Janusz S. Bień*, Marcin Woliński**

Wzbogacony korpus *Słownika frekwencyjnego polszczyzny współczesnej*

17 grudnia 2001 roku

Korpus słownika frekwencyjnego to pięć zestawów próbek po 100 000 słów wylosowanych z autentycznych tekstów z lat 1963–1967 należących do 5 stylów — tekstów popularnonaukowych, drobnych wiadomości prasowych, publicystyki, prozy artystycznej i dramatu artystycznego — na potrzeby badań frekwencji słów języka polskiego.

Pierwotnie korpus miał formę taśm papierowych wyperforowanych na dalekopisie (czego konsekwencją był brak rozróżnienia małych i dużych liter). Został on wczytany do komputera przez Bronisława Rocławskiego (wówczas na Uniwersytecie Gdańskim) i zapisany na taśmie magnetycznej; niestety, w trakcie tej operacji do korpusu wkradły się pewne przekłamania. Taśma magnetyczna została zapisana na komputerze ODRA 1204 w standardzie, który szybko wyszedł z użycia. W związku z tym taśma z korpusem trafiła w ręce Krzysztofa Szafrana, który w Instytucie Informatyki Uniwersytetu Warszawskiego na podstawie list frekwencyjnych dla poszczególnych stylów ([4, 5, 9, 6, 7]) opracowywał tzw. tom zbiorczy, opublikowany jako *Słownik frekwencyjny polszczyzny współczesnej* [8] (pewne dodatkowe informacje o historii tego projektu zawiera artykuł [13]). Krzysztof Szafran za pomocą specjalnie przygotowanego programu odczytał taśmę na komputerze SM-4 i zapisał jej zawartość na bardziej nowoczesnych nośnikach, w wyniku czego korpus stał się dostępny również na dyskietkach stosowanych w komputerach osobistych.

Choć słownictwo korpusu jest już częściowo przestarzałe, korpus ten nadal ma dużą wartość m.in. dla badań składniowych. Z tego względu Janusz S. Bień — jeszcze jako adiunkt w Instytucie Informatyki UW — wystąpił z inicjatywą dokonania korekty korpusu i udostępnienia go w bardziej nowoczesnej formie. Pierwszy krok w tym kierunku stanowiła opracowana pod jego kierunkiem praca magisterska Marty Nazarczuk ([10]). Janusz S. Bień wykonał również eksperyment polegający na przetworzeniu stylu popularnonaukowego dwoma korektorami ortograficznymi: polskiej firmy TiP i węgierskiej firmy Morphologic ([12, s. 153], [15]); współautor tego drugiego narzędzia, Robert Wołosz, przetworzył nim również inne style i udostępnił nam wyniki. Operacje te pozwoliły nie tylko wykryć błędy literowe, ale odtworzyć z dużym prawdopodobieństwem rozróżnienie dużych i małych liter.

Janusz S. Bień kierował również następnymi etapami prac nad tym zadaniem, w czym istotnie pomagał mu ostatnio Marcin Woliński. Pewne prace nad korpusem zostały wykonane w ramach projektów ELAN (*European Language*

* Zakład Zastosowań Informatycznych, Instytut Orientalistyczny Uniwersytetu Warszawskiego, patrz `http://www.orient.uw.edu.pl/~zzi/`

** Zespół Inżynierii Lingwistycznej, Instytut Podstaw Informatyki PAN, patrz `http://www.ipipan.waw.pl`

Activities Network) i projektu KBN *Zestaw testów do weryfikacji i oceny analizatorów języka polskiego* oraz w ramach pracy magisterskiej Macieja Ogrodniczuka [11]; w szczególności Katarzyna Głowińska przygotowała nową taksonomię morfologiczną inspirowaną podobnymi pracami dla języka czeskiego [1]. W latach 1999-2000 ręczna weryfikacja korpusu, utworzenie hasłowanej konkordancji i pewne inne prace z tym związane były finansowane z inicjatywy prof. dr hab. Jadwigi Sambor — kierownik Katedry Językoznawstwa Ogólnego i Bałtystyki na Wydziale Polonistyki UW — z funduszy badań statutowych Katedry.

Obecna postać korpusu różni się od pierwotnej przede wszystkim pod następującymi względami:

1. Jak już było wspomniane, wprowadzono rozróżnienie dużych i małych liter niewystępujące w oryginalnych plikach korpusu z powodów technicznych (z powodu braku polskich liter w repertuarze znaków dalekopisu odpowiednimi dużymi literami oznaczano właśnie polskie litery).
2. Skonfrontowano wersję elektroniczną korpusu z przechowywanymi przez prof. Sambor oryginalnymi fiszkami weryfikując poprawki wprowadzone komputerowo i wprowadzając pewne uzupełnienia. Dla stylu popularnonaukowego zadanie to we wzorcowy sposób wykonała Marta Nazarczuk, niestety weryfikatorzy pozostałych stylów okazali się mniej solidni.
3. Wprowadzono do komputera pełne opisy bibliograficzne źródeł i dołączono je do odpowiednich fiszek. Pracę tę wykonała Marta Nazarczuk.
4. Przygotowano elektroniczną wersję fiszek wygodną do przeglądania; pracę tę wykonał Marcin Woliński.
5. Dokonano weryfikacji kodów gramatycznych przy poszczególnych słowach oraz wprowadzono dodatkową informację gramatyczną, zgodną z wspomnianą wcześniej taksonomią. Informację tę dopisano również do tych słów, które w oryginalnej wersji w ogóle nie miały kodów gramatycznych. Pracę tę wykonał Maciej Ogrodniczuk — według wskazówek Marcina Wolińskiego — wykorzystując analizator morfologiczny SAM autorstwa Krzysztofa Szafrana [14].
6. Wprowadzenie kodów gramatycznych dla wszystkich słów, dla których było możliwe zrobić to automatycznie, pozwoliło na przygotowanie *hasłowanej konkordancji tekstu korpusu*. Pracę tę wykonał Bartłomiej Krawczyk za pomocą opracowanego przez siebie programu ([2], [3]); niestety, ze względu na eksperymentalny charakter tego programu może on być użytkowany tylko przez jego autora. Hasłowane konkordancje w nieco innym formacie zostały również przygotowane przez Marcina Wolińskiego. Oto przykład:

```
----- ---------------------------------- pies ---------------------------------------------
D0325     a dzieci pluskały się w wodzie. Pies    SSNA---------P był nieodłącznym towarzyszem
E0452       choć chodziłam za nim jak ten pies.   SSNA---------P Tyś przy nim tak nie warowała,
E0885       do nogi! Bobik, leżeć! Zdechł pies!   SSNA---------P
E1221         żeby bez plakatów, bo i tak pies    SSNA---------P z kulawą nogą nie przyjdzie.
E1245                         Co to? .. Pies ..   SSNA---------P Pies? Jaki pies? Gryzie? ..
E1245                  Co to? .. Pies .. Pies?    SSNA---------P Jaki pies? Gryzie? .. Czasami
E1245        Co to? .. Pies .. Pies? Jaki pies?   SSNA---------P Gryzie? .. Czasami .. On się
E1245         .. On się ciebie słucha, ten pies?  SSNA---------P To każ mu wyjść spod ławki
C0790    Jeśli gwizdniemy na wyszkolonego psa     SSAA---------P specjalnym gwizdkiem
C1207       psu dodatkowe serce (z innego psa)    SSGA---------P tak, że ten żył z dwoma
D0881  w szarówce ledwie ustępującej nocy psa     SSAA---------P Kolędy biegnącego ze skowytem
D1207          a kochałem tylko Grażynę i psa.    SSAA---------P Układaliśmy stóg. Najpierw na
E1017     fotografowali. Domy, cegielnię, psa     SSAA---------P przy budzie, uschnięte drzewo
C1207        skupił on uwagę i wysiłki na psach,  SPLA---------P przeszczepiając im płuca,
D1005         Pachołka, córce zaszczutego psami   SPIA---------P Pachołka? Że staraj się, aby
D1333              Poszliśmy we trzech, z psem,   SSIA---------P na nadbrzeże, przy którym
E0996               Kondycja fizyczna pod psem.   SSIA---------P Dwadzieścia lat klęczałem w
B1285      oraz schronisko dla bezdomnych psów,   SPGA---------P kotów i tym podobne.
D1221      Leosi. Już z daleka szczekanie psów    SPGA---------P przywodziło na myśl rozległe
E1251      Nie. Po prostu miałem już tyle psów,   SPGA---------P że naprawdę znam się z
E1251 jak w leśniczówce, zawsze było masę psów.   SPGA---------P Nawet w czasie wojny trzymałem
B0835           zwierzę. Tylko Ewa ruszyła psu    SSDA---------P z pomocą. Wzięła go na ręce
C1207            doktór Demichow wszczepił psu    SSDA---------P dodatkowe serce (z innego psa)
D0492  guzik. Innymi słowy, panie, to dwa psy,    SPNA---------P które usiłują jeden drugiemu
D0640    Biały promień szastnął w krzaki, psy     SPNA---------P z ujadaniem skoczyły ku
D1461      w zapadłej tylko co ciemności. Psy     SPNA---------P nagonki, które od pewnej
E1149             od niego .. To cholerne psy.    SPNA---------P
```

Tak rozbudowany korpus proponujemy nazywać wzbogaconym korpusem *słownika frekwencyjnego* (w skrócie WKSF). Jego podstawową formą jest płyta CD-ROM, której zawartość jest opisana odrębnie.

Prace nad udoskonaleniem i wzbogaceniem korpusu nie mają wyraźnego końca — kolejne eksperymenty komputerowe mogą ujawnić przeoczenia korekty, nowe potrzeby badawcze mogą wymagać wzbogacenia korpusu o dodatkowe informacje. Dlatego najwłaściwsze wydaje się publiczne udostępnienie korpusu — oraz materiałów pochodnych, takich jak konkordancje — na podstawie licencji GNU: oznacza to pełną swobodę ich wykorzystania (nawet do celów komercyjnych!), ale pod warunkiem udostępnienia na tych samych zasadach wszystkich poprawek i modyfikacji (patrz `http://www.gnu.org/copyleft/`, polskie tłumaczenie dostępne jest m.in. pod adresem `http://gnu.org.pl/text/licencja-gnu.html`). Licencja GNU jest stosowana powszechnie w informatyce od wielu lat i można śmiało stwierdzić, że jest to rozwiązanie sprawdzone w praktyce.

W razie uzyskania pisemnej zgody na jego udostępnienia na takich lub podobnych zasadach od właścicieli praw autorskich korpusu, podejmiemy kroki, aby korpus stał się również dostępny w Internecie na witrynach Instytutu Informatyki Uniwersytetu Warszawskiego i Instytutu Podstaw Informatyki PAN.

Literatura

[1] Jan Hajič and Barbora Hladka. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of ACL/Coling'98*, Montreal, Canada, Aug. 5-9, pp. 483-490. `http://shadow.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Tagging/Doc/References/col98.pdf`.

[2] Krawczyk, Bartłomiej.Indeksowanie tekstu polskiego z wykorzystaniem analizy morfologicznej. Praca magisterska napisana pod kierunkiem dra Krzysztofa Szafrana. Warszawa, 1999. Instytut Informatyki Uniwersytetu Warszawskiego. 83 s., płyta CD.

[3] Krawczyk, Bartłomiej. Konkordancje hasłowane tekstów języka polskiego. Gerhild Zybatow, Uwe Junghanns, Grit Mehlborn, Luka Szucsich, *3. Europäische Konferenz "Formale Beschreibung slavischer Sprachen, Leipzig 1999*. Linguistische Arbeits-Berichte 75, Institut für Linguistik, Universität Leipzig 2000, s. 191–199.

[4] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom I. Teksty popularnonaukowe.* Warszawa, 1974. Uniwersytet Warszawski.

[5] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom II. Drobne wiadomości prasowe.* Warszawa, 1974. Uniwersytet Warszawski.

[6] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom IV. Proza artystyczna.* Warszawa, 1976. Uniwersytet Warszawski.

[7] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom V. Dramat artystyczny.* Warszawa, 1977. Uniwersytet Warszawski.

[8] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Szafran, Krzysztof; Woronczak, Jerzy. *Słownik frekwencyjny polszczyzny współczesnej*, Kraków, 1990. Instytut Języka Polskiego PAN.

[9] Lewicki, Andrzej; Masłowski, Władysław; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom III. Publicystyka.* Warszawa, 1975. Uniwersytet Warszawski.

[10] Nazarczuk, Marta. *Wstępne przygotowanie korpusu „Słownika frekwencyjnego polszczyzny współczesnej" do dystrybucji na CD-ROM*. Praca magisterska napisana

pod kierunkiem dra hab. Janusza S. Bienia. Warszawa, 1997. Instytut Języka Polskiego Uniwersytetu Warszawskiego. 59 s., płyta CD.

[11] Ogrodniczuk, Maciej. *Wykorzystanie SGML i TEI do zapisu polskich danych lingwistycznych*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa, 2000. Instytut Informatyki Uniwersytetu Warszawskiego. 83 s., płyta CD.

[12] Prószéky, Gábor. Humor (High-speed Unification Morphology). A Morphological System for Corpus Analysis. Heile Rettig (ed.). Proceedings of the First European Seminar Language Resources for Language Technology. Tihany, Hungary, September 15 and 16, 1995, pp 149–158.

[13] Saloni, Zygmunt. *Słownik frekwencyjny polszczyzny współczesnej*. *ComputerWorld* 4 listopada 1991, s. 16-17.

[14] Szafran, Krzysztof. *Analizator morfologiczny SAM-95 — opis użytkowy*. Maj 1996. Instytut Informatyki Uniwersytetu Warszawskiego. `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/`.

[15] Robert Wołosz. Efektywna metoda analizy i syntezy morfologicznej w języku polskim. Praca doktorska (promotor prof. dr hab. Zygmunt Saloni). Wydział Polonistyki, Uniwersytet Warszawski, Warszawa 2000.