Monika Korczakowska

Polskie korpusy tekstów : wybrane zagadnienia

Prace Językoznawcze 3, 65-75

2001

Monika Korczakowska
Olsztyn

Polskie korpusy tekstów
(wybrane zagadnienia)

Polish text corpora (selected issues)

In the article, Polish text corpora are presented and the author discusses the issues connected with the corpora against the background of foreign experience in this field.

Celem artykułu jest zapoznanie polskiego czytelnika z dostępnymi polskimi korpusami tekstów oraz ukazanie problematyki korpusowej na tle doświadczeń obcych w tym zakresie. W pracy będę posługiwała się terminami „słowo", „forma wyrazowa", „leksem". Zgodnie z rozróżnieniem Zygmunta Saloniego „słowem" określam ciąg liter od znaku odstępu do znaku odstępu lub do równorzędnego separatora tekstu, „formą wyrazową" – słowo z przypisaną charakterystyką gramatyczną, „leksemem" – „zbiór form wyrazowych o identycznej lub regularnie zróżnicowanej charakterystyce semantycznej, pozostających względem siebie w regularnych opozycjach"[1].

Gromadzenie tekstów źródłowych i posługiwanie się cytatami w opisie języków naturalnych nie jest techniką XX-wieczną. Dla języka polskiego opracowania, w których pojawiają się oryginalne zdania jako przykłady użycia odpowiednich jednostek języka, powstały już w XVIII w. (*Nowy dykcjonarz to jest mownik polsko-niemiecko-francuski* Troca, 1764), a wykorzystujące teksty jako podstawę do opracowania listy haseł słownikowych jeszcze wcześniej, bo w XVII w.

Dla XIX-wiecznych słownikarzy (czy – jak mówił o sobie S.B. Linde – „zbierających tylko wyrazy") teksty literackie były podstawowym źródłem nie tylko przykładów, ale nawet definicji słownikowych. Zarówno Bandtkie, jak i Linde podali listę źródeł oraz dokładną – co do strony – lokalizację cytatów. Chociaż badacze dowiedli, że owe cytaty były często na potrzeby słownika zmieniane, nie ulega wątpliwości, że leksykografowie ci posługiwali się w swoich pracach obszernymi zbiorami tekstów.

---

[1] Por. Z. Saloni, H. Świdziński: *Składnia współczesnego języka polskiego*. Wyd. IV zmienione. Warszawa 1998, s. 85.

Gromadzenie materiału i jego ekscerpcja wymagały, przynajmniej do połowy lat sześćdziesiątych XX w., ogromnych nakładów pracy. Należało słowa tekstowe wraz z kontekstem rozpisać na oddzielne fiszki. W ten sposób gromadzono materiał jeszcze do *Słownika języka polskiego* pod red. W. Doroszewskiego (SJPDor), poddając wówczas ekscerpcji wyrywkowej ponad 3 tys. tekstów źródłowych[2]. Pełnej ekscerpcji natomiast poddano, tzn. rozpisano na tyle cytatów, ile znajdowało się w nich słów tekstowych, trzy teksty: *Niebo w płomieniach* Parandowskiego, *Dziewczęta z Nowolipek* Gojawiczyńskiej i dwa pierwsze tomy *Nocy i dni* Dąbrowskiej. W ten sposób powstała kartoteka papierowa PWN. Należy zaznaczyć, że kartoteka, w odróżnieniu od korpusu, nie jest tylko zbiorem tekstów, ale zbiorem cytatów uporządkowanych według pewnej zasady, np. kolejności alfabetycznej haseł.

Tradycyjnie w językoznawstwie **korpusem** nazywa się każdy zbiór zdań traktowany jako podstawa do analizy lingwistycznej danego języka. Obecnie terminu tego używa się na ogół w zawężonym znaczeniu. Technicznie jako korpus tekstu określa się wyłącznie zbiór zebrany w określonym celu badawczym, spełniający następujące warunki:

- udostępniony w wersji elektronicznej,
- odpowiednio obszerny,
- zrównoważony (tzn. zawierający teksty różnych stylów, gatunków, napisane przez wielu autorów na przestrzeni odpowiedniego odcinka czasu).

Zgromadzenie zbioru liczącego kilkadziesiąt milionów słów nie jest dziś zadaniem skomplikowanym i wymagającym kilkunastu lat mrówczej pracy. Większość redakcji prasowych przygotowuje gazety wykorzystując komputery. Najpoczytniejsze tytuły umieszczają je w sieci elektronicznej, tym samym dając czytelnikom natychmiastowy do nich dostęp bez względu na fizyczne utrudnienia w dotarciu do gazety w postaci papierowej. Do zebrania dużego zbioru potrzebny jest jedynie komputer z dostępem do Internetu. Zestaw zebrany w ten sposób byłby jednak reprezentatywny dla tekstów prasowych.

Dobór materiału podyktowany jest zawsze celem badań (idealny korpus). Do opisania systemu fonologicznego języka potrzebowalibyśmy stosunkowo najmniejszej liczby danych. Największego zbioru wymagałyby badania składniowe i leksykalne. Jeśli chcemy mieć korpus reprezentatywny dla współczesnej polszczyzny ogólnej, musimy zebrać próbki wszystkich odmian gatunkowych powstających na przestrzeni kilku lat, napisanych przez wielu autorów. Z oczywistych względów nie może to być zbiór w s z y s t k i c h tekstów danego języka żywego. Z ogromu materiału należy wybrać r e p r e z e n t a n t ó w.

Współczesne korpusy to w większości k o r p u s y o t w a r t e – gromadzące teksty wszystkich odmian i stylów współczesnego języka. Naturalna trud-

[2] Por. wykaz źródeł do SJPDor. T. I. Warszawa 1958, s. LXXV–CLV.

ność w kompilowaniu tego typu zbiorów polega na zrównoważeniu źródeł. Kłopotów tego rodzaju nie stwarzają korpusy języków martwych lub utworów określonego pisarza, czyli tzw. korpusy zamknięte. Zawierają one bowiem wszystkie teksty spełniające określony warunek[3].

Szczególnego korpusu wymagają badania nad językiem mówionym. Pierwszym etapem gromadzenia materiału jest nagranie różnego rodzaju rozmów, audycji radiowych i telewizyjnych, następnym – ich transkrypcja. Dopiero wtórnie zapisane zostają umieszczane w pamięci komputera[4].

Większość korpusów narodowych oraz gromadzonych na potrzeby opracowań słownikowych to zbiory najobszerniejsze i najbardziej zróżnicowane zarówno pod względem liczby danych, jak i różnorodności stylowej i chronologicznej. Zwykle zawierają one teksty dwu subkodów – pisanego i mówionego.

Tekstów zróżnicowanych jakościowo wymagają opisy zjawisk językowych w jednym języku naturalnym, porównujące te same zjawiska w dwu systemach językowych. Na potrzeby badań tego rodzaju gromadzi się korpusy wielojęzyczne, zwane korpusami równoległymi, na które składają się teksty wyprodukowane przez rodzimych użytkowników oraz tłumaczenia. Korzysta z nich językoznawstwo komparatystyczne, są nieocenionym materiałem w pracach nad tłumaczeniem automatycznym. Na podstawie korpusu języków skandynawskich w wersji mówionej porównywano użycia konstrukcji ze stroną bierną w norweskim i szwedzkim[5].

Użytkowanie współczesnego korpusu wiąże się z wykorzystaniem odpowiednich narzędzi informatycznych. Do ustalenia spisu słów zbioru, podzielonego nawet na kilkaset plików, lub listy posortowanej według podanej zasady (np. dla znalezienia wszystkich wystąpień danego słowa) konieczny jest program konkordancji. Odpowiednio ułożona konkordancja, np. według prawo- lub lewostronnego sąsiedztwa pozwala ustalić współwystępowanie odpowiednich jednostek, tzn. ich kolokacje, łączliwość. Oczywiście należy pamiętać, że nie są to dane całkowicie dokładne w tym sensie, że jednostki wchodzące ze sobą w związki strukturalne często mogą być od siebie linearnie oddalone. Problem może rozwiązać odpowiednio szeroki kontekst lub interwencja człowieka.

Przykładem korpusu narodowego jest anglojęzyczny zestaw teksów, znany jako *Bank of English*. Zbiór ten jest chyba największym – liczącym ponad 415 milionów słów – zrównoważonym korpusem języka naturalnego, istnieje

---

[3] Idealnym korpusem zamkniętym wydaje się być zbiór tekstów języka pruskiego, na który składają się dwa katechizmy. Prawdopodobieństwo odnalezienia innych źródeł jest znikome.

[4] W ten sposób teksty mówione do swojego korpusu gromadzi wydawnictwo PWN. I choć znane są już programy komputerowe do odczytywania i przetwarzania mowy brzmiącej, są one wysoko specjalistyczne, a co za tym idzie, niepowszechne i kosztowne.

[5] Cykl wykładów poświęconych temu zagadnieniu został wygłoszony przez Elisabet Engdahl podczas „Wiosennej szkoły lingwistycznej im. Vilema Mathesiusa" w Pradze w 2000 r.

od 1980 r. w Birmingham. Korpus jest częścią projektu COBUILD – współpracy wydawnictwa HarperCollins oraz uniwersytetu w Birmingham (Collins Birmingham University International Language Database). Na zbiór składają się teksty napisane i wypowiedziane po roku 1950, większość z nich pochodzi z ostatnich dziesięciu lat. Zbiór obejmuje głównie teksty brytyjskie (ok. 73,3%), ale i amerykańskie (ok. 21,1%) oraz inne, np. australijskie (ok. 5,6%). Na źródła pisane składają się: literatura piękna, niebeletrystyczna (tj. popularnonaukowa, „wspomnieniowa", poradniki), dzienniki, tygodniki, miesięczniki, broszury, ulotki, listy i inne. Język mówiony reprezentowany jest przez transkrybowane codzienne rozmowy, audycje radiowe, wywiady[6].

Wynikiem projektu jest jeden z najważniejszych słowników angielskich, zwanych pedagogicznymi, tzn. przeznaczonych dla bardziej zaawansowanych uczących się (nie native speakerów) języka angielskiego z ambicjami poprawnego pisania i mówienia w tym języku – *Collins COBUILD English Dictionary*. Korpus stał się podstawą opracowania listy haseł oraz poszczególnych znaczeń jednostek znajdujących się w słowniku.

W ramach projektu stworzono także bazę danych językowych, korpus-matkę, z rozbudowaną siecią odsyłaczy zdających sprawę z wzajemnych relacji porównywanych jednostek. Z tej bazy danych wyodrębniono kilka podkorpusów (również demonstracyjny dla zainteresowanych) i na ich podstawie opracowano inne słowniki (np. wymowy brytyjskiej i amerykańskiej, słowniki pedagogiczne), podręczniki gramatyki, materiały do nauki języka angielskiego. W bazie mieści się też *bank drzew* – reprezentacji składniowej zdań.

Z całego korpusu w celach naukowych można korzystać na miejscu – w Birmingham. Jego część (56 milionów słów) jest dostępna w każdym miejscu na ziemi za pośrednictwem Internetu. By skrócić czas oczekiwania na połączenie i ułatwić dostęp do zbioru, proponuje się sześciomiesięczną odpłatną subskrypcję (posiadacz takowej staje się wtedy użytkownikiem serwera COBUILD).

Pięćdziesięcioszesciomilionowy podkorpus można dowolnie przeszukiwać, przy czym warunki konkordancji pozwalają otrzymać listę zawierającą ciąg złożony z jednego lub kilku słów poszukiwanych i odpowiednio określonego (co do liczby słów) kontekstu. Zapytanie można sformułować tak, żeby komputer odnalazł dwa kształty oddzielone kilkoma innymi, dowolnymi oraz formy wyrazowe danego leksemu. „Mały korpus" jest oznakowany – każde słowo tekstowe jest zinterpretowane pod względem części mowy i na pewnym poziomie kodowania (niedostępnym dla przeszukującego), opatrzone odpowiednią etykietą – tagiem[7]. Dzięki temu możliwe jest znalezienie kształtu

[6] Por. *Looking up*. J. M. Sinclair (red.). Collins ELT. London and Glasgow 1987.

[7] Lista podstawowych tagów obejmuje następujące: **NOUN** – etykieta zbiorcza dla wszystkich „podtagów" rzeczownikowych, **VERB** – etykieta zbiorcza dla wszystkich „podtagów" czasownikowych, **NN** – rzeczownik pospolity, **NNS** – rzeczownik w liczbie mnogiej, **JJ** – przymiot-

z dodatkowym parametrem, np. ciągu *mak* będącego formą wyrazową leksemu czasownikowego. Program wyszukujący pozwala na umieszczenie każdorazowo wszystkich parametrów[8]. Możliwe jest także uzyskanie listy k o l o k a c j i, tzn. często powtarzających się połączeń wyrazów, oraz listy frekwencyjnej słów występujących w korpusie (takie zestawienie dla części demonstracyjnej obejmuje 200 najwyżej notowanych pozycji).

Poniżej dokonam przeglądu współcześnie dostępnych korpusów polskich, którym miałam okazję bliżej się przyjrzeć i wykorzystywać w celach badawczych.

Na szczególną uwagę zasługuje najstarszy korpus języka polskiego, stanowiący materiał badawczy **Słownika frekwencyjnego polszczyzny współczesnej** (SFPW), wyd. 1990. Zbiór obejmuje teksty języka polskiego napisane między 1 stycznia 1963 a 31 grudnia 1967 r. Prace od zgromadzenia korpusu do wydania Słownika trwały przeszło dwadzieścia lat. Różne względy natury technicznej (przystosowanie materiału zapisanego na taśmach papierowych do wymagań współczesnych komputerów, korekty merytoryczne tekstów) czy organizacyjnej sprawiły, że zbiór ten istnieje w „wersji roboczej". Do celów badawczych korpus jest udostępniany bezpłatnie (po podpisaniu oświadczenia o niewykorzystywaniu danych do celów komercyjnych).W uzasadnionych wypadkach (np. wysłanie korpusu na dyskietkach lub innym nośniku za granicę) wymagana jest rekompensata poniesionych kosztów.

Korpus zawiera próbki pięciu stylów funkcjonalnych polszczyzny – popularnonaukowego, drobnych wiadomości prasowych, publicystycznego, prozy artystycznej, dramatu – zrozumiałych dla rodzimego użytkownika języka polskiego ze średnim wykształceniem. Kryteria doboru tekstów były precyzyjnie określone dla każdej części zbioru. Na liście tekstów popularnonaukowych znalazły się 602 pozycje książkowe, wśród których nie było specjalistycznych podręczników, monografii, skryptów, jak również encyklopedii i leksykonów. Drobne wiadomości prasowe obejmują 37 tytułów dzienników wydawanych w Polsce. Materiał stanowią „teksty mające postać obiektywnych komunikatów, które występują we wszystkich działach gazet i czasopism: zagranicznym, krajowym, sportowym, kulturalnym itp. Wykluczono natomiast z kanonu źródeł wszelkie teksty zawierające jakiekolwiek elementy komentarza autorów"[9]. Do działu publicystycznego zaliczono artykuły poświęcone problematyce politycznej,

---

nik, **AT** – przedimek określony lub nieokreślony, **RB** – przysłówek, **VB** – forma podstawowa czasownika, **VBN** – forma imiesłowu przeszłego, **VBG** – forma czasownikowa na – *ing*, **VBD** – forma czasu przeszłego.

[8] Korpus dostępny jest pod adresem: http://titania.cobuild.collins.co.uk/.

[9] *Wstęp* do: *Słownik frekwencyjny polszczyzny współczesnej*. I. Kurcz, A. M. Lewicki, J. Sambor, K. Szafran, J. Woronczak, Z. Saloni (red.). Kraków 1990, s. XIV.

społecznej, gospodarczej, kulturalnej, mające na celu wyrażanie i kształtowanie opinii czytelników, a zamieszczone w różnego rodzaju tygodnikach i miesięcznikach. Włączono również stenogramy posiedzeń partyjnych. Proza artystyczna obejmuje książki beletrystyczne, reportaże, eseje, felietony, publikowane na łamach prasy. Wyłączono książki dla dzieci, gatunki prozy osobistej, takie jak pamiętniki, listy oraz utwory stylizowane i poetyckie. Przy opracowaniu listy podstawowej posłużono się *Adnotowanym Rocznikiem Bibliograficznym „Literatura Piękna"*. Na jego podstawie powstał zbiór liczący 1070 tytułów książkowych, z których wybrano łącznie 230 tys. stron druku – pierwszych 25 stron każdej pozycji. Ostatni dział zawiera nie stylizowane dramaty pisane prozą, ukazujące się w „Dialogu" lub jako druki zwarte oraz zapis dwu powieści radiowych: *Matysiakowie* i *W Jezioranach*.

Warunek reprezentatywności korpusu oprócz starannego doboru tekstów realizowany jest poprzez dwustopniowe losowanie cytatów. W każdym dziale po pierwsze losowano tytuł tekstu, z którego następnie losowano próbkę gronową – fragment tekstu ciągłego liczący około pięćdziesięciu słów. W konsekwencji każdy styl reprezentowany jest przez około stutysięczny zbiór słów.

Korpus *Słownika frekwencyjnego* jest oznakowanym przez człowieka korpusem tekstów polskich. Oznacza to, że większość słów tekstowych jest zinterpretowana pod względem gramatycznym i wyposażona w odpowiednią etykietę, zwaną środowiskowo *tagiem*, zdającą z tego sprawę.

Każda etykieta składa się z co najmniej jednego symbolu-liczby. Na pierwszym miejscu znajduje się znacznik części mowy. Autorzy korpusu wyróżnili dziewięć klas leksemów: rzeczowniki, przymiotniki (do klasy tej włączono deklinujące się derywaty odczasownikowe oraz tzw. liczebniki porządkowe), liczebniki, zaimki, czasowniki, przyimki, wykrzykniki, przysłówki, spójniki.

Miejsca drugie i trzecie wypełnia szczegółowa charakterystyka fleksyjna – dla leksemów deklinujących się są to kolejno wartości przypadka i liczby. System oznaczeń form leksemów czasownikowych dzieli formy wyrazowe na osiem typów w zależności od czasu, trybu, syntetyczności/analityczności, występowania z *się*.

Ponadto w korpusie znajdują się oznaczenia nazw własnych, skrótowców. W oddzielne tagi wyposażono formy wyrazowe składające się z więcej niż jednego słowa. Są to skostniałe, sfrazeologizowane wyrażenia przyimkowe (np. *na razie*, *w lewo*, *po polsku*), połączenia *co* i *jak* z przymiotnikami i przysłówkami w stopniu najwyższym (np. *jak najwyżej*, *co najmniej*), imiesłowy przymiotnikowe oraz rzeczowniki odsłowne występujące z *się* (np. *skarżących się*, *zastanowienie się*), nazwiska obce z pisaną osobno cząstką *von*, *de* (np. *von Beethoven*, *de Gaulle*) oraz inne jednostki co najmniej dwuwyrazowe (np. *mimo że*, *jak gdyby*).

Korpus wygląda następująco: każdy styl zapisany jest w oddzielnym pliku tekstowym, poszczególne próbki gronowe są ponumerowane i poindeksowane bibliograficznie; nawiasy zawierają etykiety odnoszące się do poprzedzającego je słowa, np.:

76~Dziennik Ludowy~05.05.1965~str. 1~kol. 4
Bosch[/] oznajmił, że Caamano[/] został[57] wybrany[211] do[62] pełnienia[121] tych[222] funkcji[122] na[66] nadzwyczajnym[261] zebraniu[161] ciała[121] ustawodawczego[221] w[66] San[+] Domingo[/]. Tymczasowy[211] prezydent powstańczy[211] zwrócił[501] się z[65] apelem do[62] prezydenta[121] Francji[/][121] de[+] Gaulle'a[/][121], do[62] rządu brytyjskiego[221] i do[62] wszystkich[222] szefów[122] państw półkuli[121] zachodniej[221] o[64] natychmiastowe[241] uznanie[141] dyplomatyczne[241].

Prace nad *Słownikiem frekwencyjnym* i korpusem, na którym bazuje *Słownik,* były bardzo starannie obmyślane i przeprowadzone. Dzięki tej staranności możliwe było wydanie *Słownika* w dwadzieścia lat po zakończeniu badań materiałowych. Ponadto korzyści płynące z korzystania ze znakowanego korpusu są oczywiste. Możliwe jest odszukanie nie tylko żądanego słowa, ale również odpowiedniej charakterystyki gramatycznej. Opisywany zbiór mimo małych rozmiarów (500 tys. słów, czyli jedna tysięczna Bank of English i jedna dziesiąta korpusu PWN) wydaje się bardzo dobrym materiałem do badań morfologicznych i składniowych – jest idealnie reprezentatywny i zrównoważony. Na niekorzyść korpusu może jedynie przemawiać jego „archaiczność" (teksty z lat 1963–1967). Staje się ona przeszkodą przy analizie materiału pod względem leksykalnym, znacznym przeobrażeniom uległ bowiem styl publicystyczny. System gramatyczny kształtuje się jednak znacznie wolniej i pod tym względem korpus się nie zdezaktualizował.

Kolejnym zbiorem tekstów języka polskiego jest **korpus Redakcji Słowników Języka Polskiego PWN**. Powstaje on od 1995 r. i w kwietniu 2000 r. liczył ponad 50 milionów słów. Jego twórcy zdecydowali, że będzie to korpus ogólny współczesnej polszczyzny, do której zaliczono teksty napisane lub wypowiedziane po 1918 r. Źródła powstałe w ostatnich dziesięciu latach stanowią 50% całości. W około 38% składają się nań literatura piękna, w 33% — naukowa, poradnikowa, pamiętnikarska, 21% stanowią teksty prasowe, 7% — przepisywane z kaset teksty mówione, 1% — pisane teksty ulotne, głównie reklamowe. W porównaniu z innymi korpusami ogólnymi ten zawiera stosunkowo dużo tekstów literackich (twórcy korpusu powołują się na autorytet kulturalny jako kryterium poprawności językowej), nawet poetyckich. Równoważy go jednak spory udział tekstów mówionych.

Korpus stał się podstawą do opracowania nowatorskiego dzieła w polskiej leksykografii – *Innego słownika języka polskiego*, pod redakcją Mirosława Bańki. „To dzięki korpusowi można było [...] wprowadzić definicje kontekstowe, precyzyjniej wyodrębnić różne jednostki opisywane dotychczas w jednej definicji

[...], zróżnicować znaczenia wielu par synonimów opisywanych do tej pory jako w pełni równoznaczne, np. *czekać* i *oczekiwać*"[10] oraz szczegółowo opisać łączliwość jednostek i użyć odpowiednich kwalifikatorów. Dane korpusowe uchroniły redaktorów słownika przed powtarzaniem błędów poprzedników, i tak np. SJP PWN jako narzędnik liczby mnogiej leksemu ***SKROŃ*** podaje formę *skrońmi*. Frekwencja tego słowa w korpusie jest bardzo niska, w części demonstracyjnej nie występuje wcale[11]. Stąd słownik Bańki na pierwszym miejscu podaje formę *skroniami*; *skrońmi* pojawia się opatrzone kwalifikatorem „rzadziej".

Od listopada 2000 r. z korpusu można korzystać za pośrednictwem Internetu. Udostępniona została jego część[12], dokładnie 1 817 058 słów-okazów i, jak zapewniają administratorzy korpusu, 167 674 słów-typów[13]. Zbiór można przeszukiwać ze względu na zadany kształt, otrzymując listę konkordancji zawierającą maksymalnie dwieście przykładów, posortowaną według poszukiwanego kształtu lub jego prawo- czy lewostronnego sąsiedztwa. Program pozwala na szukanie sekwencji dwu ciągów znaków. Otrzymane próbki są wstępnie oznakowane – przy każdej podany jest adres bibliograficzny, zawierający nazwisko autora, tytuł utworu, miejsce i rok wydania dla cytatów pochodzących z książek albo tytuł, rocznik i numer czasopisma dla przykładów prasowych.

Niewątpliwą zaletą korpusu jest jego łatwa dostępność. Korzystanie z części demonstracyjnej umieszczonej w Internecie nie wymaga specjalnych opłat z tytułu użytkowania (zainteresowany ponosi standardowe koszty połączenia teleinformatycznego). Można przypuszczać, że do celów naukowych można korzystać z całości zbioru w siedzibie wydawnictwa – w Warszawie przy ul. Miodowej.

Od niedawna (niestety twórcy nie podają dokładnej daty) powstaje w Polsce kolejny korpus. Jego właścicielem jest **Instytut Podstaw Informatyki Polskiej Akademii Nauk**. Zbiór liczy 13,4 miliona słów. Składają się nań głównie teksty prasowe oraz Stary i Nowy Testament, oprócz tego proza współczesna, rozmowy telefoniczne (materiały z książki M. Pisarkowej *Składnia rozmów telefonicznych*), wybrane utwory Konopnickiej, Sienkiewicza i Mickiewicza. Nie jest to jeszcze zbiór zrównoważony i reprezentatywny. Niedociągnięciem korpusu jest brak polskich znaków diakrytycznych lub innych znaków kodujących polskie litery. Tak więc ciąg „pisze" może jednocześnie być dekodowany jako „pisze" oraz „piszę". Należy jednak przypuszczać, że jest to tylko problem udostępnienia tekstów, który w niedalekiej przyszłości zostanie rozwiązany. Korpus

[10] M. Łaziński: *Korpus PWN*. ISJP. Warszawa 2000, s. LVII–LVIII.

[11] W części demonstracyjnej *skroniami* występuje tylko raz.

[12] Według danych dostępnych pod adresem: http://www.slowniki.vm.pl/korpus/.

[13] „Słowa-okazy" i „słowa-typy" rozróżniam za Z. Salonim: *Unilateralne i bilateralne podejście do znaków języka (naturalnego)*. [W:] *W świecie znaków*. Warszawa 1996, s. 287–294.

w całości jest dostępny w Internecie[14] i w celach niekomercyjnych może korzystać z niego każdy, z zastrzeżeniem, że w swojej pracy poda źródło danych. Materiał nie jest znakowany gramatycznie ani wstępnie ustrukturyzowany – oddzielnie jest przeszukiwany każdy plik tekstowy. Program sortujący korzysta ze składni wyrażeń regularnych, można więc każdorazowo odnaleźć więcej niż jeden kształt graficzny.

Zbiór ten na razie ma znaczenie czysto informacyjne, ze względu na przyjęty system kodowania („zlepiający" litery) nie może być wykorzystywany do badań nad współczesną polszczyzną ogólną. Z pewnością przez najbliższe trzy lata zestaw ten zostanie udostępniony w postaci poprawnej i będzie źródłem danych dla językoznawców. IPI PAN realizuje bowiem projekt badawczy finansowany przez Komitet Badań Naukowych.

Korpusy albo zalążki korpusów powstają równolegle w kilku ośrodkach naukowych w Polsce. W **Instytucie Języka Polskiego Polskiej Akademii Nauk w Krakowie** realizuje się projekt komputerowego korpusu współczesnych tekstów polskich[15]. W **Instytucie Filologii Polskiej Uniwersytetu Gdańskiego** powstaje wielka biblioteka internetowa literatury polskiej (projekt badawczo-naukowy *Literatura polska w internecie*), gromadząca zarówno najstarsze teksty literackie, jak i współczesne. Z kolei w **Instytucie Anglistyki Uniwersytetu Łódzkiego** zgromadzono największy zestaw tekstów polskich, na który składają się w większości teksty prasowe.

Dużymi korpusami, rzędu kilku milionów słów, dysponują osoby prywatne. Zbiory gromadzone na użytek własnych badań językowych są nierzadko udostępniane również innym badaczom[16].

Wyszukiwanie w kilkudziesięciumilionowych nie oznakowanych wcześniej zbiorach tekstowych wyrafinowanych informacji możliwe jest dzięki odpowiednim programom komputerowym. Obecnie prowadzi się intensywne prace nad takimi narzędziami informatycznymi. Dla przykładu podam dwa z nich, które miałam okazję bliżej poznać m.in. podczas zajęć w Zakładzie Lingwistyki Komputerowej w Uniwersytecie Warszawskim, prowadzonych przez prof. Marka Świdzińskiego.

Pierwszym etapem interpretacji gramatycznej języków fleksyjnych jest a n a l i z a m o r f o l o g i c z n a. Polega ona na tym, że słowo, jako jednostka

[14] Zbiór jest dostępny pod adresem: http://ling.ohio-state.edu/adamp/searchpage.

[15] Por. K. Węgrzynek: *Projekt komputerowego korpusu współczesnych teksów polskich*, „Język Polski" 1995, LXXV, z. 4–5, s. 332–341.

[16] Np. Robert Wołosz na podstawie swojego korpusu, liczącego ok. 80 milionów słów, dostarczył danych materiałowych Elżbiecie Awramiuk. Por. E. Awramiuk: *Systemowość polskiej homonimii międzyparadygmatycznej*. Białystok 1999, s. 9.

unilateralna, zostaje zinterpretowane pod względem morfologicznym, stając się bilateralną jednostką tekstu. Takim programem do analizy automatycznej tekstów języka polskiego jest POMOR, którego część lingwistyczną opracował Robert Wołosz. Analizator rozpoznaje i podaje kompletną charakterystykę gramatyczną ponad 140 tys. leksemów języka polskiego. Budowa słownika jest dwudzielna, składa się bowiem ze zbioru tematów (niezmiennych części w ramach paradygmatów) oraz ze zbioru zakończeń (zmiennych elementów w budowie paradygmatów). Dzięki odpowiedniemu kodowaniu elementów obu części analizator nie dopuszcza do łączenia ze sobą przypadkowych tematów i zakończeń[17].

Drugim etapem jest a n a l i z a s k ł a d n i o w a . Jej celem jest zinterpretowanie zdania rozumianego ortograficznie jako pewnej konstrukcji składniowej i reprezentowanie jej w sposób graficzny, np. w postaci drzewa. Przykładem takiego programu jest parser AS, autorstwa Marcina Wolińskiego, będący implementacją *Gramatyki formalnej języka polskiego* Marka Świdzińskiego[18]. Zbudowany jest ze słownika gramatycznego, zawierającego wymagania konotacyjne poszczególnych jednostek, oraz stosownych reguł, na mocy których ciąg znaków rozpoczynający się wielką literą i kończący kropką jest interpretowany jako zdanie języka polskiego. Punktem wyjścia analizy są informacje ze słownika na temat wymagań składniowych poszczególnych leksemów. I tak np., jeśli w słowniku przy czasowniku ZNAĆ zapisano wymaganie frazy rzeczownikowej w bierniku, to ciąg „tego" w zdaniu „Znam tego." zostanie zinterpretowany jako biernik liczby pojedynczej leksemu TEN. Wynik uzyskany w ten sposób określa się jako *sterowany przez dane*[19].

Wspomniane analizatory są programami opracowanymi na podstawie opisu lingwistycznego języka polskiego. Dlatego dla jednego słowa tekstowego podają w s z y s t k i e alternatywne interpretacje gramatyczne, np. analizator morfologiczny kształt *zajęczy* zidentyfikuje jako formę trzeciej osoby liczby pojedynczej czasu teraźniejszego czasownika ZAJĘCZEĆ oraz formę mianownika lub biernika liczby pojedynczej rodzaju męskiego przymiotnika ZAJĘCZY.

Na zupełnie innej zasadzie konstruuje się tagery – programy przewidujące charakterystykę gramatyczną słowa, dające j e d n ą „słuszną" odpowiedź dla każdego ciągu znaków między dwoma separatorami. Rozpatrując rzecz etymologicznie, tager to coś, co przypisuje tagi, tzn. etykietki, znaczniki. Jego wyniki mogą być przybliżone lub niekompletne. Analizator sugeruje wnikliwe przetwarzanie.

[17] Por. R. Wołosz: *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*. Warszawa 2000. (Nie publikowana praca doktorska).

[18] M. Świdziński: *Gramatyka formalna języka polskiego*. Warszawa 1992.

[19] Por. A. Wachowski: *Adekwatność lingwistyczna analizatorów składniowych języka polskiego*. Warszawa 2000. (Nie publikowana praca magisterska. Wydział Matematyki, Informatyki i Mechaniki UW).

Oczywiście im więcej możliwości interpretacyjnych, tym poważniejszy staje się problem ujednoznacznienia wyników (dyzambiguacji).

Tagery znalazły już zastosowanie w badaniach języków o okrojonej fleksji, np. tagerem statystycznym oznakowano brytyjski *Bank of English*. Podejmuje się próby budowania tagerów dla języków fleksyjnych, np. dla czeskiego.

Korpusy odgrywają znaczącą rolę w opracowywaniu narzędzi informatycznych do automatycznego przetwarzania teksów języków naturalnych. Zarówno analizatory, jak i tagery budowano i testowano na zbiorach tekstowych.

Szybki rozwój nauk technicznych, powszechna dostępność szybkich komputerów o coraz większej mocy obliczeniowej z pewnością sprzyja archiwizowaniu obszernych danych językowych w postaci elektronicznej i automatycznemu ich przetwarzaniu. Należy przypuszczać, że powstająca właśnie lingwistyka komputerowa zyska na znaczeniu i stanie się niezależną dyscypliną naukową.

Prawdopodobnie istniejące polskie korpusy będą się rozrastały. Doświadczenia innych krajów pokazują jednak, że równorzędne zestawy nie rozwijają się równomiernie. Z czasem zyskuje na znaczeniu jeden z nich, pretendując w ten sposób do miana „dominującego" zbioru tekstów. Najbardziej zaawansowany w pracach korpusowych w Polsce jest ośrodek warszawski, tu zatem są największe szanse na powstanie polskiego korpusu narodowego.

Summary

The aim of the article it to introduce the Polish reader to the problems of corpora of natural languages and to present three sets of Polish texts: the *Słownik frekwencyjny polszczyzny współczesnej* (*Dictionary of Word Appearance Frequency in Contemporary Polish*), the edition of dictionaries of the Polish language of the PWN (Polish Scientific Publishers) and the Institute of Rudiments of Computer Science of the Polish Academy of Sciences. In the general part, also a British collection, the *Bank of English*, is characterised. The author gives examples of how to use text corpora in such researches of the linguistic system as: dictionary compilation, text tagging and automatic morphological and syntactic analysis. The work includes a description of how to use particular sets and their contents. In the conclusion, the author predicts in which direction Polish corpora will be developing.