fact that it focuses
primarily on the pragmatic approach to grammar as opposed to the one embodied by universal
grammar. The concepts like emergent grammar (Hopper, 67–92) and functional grammar
(Halliday, 274) point out that language structures stem out of language usage and that they are
basically formed by the previous discourse and the previous language experience combined
with the contemporary discourse. Language formations are only partly or fully hardened
formations created by frequent usage governed by a suitable context. And there is hardly a better source to investigate frequency and language performance coupled with the social context than large corpora.

The list of areas of linguistics that can be addressed with the empirical corpus-based approach represents the core disciplines of linguistics:

− lexicography: it enables the examination of the linguistic and non-linguistic associations of particular words and can greatly help dictionary makers in their work;
− sociolinguistics: the corpus-based techniques allow exploration of dialects and registers that previously could not have been undertaken;
− discourse analysis: such an approach provides us with a sample of language large enough that we can finally make satisfactory generalizations about a language's characteristics beyond its structure;
− morphology: results of research through the application of this method can tell us a lot about the frequency, distribution and the function of word variants;
− phonology: it can give us great insight into the various means of phonetic distribution;
− semantics: no other existing approach can yield such comprehensive and insightful evidence of the meaning of words than this;
− syntax: investigating language structure in this way can give empirical evidence about how we construct sentences and express ourselves through language;
− comparative and contrastive linguistics: it can show us the similarities and differences in language use between languages, provided the existence of parallel corpora;
− educational linguistics: such large scale studies can help in designing effective learning materials and activities for language learners.

To sum up, practically any area of linguistics can be researched from the perspective of usage, and this is quickly being discovered and used by linguists in all of the aforementioned fields. The empirical nature of corpus based linguistics allows for all of the research to have a solid base of concrete and statistical evidence for any manner of language analysis.

6. LIMITATIONS OF CORPORA
The biggest fault of corpora can be seen from the following quote “I don't think there can be any corpora, however large, that contain information about all of the areas of [English] lexicon and grammar that I want to explore [...]“ (Fillmore, 35). It is of course true that the sheer volume of natural language will never be able to be captured inside a database because it is truly mathematically infinite. It is true that even the most representative corpus represents only one, smaller or bigger, cross-section of the absolute discourse and that every corpus suffers from overrepresentation or underrepresentation. However, the second part of the previous shows that even if we acknowledge all of the limitations of corpora it still remains as the one of the best sources for language research “[...] but every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. “ (Fillmore, 35).

REFERENCES

Nikola Dobrić
Korpusna lingvistika – osnovni oblik lingvističke analize

Rezime

Korpusni pristup u svom savremenom okviru obeležava povratak lingvistike u okvire empirijskih nauka posle duge dominacije introspektivne analize. Prvi deo rada prikazuje početke korpusnog pristupa u lingvistici. Drugi deo opisuje pojam i tipove korpusa. Treći deo rada predstavlja prednosti korpusnog pristupa istraživanju i osnovne karakteristike korpusne lingvistike. Poslednji deo rada daje određena ograničenja korpusne analize i samih korpusa.

Ključne reči: korpus, empirijski, uslovljena gramatika, funkcionalna gramatika

For a full list see Dobrić, 2009, 18.

The author has written more on the topic of importance of a national corpus in a paper which is due to be published soon.

Brown korpus (http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html)

It is important to note that the Corpus of Serbian is completely manually annotated for grammar, number of graphemes, phonological structure and for start and end of every sentence. This kind of level of annotation is very rare and this corpus is considered as one of the largest manually annotated corpora in the world (Kostić, 260).