1

Narodowy Korpus Języka Polskiego: geneza i dzień dzisiejszy

Barbara Lewandowska-Tomaszczyk, Mirosław Bańko, Rafał L. Górski, Marek Łaziński, Piotr Pęzik, Adam Przepiórkowski

Nareszcie mamy narodowy korpus polszczyzny – dostępny publicznie i w dodatku bezpłatnie. Użytkownicy, którzy zechcą wykorzystać go do swoich celów, mogą uznać, że jest niedostatecznie zróżnicowany, nie dość precyzyjnie oznakowany – te i inne zarzuty łatwo sobie wyobrazić. Łatwo je także odparować: praca takich rozmiarów jest wynikiem różnych kompromisów, poza tym trudno jest wszystkich zadowolić.

Być może niektórzy uznają, że NKJP się niewłaściwie nazywa, gdyż słowo *narodowy* ma w polskim uchu konotacje patriotyczne i w nazwie projektu naukowego brzmi pretensjonalnie. Wątpliwości związane z nazwą twórcy korpusu sami mieli na początku prac, ale wydaje się, że niesłusznie; słowo *narodowy* w nazwie NKJP kontynuuje tradycje polskiej leksykografii. Przypomnijmy, że Linde swój słownik nazywał narodowym (zob. Matuszczyk 2006: 99–104) i że pod nagłówkiem „słowniki narodowe" zgrupował podobne dzieła Grzegorczyk (1967) w swojej bibliografii. Także w leksykografii encyklopedycznej funkcjonuje kategoria encyklopedii narodowych (np. Olkiewicz 1988). Co więcej, jak wiele innych narodów mamy Bibliotekę Narodową, której zbiory z upływem czasu będą coraz bardziej dygitalizowane, co w końcu upodobni ją do korpusu, a nawet uczyni korpusem *in potentia*.

1.1. Krótki rys historyczny

Od lat sześćdziesiątych dwudziestego wieku wzrastało zainteresowanie językoznawców badaniem częstotliwości użycia różnych form językowych oraz prawdopodobieństwa ich występowania w różnych odmianach języka. Zainteresowanie

to zbiegło się w czasie z dynamicznym rozwojem komputerów i nowych technik informacyjnych. Komputer pozwolił językoznawcom na magazynowanie i szybką obróbkę dużej ilości danych językowych. Nowo powstała dyscyplina lingwistyczna, *językoznawstwo korpusowe*, rozwijająca się obecnie niezwykle dynamicznie w wielu krajach świata, pozwoliła na badania języka na szeroka skalę. Zbiory językowe, gromadzone w *korpusach językowych* czyli komputerowych zbiorach autentycznych tekstów językowych, mówionych i pisanych, reprezentują różne odmiany, style i typy tekstu.

Pierwszy powszechnie używany językowy korpus komputerowy to korpus amerykańskiej odmiany języka angielskiego, zawierający zaledwie jeden milion wyrazów, ale bezprecedensowy wtedy, zebrany w latach sześćdziesiątych ubiegłego wieku przez Henry'ego Kučerę i Nelsona Francisa (Kučera i Francis 1967). Obecnie zbiory językowe sięgają setek milionów, a nawet miliardów, jednostek. Autentyczne materiały korpusowe używane są dziś do różnych zadań, takich jak opracowywanie słowników i materiałów dydaktycznych, przekładu, jak również do studiów literackich, kulturowych i in. Komputery umożliwiają magazynowanie i analizowanie dużo większych zbiorów materiału językowego niż tradycyjne metody językoznawcze. Ponadto, co jest szczególnie ważne, pozwalają na weryfikowanie intuicji językowych, pokazując użycie i frekwencję zarówno form krótszych, jak i wzorców współwystępowania różnych form językowych i ich częstotliwości użycia. Pozwalają także na szeroko zakrojone badania języka różnych grup użytkowników, jak i kwantyfikację wyników badań dialektologicznych czy socjolingwistycznych na dużą skalę w zależności od typu dyskursu, stylu, czy indywidualnych preferencji użytkownika języka. Zbiory językowe w korpusach są także niezbędne do zastosowań w szerszej gamie działań językoznawstwa komputerowego, np. dla celów przetwarzania języka naturalnego z wykorzystaniem metod uczenia maszynowego.

Najliczniejsze i najbardziej różnorodne zbiory zawierają narodowe korpusy angielskie – brytyjskie (British National Corpus, Bank of English i in.) i amerykańskie (Corpus of Contemporary American English, American National Corpus, Google Books: American English i in.). Zbiory innych języków rozwijają się dynamicznie. Są to m.in. chiński korpus mandaryńskiego, korpusy języka niderlandzkiego i duńskiego, estoński korpus języka prawa, korpusy odmian języka francuskiego, zbiory tekstów niektórych języków celtyckich (język irlandzki oraz język mański), włoski korpus CORIS/CODIS, teksty norweskie *Oslo Corpus of Tagged Norwegian Texts*, korpus brazylijskiej odmiany języka portugalskiego, „bank językowy" tekstów szwedzkich, korpusy europejskiej i południowo-amerykańskich odmian języka hiszpańskiego, korpus języka tagalog, korpus literatury malajskiej i korpusy innych języków południowo-wschodniej Azji.

Od kilku lat szczególnie dynamicznie rozwijają się korpusy narodowe innych języków słowiańskich. Od roku 1994, kiedy rozpoczęły się prace nad Czeskim Korpusem Narodowym, powstały wielkie referencyjne korpusy większości języków słowiańskich, z których chorwacki, czeski, polski, rosyjski, słowacki i słoweński dostępne są w Internecie z wyszukiwarkami morfologicznymi (trzeba podkreślić, że morfologia słowiańska stanowi dla technologii większe wyzwanie niż romańska czy germańska). Narodowy Korpus Języka Poskiego współorganizował konferencję założycielską Komisji Korpusowej Międzynarodowego Komitetu Slawistów (por. Slavicorp 2012).

Najbardziej naturalnym użyciem korpusów językowych było zawsze zastosowanie autentycznych materiałów językowych w pracach słownikowych. Pierwszy słownik języka polskiego oparty na korpusach w nowoczesnym rozumieniu tego pojęcia to listy frekwencyjne różnych odmian polszczyzny z lat sześćdziesiątych dwudziestego wieku (Kurcz i in. 1974a, b, 1976, 1977, Lewicki i in. 1975). Dane te posłużyły także do opracowania nowego polskiego słownika frekwencyjnego (Kurcz i in. 1990).

Budowa narzędzi korpusowych jest równoległym do zbierania danych zestawem działań językoznawstwa korpusowego, niezbędnym do efektywnego używania korpusów. Języki fleksyjne, takie jak język polski, stanowią spore wyzwanie dla automatycznej analizy morfologicznej i składniowej. W języku polskim częsta jest wieloznaczność form gramatycznych, zarówno między częściami mowy, jak również w kręgu kategorii gramatycznych, takich jak liczba i przypadek. Szczególnie istotne dla automatycznego rozpoznawania form wyrazowych są więc propozycje autorów analizatorów morfologicznych języka polskiego, programów ujednoznaczniających oraz testów do ich weryfikacji i oceny.

W Polsce aktywnie działało od lat dziewięćdziesiątych kilka grup językoznawczych, informatycznych i leksykograficznych, które zajmowały się zarówno zbieraniem danych korpusowych, jak i tworzeniem narzędzi do ich opracowywania, m.in. zespoły w Instytucie Podstaw Informatyki Polskiej Akademii Nauk (`http://nlp.ipipan.waw.pl/`), zespół w Instytucie Języka Polskiego Polskiej Akademii Nauk w Krakowie (`http://www.ijp-pan.krakow.pl/`), zespół korpusowy PELCRA (`http://pelcra.pl/`) w Katedrze Języka Angielskiego i Językoznawstwa Stosowanego w Uniwersytecie Łódzkim oraz zespół korpusowy w Redakcji Słowników Języka Polskiego Wydawnictwa Naukowego PWN (`http://korpus.pwn.pl/`).

IPI PAN Zespół Inżynierii Lingwistycznej został założony – i był przez wiele lat kierowany – przez prof. Leonarda Bolca. W latach dziewięćdziesiątych ubiegłego wieku działalność zespołu dotyczyła głównie przetwarzania składniowego języka polskiego, przede wszystkim w ramach teorii formalnej Head-driven Phrase

Structure Grammar. Ukoronowaniem tej działalności była wydana w 2002 roku monografia *Formalny opis języka polskiego: teoria i implementacja* (Przepiórkowski i in. 2002).

Zespół rozpoczął badania korpusowe na początku obecnego stulecia. Ponieważ jednak w tym czasie nie był publicznie dostępny duży i odpowiednio znakowany korpus języka polskiego, w lipcu 2000 roku został złożony do Komitetu Badań Naukowych wniosek o finansowanie budowy Korpusu IPI PAN. Realizacja projektu trwała od kwietnia 2001 do marca 2004 roku i zaowocowała stworzeniem ówcześnie największego – choć nie zrównoważonego ani nie reprezentatywnego w żadnym sensie – korpusu języka polskiego. Korpus ten liczył 250 milionów segmentów (ponad 200 milionów tradycyjnie rozumianych słów) i był pierwszym dużym polskim korpusem znakowanym morfosyntaktycznie (`http://korpus.pl/`).

Obecnie prace zespołu koncentrują się wokół kilku projektów europejskich: CLARIN, CESAR (w ramach konsorcjum META-NET) i ATLAS, oraz krajowych: NEKST (*Adaptacyjny system wspomagający rozwiązywanie problemów w oparciu o analizę treści dostępnych źródeł elektronicznych*), *Budowa banku drzew składniowych dla języka polskiego z wykorzystaniem automatycznej analizy składniowej* i *Komputerowe metody identyfikacji nawiązań w tekstach polskich*, a także – w mniejszym stopniu – SYNAT (*Utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy*).

IJP PAN Początki prac nad korpusem IJP PAN ściśle wiążą się z zarzuconym później pomysłem słownika języka polskiego. Korpus miał być dla niego bazą empiryczną. Gdy w początkach lat dziewięćdziesiątych okazało się, że w ówczesnej sytuacji projekt nie może dojść do skutku, korpus zaczęto wykorzystywać do badań prowadzonych w Instytucie, w szczególności stanowił podstawowe źródło dla badań nad semantyką i składnią czasowników polskich, projektu badawczego, który ma zaowocować słownikiem. Był też przez cały czas rozbudowywany, nie planowano natomiast publicznego udostępnienia tego korpusu, jakkolwiek – szczególnie zanim powstał korpus PWN – korzystało z niego wielu naukowców spoza IJP. Na marginesie warto dodać, że w IJP PAN został stworzony również korpus staropolski, obejmujący wszystkie ciągłe teksty do roku 1500. Korpus ten jest publicznie dostępny na płytach CD (Twardzik 2006). To jak dotąd jedyny korpus historyczny języka polskiego.

PELCRA UŁ Jedną z najstarszych grup językoznawstwa korpusowego jest zespół PELCRA, działający w Katedrze Języka Angielskiego Uniwersytetu Łódzkiego od 1995 roku, początkowo we współpracy ze znanym językoznawcą

korpusowym Anthonym McEnerym i pracownikami Katedry Językoznawstwa i Współczesnego Języka Angielskiego w Uniwersytecie w Lancaster w Wielkiej Brytanii. Były to też lata, w których powstawał największy znany korpus językowy British National Corpus, czyli Brytyjski Korpus Narodowy. Kontakty te zaowocowały wspólnymi badaniami rozpoczętymi w ramach projektu, który został nazwany PELCRA (*Polish and English Language Corpora for Research and Application*). Obecne zasoby zespołu PELCRA to blisko stumilionowy korpus PELCRA, reprezentujący proporcjonalnie różne typy i odmiany tekstów, zarówno języka mówionego jak i pisanego, dostępny w witrynie `http://korpus.ia.uni.lodz.pl/`, z wyszukiwarką, której współautorem jest Piotr Pęzik, oraz korpus tekstów mówionych liczący ponad 600 tysięcy słów.

Polski Korpus Uczniowski Języka Angielskiego w zasobach PELCRA, który liczy obecnie 1,5 miliona słów, docelowo zaś – 3 miliony, zawiera angielskie dane językowe Polaków uczących się języka angielskiego (grant MNiSW nr N N104 205039, kierowany przez Piotra Pęzika). W toku są prace nad budową korpusów paralelnych i porównywalnych angielsko-polskich i polsko-angielskich (europejski projekt CESAR / META-NET, numer ICT-PSP 271022).

Prace badawcze członków zespołu PELCRA koncentrują się na kilku kręgach tematycznych. Należy do nich analiza i opis języka angielskiego, także w aspekcie kontrastywnym w porównaniu z językiem polskim oraz języka polskiego – na tle różnic i podobieństw z językiem angielskim. Ponadto znajdują się wśród nich badania nad rozwijaniem technik i materiałów do nauczania języków obcych, analizy przekładów oraz studia nad problematyką interkulturowości.

PWN Zespół korpusowy utworzono w Redakcji Słowników Języka Polskiego w roku 1997 w trakcie prac nad pierwszym słownikiem języka polskiego opartym na korpusie – *Innym słownikiem języka polskiego* (Bańko 2000) oraz bardziej tradycyjnym *Uniwersalnym słownikiem języka polskiego* (Dubisz 2003). Personalnie zespół Korpusu Języka Polskiego PWN i Redakcji Słowników Języka Polskiego był w naturalny sposób związany z Wydziałem Polonistyki Uniwersytetu Warszawskiego. Korpus PWN miał być podstawą opisu leksykograficznego i bazą różnorodnych przykładów ilustrujących znaczenie haseł w słownikach. Dlatego największy nacisk położono w nim na różnorodność tematyczną i stylistyczną tekstów, a także na ich reprezentatywność z punktu widzenia polskiej tradycji literackiej (w skład korpusu weszła m.in. cała lista lektur z literatury polskiej na poziomie programu maturalnego). Korpus PWN był także wykorzystywany przez językoznawców spoza wydawnictwa. Zrównoważony korpus z czasem osiągnął nieco ponad 100 milionów słów, a 40-milionowa próbka dostępna jest w Internecie z prostą wyszukiwarką: `http://korpus.pwn.pl/`.

1.2. Narodowy Korpus Języka Polskiego

Mimo opisanych powyżej sukcesów sytuacji nie można było uznać za zadowalającą. Wciąż bowiem nie dysponowaliśmy korpusem, który byłby równocześnie duży, zróżnicowany, reprezentatywny i anotowany morfosyntaktycznie. Realizowane projekty stworzyły narzędzia badawcze, które spełniały tylko niektóre z powyższych postulatów. Równocześnie narastało przekonanie, że dotychczasowe rozproszenie wysiłków jest niekorzystne[1]. W roku 2006, z inspiracji Prezydium Komitetu Językoznawstwa Polskiej Akademii Nauk i przy poparciu jego ówczesnego przewodniczącego prof. dr. hab. Stanisława Gajdy, zawiązało się konsorcjum *Narodowy Korpus Języka Polskiego* (`http://nkjp.pl/`), w skład którego weszły trzy ośrodki akademickie: IPI PAN, reprezentowany przez Adama Przepiórkowskiego, IJP PAN (Rafał Górski) i UŁ (Barbara Lewandowska-Tomaszczyk), oraz wydawnictwo PWN (Mirosław Bańko i Marek Łaziński). W następnym roku zespół NKJP uzyskał grant rozwojowy (nr R17 003 03, od grudnia 2007 do grudnia 2010, przedłużony do czerwca 2011 roku), którego głównym celem była budowa NKJP oraz opracowanie narzędzi do jego wykorzystywania. Za zgodą wszystkich członków nowo powstałego zespołu koordynatorem projektu został IPI PAN, jego kierownikiem zaś dr hab. Adam Przepiórkowski.

Znaczące części korpusów partnerów projektu weszły w skład Narodowego Korpusu Języka Polskiego. IPI PAN przekazało NKJP cały zebrany wcześniej korpus, liczący ponad 200 milion słów. Z zasobów Korpusu Języka Polskiego PWN do NKJP weszło na początku projektu 50 milionów słów, kolejne 200 milionów tekstów książkowych, w tym w dużej części literackich oraz prasowych, zebrano w trakcie realizacji projektu. Z zasobów korpusu PELCRA do Narodowego Korpusu Języka Polskiego także weszło kilkadziesiąt milionów słów, dalsze ponad 660 milionów zostało zebranych w pracach projektowych. Ponad 2 miliony słów stanowią nowe dane mówione (w skład korpusu zostaje włączonych 1,9 miliona słów danych konwersacyjnych, które NKJP zamierza udostępnić na licencji niekomercyjnej). Najnowsze materiały zespołu PELCRA to kilkusetmilionowe zbiory tekstów internetowych, których istotna część zasiliła zasoby NKJP.

W ramach projektu powstały i były rozwijane różne formy narzędzi korpusowych. Piotr Pęzik z zespołu PELCRA jest autorem kilku aplikacji korpusowych dla NKJP: wyszukiwarki korpusowej do danych NKJP, wyszukiwarki do danych

[1] Z pewnością trudno to początkowe rozproszenie wysiłków uznać za zjawisko pozytywne. Należy na to popatrzeć także z drugiej strony. NKJP tworzył zespół badaczy, którzy po pierwsze, mieli już niemałe doświadczenie w zakresie opracowywania korpusów, a po drugie intensywnie używali ich w swoich badaniach, zdawali więc sobie sprawę z tego, czego od korpusu oczekuje jego użytkownik.

mówionych NKJP (`http://nkjp.uni.lodz.pl/spoken.jsp`) oraz dla zainteresowanych wyszukiwarki SlopeQ do BNC. Jest też autorem kolokatora do NKJP oraz Automatycznego Słownika Kolokacji (ASK) (dostępna wersja demo). Zmodyfikowana została także przeszukiwarka Poliqarp, służąca wcześniej do obsługi Korpusu IPI PAN. Stworzono także liczne narzędzia do przetwarzania tekstów, szczegółowo opisane w dalszych rozdziałach; są to m.in.: tager morfosyntaktyczny PANTERA, nowa gramatyka powierzchniowa języka polskiego dla parsera Spejd, narzędzia identyfikujące w tekstach jednostki nazewnicze oraz prototypowy system ujednoznaczniania sensów słów.

Dziś, w roku 2011, możemy stwierdzić, że Narodowy Korpus Języka Polskiego, największy, morfologicznie anotowany zbiór danych języka polskiego, jest faktem. NKJP jest dostępny bezpłatnie dla wszystkich chętnych i zainteresowanych, tom zaś, który Państwu przedstawiamy to pokłosie prawie czteroletnich wysiłków zespołu NKJP.

Z korpusu korzysta już bardzo wielu użykowników. Danymi NKJP posługuje się przede wszystkim zespół powstającego *Wielkiego słownika języka polskiego* (Żmigrodzki i in. 2007), jest on również wykorzystywany w badaniach kontrastywnych z udziałem języka polskiego (np. Dziwirek i Lewandowska-Tomaszczyk 2010), w praktyce i dydaktyce translatoryki (Pęzik 2011), a także w Poradni Językowej PWN. Wydawnictwa UW zapowiadają już nawet pierwszy słownik powstały na podstawie NKJP – nosi on tytuł *Ludzie i miejsca w języku*, a gromadzi frazy odimienne typu *puszka Pandory* lub *jajko Kolumba* (autorami są Maciej Czeszewski i Katarzyna Foremniak).

Nie jest to naturalnie koniec naszych prac. Czekają nas dalsze wyzwania. Aby pozostał największym narzędziem referencyjnym dla rzeszy użytkowników, NKJP musi być ustawicznie uzupełniany, ulepszany, nadzorowany i monitorowany. Ponadto konstrukcja zarówno programów automatycznego wydobywania znaczeń z danych korpusowych, jak i programów ujednoznaczniających to problematyka, z którą zmaga się duża część językoznawstwa informatycznego. Naturalne jest także, że dla rozpowszechniania zarówno samych danych NKJP, jak i narzędzi do ich wykorzystywania konieczne są wersje różnych aplikacji, mogących zdalnie obsługiwać jednocześnie wielu użytkowników, o różnych potrzebach poznawczych i celach aplikacyjnych.

1.3. Podziękowania

Prace opisane w niniejszej publikacji były finansowane ze środków na naukę w latach 2007–2011 w ramach projektu rozwojowego „Narodowy Korpus Języka Polskiego" (nr R17 003 03). Spis licznych dobroczyńców NKJP – przede wszystkim

wydawnictw i autorów, ale także innych grup badawczych i zespołów projektów, z którymi NKJP współpracował – oraz wykonawców niniejszego projektu, znajduje się na stronach `http://nkjp.pl/` (w zakładkach PODZIĘKOWANIA, TEKSTY KORPUSU i ZESPÓŁ). Wszystkim tym osobom i instytucjom oraz rzeszy innych, którzy pomagali nam w pracach, w tym miejscu raz jeszcze dziękujemy.