Przeszukiwanie języka

Polski korpus

Doc. dr hab. Rafał L. Górski jest pracownikiem Instytutu Języka Polskiego PAN w Krakowie

RAFAŁ L. GÓRSKI
Instytut Języka Polskiego, Kraków
Polska Akademia Nauk
rafalg@ijp-pan.krakow.pl

Doc. dr hab. Adam Przepiórkowski jest Kierownikiem Zespołu Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN

ADAM PRZEPIÓRKOWSKI
Instytut Podstaw Informatyki, Warszawa
Polska Akademia Nauk
adamp@ipipan.waw.pl

Prof. dr hab. Barbara Lewandowska-Tomaszczyk jest kierownikiem Katedry Języka Angielskiego i Językoznawstwa Stosowanego oraz Zakładu Językoznawstwa Komputerowego i Korpusowego UŁ

BARBARA LEWANDOWSKA-TOMASZCZYK
Uniwersytet Łódzki, Łódź
blt@uni.lodz.pl

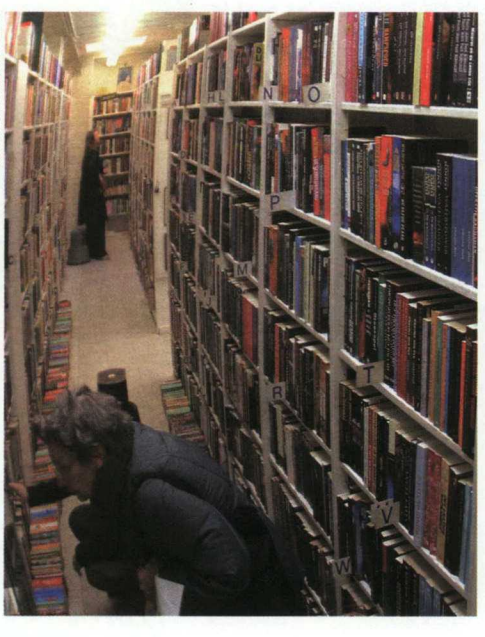
Dr hab. Marek Łaziński jest zatrudniony w Instytucie Języka Polskiego UW

MAREK ŁAZIŃSKI
Instytut Języka Polskiego, Warszawa
Uniwersytet Warszawski
M.Lazinski@uw.edu.pl

Jeśli przedmiotem badań językoznawcy jest język, to powstaje pytanie, jak ma on dotrzeć do przedmiotu swych badań

Do poznania języka prowadzą dwie drogi - podobnie jak on sam istnieje na dwa sposoby. Z jednej strony jest to pewna dyspozycja psychiczna, w jaką są wyposażeni niemal wszyscy ludzie (wyjątek stanowią osoby głęboko upośledzone lub takie, które doznały głębokiego uszkodzenia mózgu), umiejętność tworzenia i rozumienia wypowiedzi, w lingwistyce zwana kompetencją językową. Z drugiej strony język to produkt tej dyspozycji psychicznej, czyli mówione i pisane teksty. Językoznawca może więc albo badać tę psychiczną stronę języka, mówiąc obrazowo - gramatykę i słownik, jaki nam wszystkim, gdyśmy byli dziećmi i uczyliśmy się języka, włożono do głowy, albo też może badać teksty, po to by na ich podstawie ową gramatykę i ów słownik rekonstruować. Obie metody dają się uzasadnić, zapewne obie są jednakowo poprawne, choć nie każda daje się zastosować do konkretnego problemu badawczego.

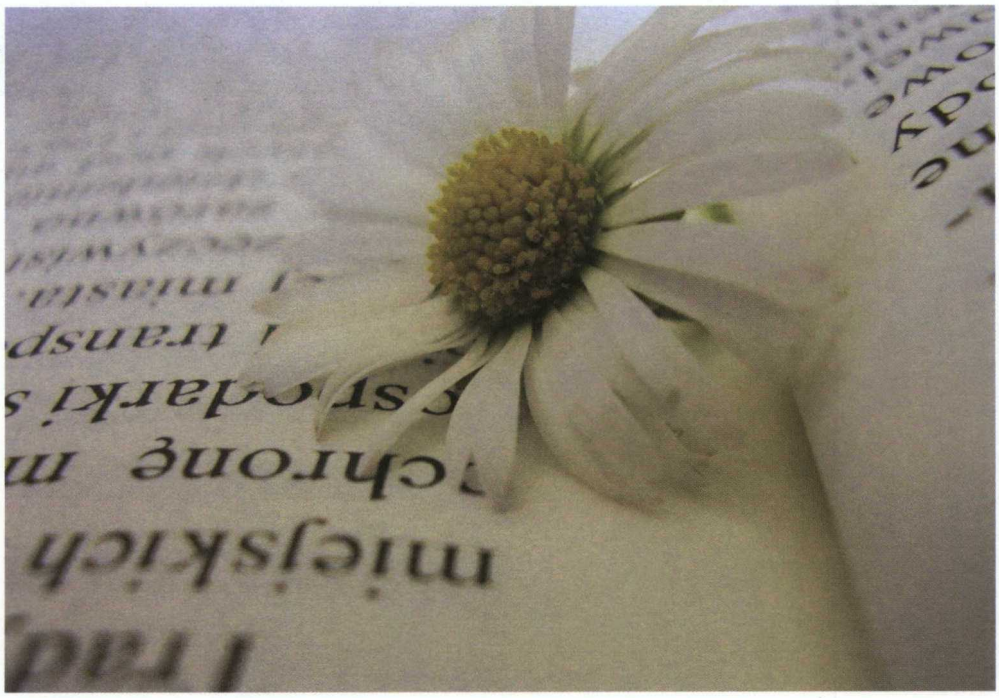
Aby dotrzeć do kompetencji językowej, najprościej zapytać rodzimego użytkownika języka o to, czy dane zdanie jest - jego zdaniem - zrozumiałe, „w porządku", „udane", czyli zgodne lub niezgodne z jego kompetencją językową. Nas jednak interesuje to drugie podejście - badanie tekstów. Przyznajmy od razu, że ta metoda przez co najmniej pół wieku była w odwrocie. Pomijając względy czysto naukowe, naukowców zniechęcało pracochłonne przeszukiwanie tekstów karta po karcie w poszukiwaniu interesujących ich zjawisk. Na szczęście ten nudny i żmudny etap pracy to wymarzone zajęcie dla komputera. Na taki pomysł wpadli językoznawcy z Uniwersytetu Browna w połowie lat sześćdziesiątych. Postanowili oni zdygitalizować - przenieść do komputera - pewną próbkę tekstów wziętą z amerykańskich książek i gazet. I tak powstał pierwszy elektroniczny korpus, tzw. Brown University Standard Corpus of Present-Day American English, notabene dotychczas niekiedy używany przez naukowców na całym świecie. Miał on wtedy niewyobrażalną liczbę 1 miliona słów. Dla porównania jest to około 700 razy więcej niż niniejszy tekst. Co ciekawe, niewiele później zaczęto tworzyć pierwszy polski korpus elektroniczny, niestety, prace nad nim trwały niezwykle długo. I trzeba szczerze przyznać, że polszczyzna, jeden



Herman Brinkman/www.src.hu

Mówimy żartem „ojca szwagra żony brat", ale tak poważnie: ile rzeczowników w dopełniaczu może się pojawić realnie w tekście? Prosta kwerenda w korpusie zwraca nam ciąg np. propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju

Anna Piątkowska



Budowany przez nas korpus już teraz znajduje zastosowanie w leksykografii. Redaktor słownika dysponuje dziś znacznie obfitszą egzemplifikacją użycia opracowywanego słowa, niż to miało miejsce w epoce fiszek. Komputer pozwala ten niezwykle obfity materiał wstępnie analizować

z ważniejszych (przynajmniej jeśli chodzi o liczbę rodzimych użytkowników) języków Europy, nie doczekała się korpusu spełniającego wszystkie wymogi współczesnej nauki.

Trochę historii

Prace nad polskimi korpusami nabrały impetu dopiero na przełomie tysiąclecia. Jako pierwszy powstał korpus Instytutu Języka Polskiego PAN (nie jest on dostępny publicznie), kolejno korpus Wydawnictwa Naukowego PWN, Korpus grupy PELCRA z Uniwersytetu Łódzkiego i wreszcie Korpus Instytutu Podstaw Informatyki PAN. Te cztery zespoły połączyły swoje siły (i zasoby) w roku 2006, by wspólnie wystąpić o finansowanie projektu. Projekt został wsparty grantem Ministerstwa Nauki i Szkolnictwa Wyższego (nr R17 003 03). Trzeba zaznaczyć, że każdy z korpusów miał inne mocne i słabe strony, toteż każdy z zespołów wnosi inne doświadczenia.

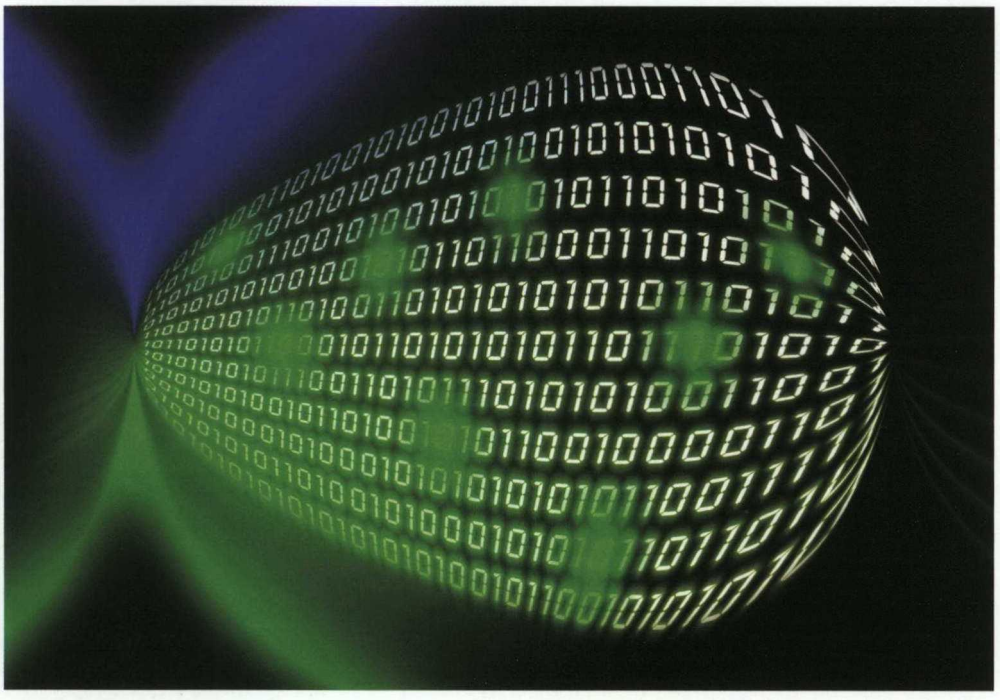
Konsorcjum Narodowego Korpusu Języka Polskiego tworzą: Instytut Podstaw Informatyki PAN, Instytut Języka Polskiego PAN, Katedra Języka Angielskiego i Językoznawstwa Stosowanego Uniwersytetu Łódzkiego oraz Wydawnictwo Naukowe PWN. Projektem realizowanym w IPI PAN kieruje doc. dr hab. Adam Przepiórkowski.

Projekt zakłada stworzenie korpusu wielkości 800 mln - 1 mld słów, a więc o 3 rzędy wielkości większego od korpusu Browna. Oczywiście nie da się stworzyć korpusu tej wielkości inaczej jak z przypadkowo dobranych tekstów. Dla wielu zastosowań istotniejsza będzie część zrównoważona, to znaczy taka, która stanowi reprezentatywną próbkę tekstów, a ściślej reprezentację języka, jaki dociera do przeciętnego członka polskiej społeczności językowej. W tym wypadku teksty będą dobierane według pewnego klucza, choć na pewno nie będzie to dobór czysto losowy, jesteśmy bowiem związani koniecznością honorowania praw autorskich i możemy włączać do korpusu teksy tylko za zgodą ich właścicieli. Część zrównoważona ma liczyć 300 mln słów tekstowych. Warto dodać, że tego rodzaju korpusy tworzone poza Polską liczą najczęściej 100 mln słów - jest to nieformalny standard narzucony przez British National Corpus. Szczególnym komponentem będzie kilkumilionowy zapis spontanicznych rozmów nagrywanych, a następnie transkrybowanych na użytek korpusu. Pozwoli to badać język mówiony, który od pisanego różni się nie tylko kanałem, lecz także wieloma specyficznymi cechami.

Poszukiwanie tekstów

Już dowolny zdygitalizowany zbiór tekstów może oddać językoznawcom pewne usługi. Im więcej jednak informacji jest zaszytych w korpusie, tym bardziej zwiększa się jego potencjał. Zacznijmy od tego, że

Przeszukiwanie języka



www.sxc.hu

Także zainteresowany językiem laik może dzięki korpusowi zaspokoić swoją ciekawość choćby po to, by zlokalizować interesujący go cytat, przyjrzeć się funkcjonowaniu tzw. skrzydlatych słów, czy wreszcie rekonstruować znaczenia wyrazów

w planowanym korpusie każdy tekst będzie opatrzony dokładnym adresem bibliograficznym, a także danymi dotyczącymi daty powstania czy typu tekstu. To pozwala każdemu cytatowi z korpusu przypisać autora i tytuł tekstu, z którego pochodzi. Dodatkowe informacje w rodzaju daty powstania tekstu lub jego przynależności gatunkowej pozwalają uchwycić różnice między stylami języka czy też subtelne zmiany, jakie zaszły w ciągu ostatnich lat. Podobnemu celowi służą dane demograficzne uczestników rozmów.

Każde słowo w korpusie jest opatrzone opisem gramatycznym, dzięki czemu można w nim poszukiwać nie tylko wyrazów, lecz także form gramatycznych. Można więc poszukać w korpusie np. wszystkich przymiotników w celowniku i liczbie mnogiej. Planujemy identyfikację w korpusie grup składniowych, choć z pewnością nie będzie to pełen rozbiór gramatyczny zdania. Inne narzędzie informatyczne, które powstaje na użytek projektu, to program rozpoznający nazwy własne (również wieloczłonowe) w tekście. Wreszcie - raczej w charakterze prototypu zostanie wdrożony system automatycznego rozpoznawania znaczeń: dla kilkuset słów zostanie wyznaczona pewna liczba znaczeń i w konkretnych zdaniach będzie wskazane, które znaczenie tu występuje. Naturalnie ze względu na wielkość korpusu wszystkie te zadania mogą być wykonane tyko automatycznie.

Do przeszukiwania korpusu mamy do dyspozycji dwa narzędzia: Poliqarp stworzony w IPI PAN oraz PELCRA stworzony na UŁ. Ich główne zadanie to tworzenie konkordancji, czyli pokazują one wszystkie wystąpienia poszukiwanego słowa (czy słów) z najbliższym, kilkuwyrazowym kontekstem. Ponadto pozwalają sortować konkordancje i wyszukiwać typowe połączenia, lokalizować cytaty czy wreszcie podają one liczbę wystąpień w korpusie szukanego elementu. Dla znakomitej większości użytkowników dostęp przez te narzędzia będzie wystarczający. Dla pewnych specyficznych zastosowań niezbędne okaże się przeszukiwanie za pomocą tworzonego ad hoc oprogramowania. O ile w tym drugim wypadku dostęp może być w pewnej mierze ograniczany ze względu na bezpieczeństwo tekstów (autorzy i wydawcy godzą się na współpracę z nami, pod warunkiem że nie stanowi to konkurencji dla wydanej książki), to dostęp przez Internet za pośrednictwem obu wyszukiwarek jest bezpłatny i nieograniczany wymogami rejestracji.

Niezwykłe przykłady

Jakie zastosowania może mieć korpus? Oczywiście przede wszystkim językoznawcze. Można go potraktować jako źródło przykładów, np. wyszukując wystąpienia jakiegoś słowa czy konstrukcji. Mówimy żartem „ojca szwagra żony brat", ale tak poważnie: ile rze-

czowników w dopełniaczu może się pojawić realnie w tekście? Prosta kwerenda w korpusie zwraca nam ciąg np. *propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju*. Jest to jednak zastosowanie najbardziej elementarne. Szersze wprowadzenie korpusów do badań językoznawczych pozwala uzupełnić opis jakościowy o opis ilościowy, a tym samym rekonstruować nie tylko „sztywne" reguły językowe, lecz także tendencje, które bywają nieraz przekraczane, ale które dają się obserwować. Pozwalają one oddzielić to, co typowe w języku, od tego, co - wprawdzie najzupełniej akceptowalne - jednak marginalne. Dzięki nim dostrzega się, jak dalece mowa składa się z pewnych prefabrykatów. Z jednej strony jesteśmy niezwykle twórczy, tworzymy wciąż nowe, niepowtarzalne zdania i połączenia wyrazowe, z drugiej strony jednak w naszych wypowiedziach pojawiają się wciąż te same ciągi słów. Wyobraźmy sobie zjawisko językowe, które występuje przeciętnie raz na pół miliona słów. Kiedy czytamy teksty, zjawisko to niknie wśród tysięcy innych. Jeśli jednak stworzymy wspomnianą konkordancję, zobaczymy naraz kilkaset przykładów tego zjawiska.

Nastroje w komputerze

Budowany przez nas korpus już teraz znajduje zastosowanie w leksykografii. Słowniki zawsze były tworzone na podstawie tekstów, niemniej korpusy stanowią w leksykografii nową jakość. Redaktor słownika dysponuje zazwyczaj znacznie obfitszą egzemplifikacją użycia opracowywanego słowa, niż to miało miejsce w epoce fiszek, ale też komputer pozwala ten niezwykle obfity materiał wstępnie analizować.

Nie sposób pominąć roli korpusów w dydaktyce. Na świecie podejmowane są próby zastosowania korpusów w nauczaniu języka jako obcego - dotyczy to głównie angielskiego - na poziomie zaawansowanym (uniwersyteckim). Korpus to też podstawa lepszych (bo opartych na rzeczywistym użyciu) materiałów dydaktycznych. Wreszcie z pewnością może stać się ciekawym urozmaiceniem nauczania wiedzy o języku w liceach. Także zainteresowany językiem laik może zaspokoić swoją ciekawość choćby po to, by zlokalizować interesujący go cytat, przyjrzeć się funkcjonowaniu tzw. skrzydlatych słów czy wreszcie rekonstruować znaczenia wyrazów.

Najszersze kręgi społeczne zainteresuje jednak przede wszystkim zastosowanie korpusu i narzędzi do przetwarzania języka naturalnego powstających na jego potrzeby w inżynierii lingwistycznej. W dobie, gdy większość tekstów pisanych powstaje w wersji elektronicznej, automatyczne wyszukiwanie informacji przekazywanej w języku naturalnym jest zadaniem coraz istotniejszym. Wszystkie opisane wyżej narzędzia, które służą do przygotowania i przetwarzania korpusu, są przydatne w inteligentnym wyszukiwaniu informacji, tłumaczeniu automatycznym, badaniu nastrojów społecznych (drogą analizowania w czasie rzeczywistym wielu tekstów gazetowych i internetowych), telefonii (komponent pisany to 3 mln tekstów, które są dostępne równolegle jako nagranie i tekst pisany) itp.

Wspomnieliśmy, że polska nauka pozostawała nieco w tyle, zarówno jeśli chodzi o tworzenie korpusów, jak i oparte na nich badania. Opisywany projekt pozwala żywić nadzieję, że stan ten ulegnie radykalnej zmianie.

Wstępna wersja korpusu jest dostępna pod adresem http://www.nkjp.pl ■

Narzędzia, które służą do przygotowania i przetwarzania korpusu, są również przydatne w badaniu nastrojów społecznych, pozwalają bowiem na analizowanie w czasie rzeczywistym wielu tekstów gazetowych i internetowych

Chcesz wiedzieć więcej?

Lewandowska-Tomaszczyk B. (Red.). (2005). *Podstawy językoznawstwa korpusowego*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Przepiórkowski A. (2004). *Korpus IPI PAN. Wersja wstępna* [*The IPI PAN Corpus: Preliminary version*]. Warszawa: IPI PAN.

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

sanja gjenero/www.sxc.hu