Computer tools for language analysis

Weaving a Web of Words

MARCIN MIŁKOWSKI
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw
marcin.milkowski@gmail.com
Dr. Marcin Miłkowski, an assistant professor at the Logic and Cognitive Science Department of the PAS Institute of Philosophy and Sociology, works on the philosophy of mind and computational linguistics; he has won a scholarship from the Foundation for Polish Science and the weekly *Polityka*.

We use our computers to write, to make calculations, to search for information, and so much more. But we do not always realize how more and more of their capabilities crucially hinge on successful natural language processing technologies

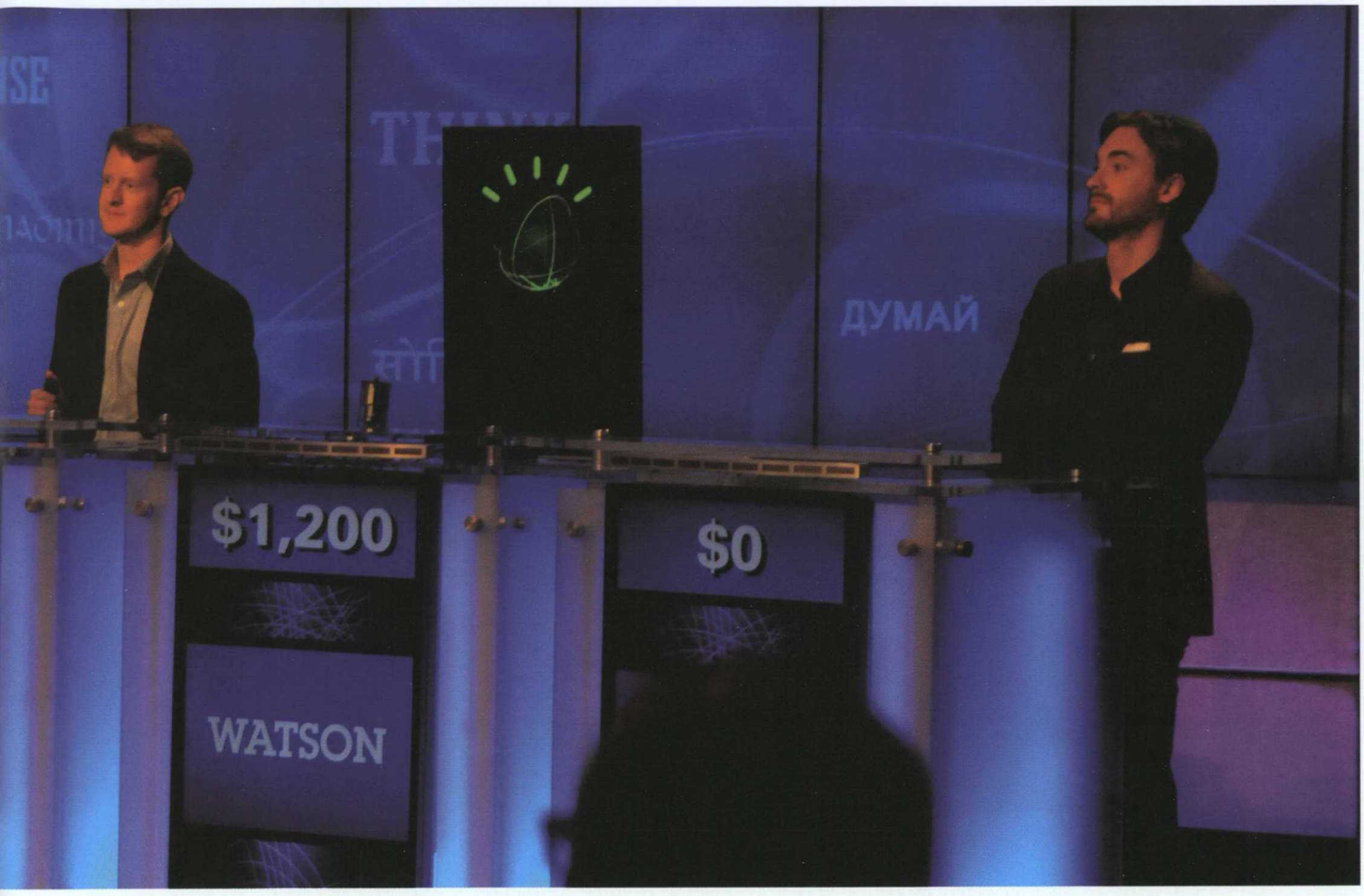
If we want to read the latest news from exotic countries, we can easily use an online translation services. Although the output will probably make us laugh to tears, such services enjoy increasing popularity, since they do in fact provide a certain approximation of the original meaning. In this sense, the success of computational linguistics is undeniable, and it is no wonder that IBM Watson – a supercomputer crunching gigantic amounts of information, combining it and drawing conclusions – ended up winning the TV game show "Jeopardy."

However, one has to admit that the most spectacular computational advances have focused on English. Although the language has no more than 400 million native speakers, it is nevertheless estimated to be spoken by at least a billion and a half people on the planet. This is why investments and research have for years concentrated on English.

Tools developed for one language do not necessarily work for another. Polish differs from English in terms of a relatively free word-order, a complex declension system, the existence of grammatical genders, and letters with diacritical marks. IBM Watson would have a much harder time defeating the contenders on the Polish TV show "Va banque" (the counterpart to America's "Jeopardy). Watson is able to analyze an English text grammatically, utilizing an advanced formalism and dictionary that represents the dependencies between the expressions in a sentence (called a "dependency grammar"). Such a grammar is only now being created for Polish, at the PAS Institute of Computer Science.

Thousands of rules – not enough

That does not mean that all the tools for processing English text and speech are better than those available for Polish. That is because computers can process language in two ways. The first involves statistical techniques. Sometimes an algorithm can be devised in a way that is language-neural. Such algorithms, for instance, lie at the heart of speech synthesizers, which can add an automatic overdubbed track to a movie or provide a voice for a GPS device. The Polish program IVONA is among the best in the world and can handle a large number of languages. Speech recognition, on the other hand, is unfortunately a different story: here the task is much harder and perhaps requires more linguistic knowledge than building a voice synthesizer. Speech recognition systems for English frequently also rely on grammatical rules, to help more accurately fish the individual spoken words out of a speech stream. This is the other (historically earlier) concept for teaching computers to handle language – using rules. If you want to realize how tough a task automatic speech recognition is, just listen closely to how we actually speak: we do not in fact make any breaks between words, like the spaces in written texts, and moreover we slur our words together. What a computer has to "listen to" is therefore one big mush, which we could try to more faithfully transcribe as "heissomthinspokenby-



IBM

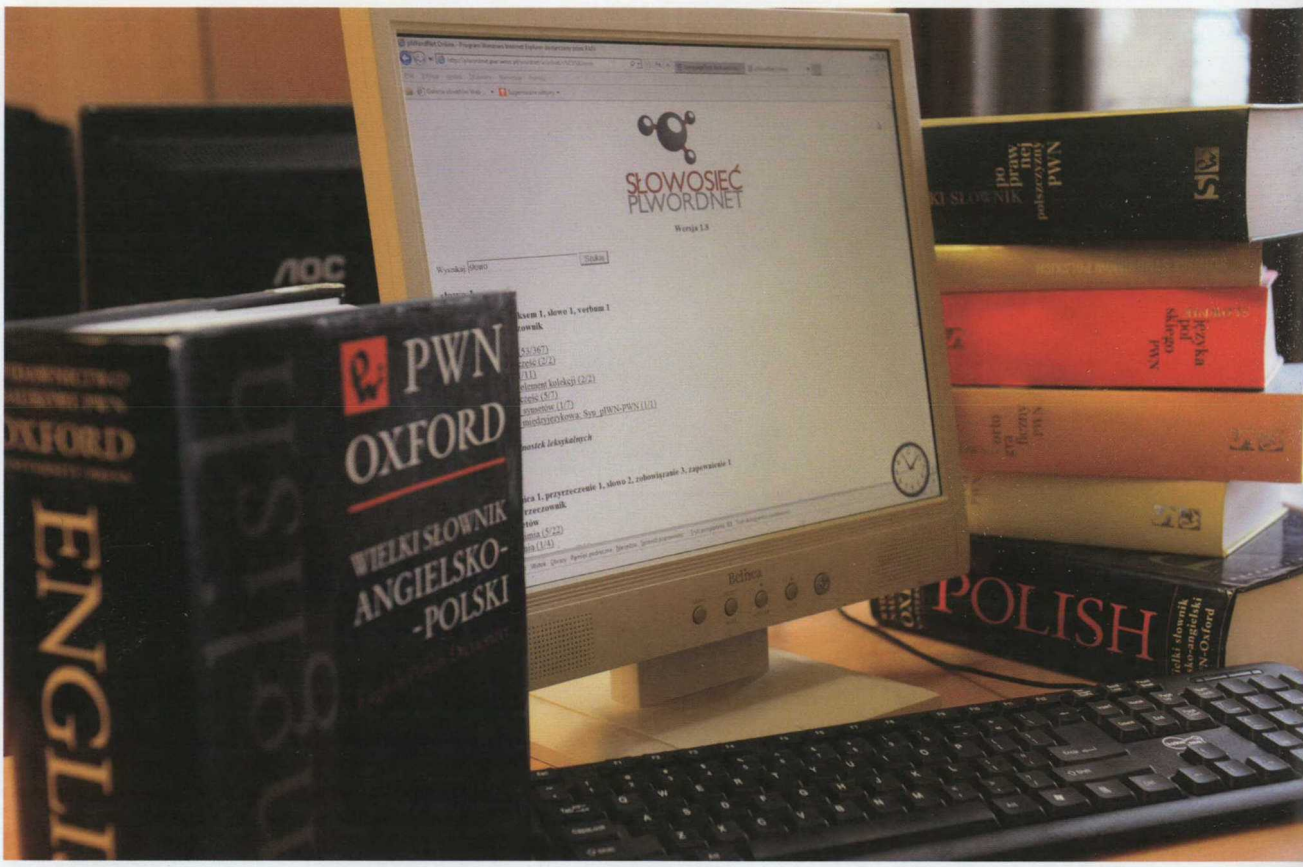
Watson defeating human contestants on "Jeopardy"

aprson." Yet we humans can understand it without trouble, and even with music playing in the background.

The spell-checking tools that we use in our text-editing programs, web browsers, and telephones are likewise relatively language-independent. They can easily start to cope with Polish, although simple spell-checkers that merely look up each word separately in a dictionary not be able to identify every mistake (unable to discern, for instance, *żądny władzy* from *rządny władzy* or between *prosimy o niepalenie* and *prosimy o nie palenie*). Only grammatical or style checkers, like LanguageTool, which the present author was involved in developing, can detect such mistakes and suggest sensible corrections. Although LanguageTool uses more than 1000 rules to check a Polish text, it is still far from being complete. Generating more such rules automatically, although possible in principle, requires huge quantities of text for processing. Contemporary linguistics studies actual language use empirically, using specially prepared collections of texts or other linguistic records, called "corpora."

The world's largest corpora, for example Google Books created by scanning volumes, may grow to the size of hundreds of billions of words, although it is not just size alone that determines their usefulness. Also important is the balanced selection of language samples. For Polish, the largest such corpus is the National Corpus of Polish, which is accessible at www.nkjp.pl and which has already been featured in an *Academia* magazine article (issue 2(22) in 2009). The corpus website offers search tools for studying the real usage of individual words or complex phrases, and even to ask which words "like" one another (frequently appearing collocations). The latter tool may be especially useful in the work of an editor or translator. If we want to check, for instance, whether it is better to write *w porównaniu do* ("compared to") or *w porównaniu z* ("com-

Jakub Ostałowski



pared with"), we just have to check which phrase is used in well-written Polish texts (mainly published books). Perhaps it will turn out that both phrases are correct, but used in different contexts? Feel free to look and analyze the results yourself.

However, a corpus would not be terribly useful without the ability to look for all phrases of a specific grammatical form, or call up all occurrences of a given word irrespective of how they are declined. But for that to be possible, the corpus texts first need to be appropriately preprocessed using, for instance, a morphosyntactic dictionary, which supplies grammatically tagged word-forms. Many such dictionaries have been created. Until recently the largest such resources were the Grammatical Dictionary of Polish (SGJP) developed by Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński and Robert Wołosz, and Morfologik, a resource that I created and made available under a free license on the Internet. The advantage of the Grammatical Dictionary of Polish was of course the higher quality of its grammatical description, whereas Morfologik had the advantage of being available under a liberal license, and could thus be built into open source programs. It therefore seemed natural to combine the two resources - and contrary to the unfortunate trends of fragmentation and reduplication that often prevail in science, a combined resource called PoliMof was indeed jointly created under the European project CESAR. It is currently the largest morphosyntactic dictionary for Polish, containing over 400,000 words (with more than 4 million declined forms).

The latest dictionaries, including WordNet, assist not just people but also artificial intelligence

Another type of corpus is called a parallel corpus - containing the same texts in more than one language. Because many texts get translated, and often using computer tools, it is possible to create such resources relatively cost-effectively. The EU, for instance, provides free access to the translations of its legal regulations - and to the parallel corpora so formed. Legal documents are not protected by copyright, unlike other types of text (which is a source of problems for

computational linguistics; one cannot simply gather together collections of texts and publish a corpus of them, if the texts are only available under a limited license). Such parallel corpora are also used by programs for statistically-based machine translation, which learn not from rules but are based on new statistical tendencies they observe in real text. You may notice, for example, that Google's automatic translation service is distinctly better at handling economic and legal texts. That is thanks to the EU corpora used by Google for training.

"The box is in the pen"

In fact, satisfactory machine translation was expected to be achieved a long time ago (at least by the optimists). Significant investments were made in this domain of artificial intelligence back in the 1950s, in the hope of quickly developing easy and rapid translation of texts (especially between Russian and English – these were, after all, frequently military projects). However, progress was very slow in coming. Some people even harshly criticized the very notion of translation ever being done by computers. The eminent logician and linguist Yehoshua Bar-Hillel, for instance, claimed that it would be impossible for a computer to properly translate the sentence "the box is in the pen," given that "box" and "pen" both have many senses and a computer would never be able to cope with such ambiguity. We can quite easily test how today's major machine translation resources cope with that task into Polish. Google Translate performs quite decently (suggesting: *Skrzynka jest w zagrodzie*), being a statistical system, although that means it may sometimes create ungrammatical sentences. The rule-based Translatica system being developed in Poznan fares somewhat less well in this case (offering: *Pudło jest w piórze*). Translatica's advantage, on the other hand, is its huge vocabulary (it draws upon the dictionaries produced by publisher PWN, further augmented by data gathered by the company PolEng) and its good handling of Polish grammar. Perhaps, in order to generate a larger number of rules, Translatica will have to analyze large quantities of text statistically.

It is indeed statistical data processing that it is now the driving force behind practical solutions in computational linguistics. Another example is what is known as WordNet – a type of dictionary representing the relationships between expressions (synonyms, antonyms, hyperonyms, etc.), developed in the United States. The Polish version of Wordnet, also known as *Słowosieć*, is available at http://plwordnet.pwr.wroc.pl/wordnet. It is among the largest such resources in the world, even though it was created at Wrocław University of Technology quite recently. The Polish WordNet was created partially automatically – using special algorithms for detecting the relationships between words in huge bodies of text. This type of dictionary represents words as a dense network of relations, capturing their meaning and interconnections. Such a dictionary is somewhat different than the typical thesaurus intended to help writers find a better word. No, the point here is not to avoid stylistic monotony, but to facilitate automated reasoning about a text. For instance, a web search engine equipped with such a dictionary can provide results that include not just the literally formulated search terms, but also their synonyms, or even other, related results. Briefly put, WordNet may be used to create a more semantic Web.

Computational linguistics faces many challenges. Although the existing tools and resources for Polish have made tremendous progress thanks to numerous research projects and commercial applications, it is still out of the question that a computer might be able, today or tomorrow, to fully interpret or generate a sentence in Polish at the same level as a typical Polish speaker. Much time will have to pass before computers are able to translate well the various tricky examples that Bar-Hillel invented. And that is not just due to insufficient funding or support from the largest corporations. It is also because full computer understanding of natural language is the Holy Grail of linguistics – is it actually attainable? ■

Further reading:

Miłkowski M., *The Polish Language in the Digital Age. Język polski w erze cyfrowej*, (red. Rehm G. i Uszkoreit H.), Springer, Berlin, Heidelberg, 2012 (dostępny bezpłatnie w całości pod adresem http://www.meta-net.eu/whitepapers/volumes/polish).

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.). *Narodowy korpus języka polskiego*, Wydawnictwo Naukowe PWN, Warszawa, 2012