Knowledge gleaned from sifting huge amounts of text

# A Corpus of Polish


Assoc. Prof. Rafał L. Górski is a staff member at the Institute of the Polish Language, Polish Academy of Sciences, in Kraków

RAFAŁ L. GÓRSKI
Institute of the Polish Language, Kraków
Polish Academy of Sciences
rafalg@ijp-pan.krakow.pl

Assoc. Prof. Adam Przepiórkowski heads the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences

ADAM PRZEPIÓRKOWSKI
Institute of Computer Science, Warsaw
Polish Academy of Sciences
adamp@ipipan.waw.pl

Prof. Barbara Lewandowska-Tomaszczyk heads the Chair of English and Applied Linguistics and the Department of Computer and Corpus Linguistics at the University of Łódź

BARBARA LEWANDOWSKA-TOMASZCZYK
University of Lódź
blt@uni.lodz.pl

Assist. Prof. Marek Łaziński is a staff member at the Institute of the Polish Language, Warsaw University

MAREK ŁAZIŃSKI
Institute of the Polish Language, Warsaw
University of Warsaw
M.Lazinski@uw.edu.pl

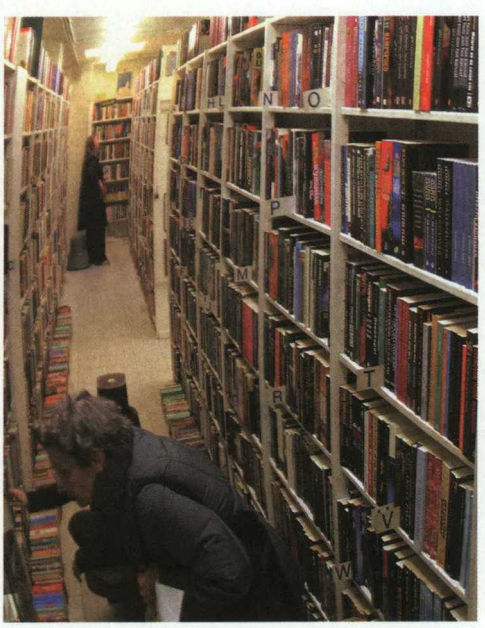

**If the object of study in linguistics is language, the question arises of how a linguist is to access that object**

There are two routes to studying language, just like language itself exists in twofold fashion. On the one hand, language can be thought of as a certain mental faculty that nearly all humans possess (bar those with severe disabilities or severe brain damage), an ability to generate and understand utterances. This is known in linguistics as "competence." On the other hand, language can also be seen as the output of that mental faculty, meaning spoken and written texts themselves ("performance"). The linguist may therefore study either this mental side of language - figuratively speaking, the grammar and lexicon that got placed into our heads back when we were learning language in our childhood - or he or she may study texts, as a basis for reconstructing this grammar and lexicon. Both methods can be justified and both are definitely equally warranted, although not always are both equally applicable to a specific research problem.

In order to access linguistic competence, the simplest approach known as elicitation is to ask a native speaker of a language whether a given utterance is in their opinion comprehensible, "correct," "well-formed," in other words consistent or inconsistent with their linguistic competence. Our research, however, is concerned with the other approach: studying performance, or texts themselves.

It should be said that at a time when this meant for the linguist to have to go through mountains of data armed with pen and paper, i.e. roughly before powerful personal computers became available and affordable, the approach was not enthusiastically embraced, hence the popularity of armchair linguistics and elicitation techniques.

Nevertheless, corpus linguistics is considerably older than the modern PC, having been launched at Brown University in the mid 1960's, where a group of linguists decided to digitize a sample of texts drawn from US books and newspapers. The project gave rise to the first electronic corpus, known as the Brown University Standard Corpus of Present-Day American English, or Brown Corpus as it is known which incidentally is still sometimes used by researchers the world over. It then contained the unimaginable figure of 1 million words. By comparison,



Herman Brinkman/www.sxc.hu

**How many consecutive nouns in the genitive case can be realistically strung together in Polish? A simple corpus search digs up such surprising real examples as the series *propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju* ("proposal for setting the date of launching the process of introducing reform of the system")**

Anna Piątkowska

**The corpus we are developing is already being used in lexicography. Dictionary editors nowadays have a much richer set of examples at their fingertips than back in the days of index-card work, plus computers to help preliminarily process this rich material**

that is around 500 times larger than the present article. Interestingly, work on the first electronic corpus of Polish texts begin not long thereafter, although unfortunately this project proceeded extraordinarily slowly. Frankly, we must concede that Polish, one of the major languages of Europe (at the very least in terms of its number of speakers), has yet to gain a corpus capable of meeting all the demands of modern linguistic science.

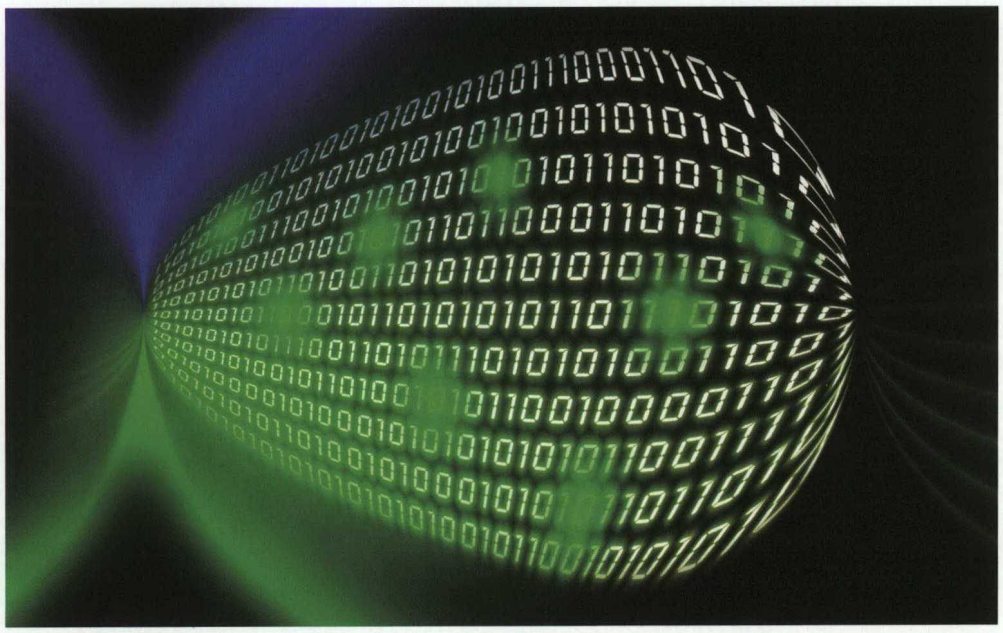
## Past efforts

It was not until the beginning of the new millennium that work on Polish corpora gained significant impetus. The first corpus to emerge was developed by the Institute of the Polish Language, Polish Academy of Sciences (not publicly available), followed by the corpus of PWN publishers, then the corpus of the PELCRA group at the University of Łódź, and finally the corpus of the Institute of Computer Science, Polish Academy of Sciences. All four teams decided to join forces (and resources) in 2006, forming the Consortium for the National Corpus of Polish to seek funding jointly and ultimately winning a grant from the Polish Ministry of Science and Higher Education (no. R17 003 03). Each of the individual corpora had its strong and weak points, and thus each of our teams can contribute different expertise.

The Consortium is comprised of the Institute of Computer Science (Polish Academy of Sciences), the Institute of the Polish Language (Polish Academy of Sciences), the Institute of the Polish Language (University of Warsaw), and the Chair of English and Applied Linguistics (University of Lódź). The first of these institutions, ICS PAS, is the coordinator of the project, which is headed by Assoc. Prof. Adam Przepiórkowski.

The project plans to develop a corpus of from 800 million to 1 billion words, i.e. three orders of magnitude larger than the Brown corpus. Of course, the only way to build a corpus of this size is from randomly selected texts. For many applications the most important part will be the "balanced" portion - the part that constitutes a representative sample of texts, or more precisely a representation of the language actually read and heard by the average member of the Polish-speaking community. In this case the texts will be chosen according to set criteria, although this will definitely not be a truly random sample since we need to respect copyright and we may only include texts into the corpus with their owners' consent. The balanced part is intended to encompass 300 million words of text. Notably, existing corpora of this sort for other languages mostly encompass around 100 million words - an informal standard set by the British National Corpus. Another special component of the Polish corpus will contain several million words of spontaneous conversations, recorded and transcribed.

Knowledge gleaned from sifting huge amounts of text



www.sxc.hu

**Even nonspecialists interested in language can use a corpus to satisfy their own curiosity, to track down a particular quote, to investigate how well-known quotations get reused and paraphrased, or try to pin down the meanings of unfamiliar words and expressions**

This will make it possible to study spoken language, which differs from written Polish in terms of many specific characteristics.

## Texts and more

While any digitized set of texts may be useful to linguists in certain ways, the more information gets built into a corpus, the greater its potential. For starters, our planned corpus will be outfitted with precise bibliographic references plus information about the type of text. This will enable every citation from the corpus to be ascribed to its author and title of the text it derives from. Additional information, such as the date of the text and its genre, will enable researchers to capture differences between language styles and identify changes that have occurred in recent years. For oral texts, demographic data on the speakers will play a similar role.

Each word in the corpus will be "tagged" with a grammatical description, so that in addition to specific words and phrases users can also formulate queries in terms of grammatical forms - such as a search to find all plural adjectives in the dative case, for instance. We also plan to devlop corpus tools to tag certain syntactic groups, although it will definitely not perform full grammatical parsing of each sentence. Another tool now being developed for the purposes of the project is a program to recognize proper names (especially multiword names) within the texts. Lastly, an automatic meaning recognition system will be implemented as more of a prototype: a certain list of meetings will be established for several hundred words, then occurrences in specific sentences will be marked for which particular meaning they bear. Naturally, due to the size of the corpus, all these tasks can only be performed automatically.

There are two tools to use for searching the corpus: Poliqarp developed by ICS PAS, and PELCRA developed at the University of Łódź. Their main task is to generate concordances, i.e. lists showing all the occurrences of a target word (or words) within the context of the closest several phrases. Moreover, they enable users to sort and search these concordances for typical collocations (word combinations), to locate specific quotations, and identify the number of times a target element occurs within the whole corpus. Users will be satisfied to access these tools online via web browsers, free of charge, and without user registration. Certain specific applications, on the other hand, will require ad hoc software to be developed. In the latter case, access may be limited to some degree out of concern for text security (authors and publishers permit the use of texts on condition that the project poses no competition to published books).

## Surprising real examples

What applications can such a corpus have? Chiefly linguistic ones, of course, as a source of examples for studying the actual use of a certain word or construction. To take a

slightly sophisticated example, note how the structure of Polish allows a series of nouns to be strung together in the genitive case – such as in the jocular expression *ojca szwagra żony brat* ("father's brother-in-law's wife's brother") essentially used to indicate that a person is simply a very, very distant relative. We might therefore ask how many such consecutive nouns in the genitive might realistically occur in actual Polish texts. A simple corpus search turns up such surprising finds as the series of eight consecutive genitive-case nouns in *propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju* ("proposal for setting the date of launching the process of introducing reform of the system").

However, being able to find such curiosities is just the tip of the iceberg. The broader use of corpora in linguistics research enables traditional qualitative descriptions to be accompanied by quantitative descriptions, thus reflecting not just the "rigid" linguistic rules but also trends that may sometimes be broken yet are nevertheless observable. Corpora enable us to distinguish what is typical in language from what is entirely acceptable albeit marginal. They help us to perceive to what extent language is comprised of certain "prefabricated" (ready-made) elements. On the one hand we are extremely creative when we speak, constantly piecing together new sentences and unique combinations of phrases, yet on the other hand many of the same combinations reappear time and again in what we say.

Let's imagine a linguistic phenomenon that occurs once every half a million words on average. When we sit down and read texts such phenomenon will disappear among thousands of others. But when we generate a concordance focusing on this one phenomenon, we can look at several hundred examples all at once.

## Goldmine of data

The corpus we are building is already being used in lexicography. Dictionaries have always been developed based on texts, but corpora have taken lexicography to a new level. Dictionary editors nowadays have a much richer set of examples at their fingertips than back in the days of index cards, plus computers to help preliminarily process all this rich material.

Corpora also have a strong role to play in teaching. Worldwide, attempts are being made to use corpora in foreign language teaching (so far mainly for English on the advanced, university level). Corpus use can also help in developing better teaching materials based on real usage, and they can definitely offer an interesting enhancement to high school writing and literature classes. Even nonspecialist adults can use a corpus to satisfy their own curiosity, to track down a particular quote, to investigate how well-known quotations get reused and paraphrased, or try to pin down the meanings of unfamiliar words and expressions.

The greatest public interest, however, is sparked by how corpora and the natural language processing tools developed for them can be used in linguistic engineering. Because most written texts are created in electronic form these days, finding ways to automatically "mine" the information contained in natural language is becoming an increasingly important task. All the above-mentioned tools used for preparing and processing a text corpus can be useful in intelligent data mining, machine translation, even gauging the public mood (by subjecting large numbers of newspaper and Internet texts to real-time analysis and spotting trends).

As mentioned above, Polish science has been somewhat behind in terms of creating corpora and using them in research. The project we have outlined here hopes to change that soon, and radically so.

A preliminary version of the corpus is accessible online at http://www.nkjp.pl. ■

**Further reading:**

Sinclair J. (1991). *Corpus, Concordance, Collocation (Describing English Language)*. Oxford: Oxford University Press.

sanja gieniero/www.sxc.hu

**The tools used for preparing and processing a corpus can even be useful in gauging the public mood – by subjecting large numbers of newspaper and Internet texts to real-time analysis and spotting trends**