Nowa edycja wzbogaconego korpusu słownika frekwencyjnego

Maciej Ogrodniczuk

4 kwietnia 2003

1 Wstęp

Dane korpusu frekwencyjnego zostały zebrane w latach 1963–1967 na potrzeby badań nad częstością występowania wyrazów w języku polskim. Składa się na nie 10 000 próbek po około 50 słów każda, czyli ogółem ok. 500 000 słów z tekstów współczesnego języka pisanego, zgromadzonych w pięciu transzach odpowiadających najważniejszym stylom polszczyzny pisanej. Wyniki dla poszczególnych stylów zostały w latach 1974-77 udostępnione w postaci list frekwencyjnych [13], natomiast w roku 1990 ukazał się słownik frekwencyjny polszczyzny współczesnej w postaci tomu zbiorczego [14].

W roku 2001 autorzy wyrazili zgodę na udostępnienie korpusu na zasadach licencji GNU (opisanej szerzej w dalszej części artykułu) i w konsekwencji było możliwe umieszczenie go w Internecie. Z inicjatywy Janusza S. Bienia powstała płyta CD pod nazwą *Wzbogacony Korpus Słownika Frekwencyjnego* [2], zawierająca oprócz danych korpusowych w wersji podstawowej i rozszerzonej (wzbogaconej) także związane z nimi dodatkowe materiały i dokumentację korpusu.

W chwili obecnej[1] przygotowywane jest drugie wydanie tej płyty, poszerzone m. in. o dodatkowe narzędzia wspomagające przeglądanie tekstów korpusowych. Informacja o sposobie pobrania aktualnej wersji obrazu płyty z Internetu została omówiona w dodatku.

2 Wersje korpusu

Korpus słownika frekwencyjnego został zamieszczony na płycie w dwóch wersjach. Pierwsza, nazywana *wersją podstawową* i zapisana w katalogu `Korpus`

[1] Kwiecień 2003 r.

`surowy`, zawiera oryginalne, dopisane ręcznie trzycyfrowe[2] kody gramatyczne wraz z dodatkowymi oznaczeniami wyróżniającymi np. formy złożone i nazwy własne (ich dokładny opis jest dostępny m. in. na płycie [10]). Oto fragment próbki w wersji podstawowej:

```
37~Dziennik Bałtycki~02.10.1965~str. 1~kol. 7
W[66] imieniu dwu[32] tysięcy[122] studentów[122] Uniwersytetu
Warszawskiego[221] zgromadzonych[222] w[66] auli[161]
Auditorium[+]Maximum[&], rektor uczelni[121], profesor Stanisław[/]
Turski[/] powitał wicepremiera[141] Zenona[/][141] Nowaka[/][141].
```

Wersja wzbogacona, zapisana na płycie w katalogu `Korpus wzbogacony`, powstała poprzez automatyczne uzupełnienie i rozbudowanie informacji morfologicznej zgodnie z przyjętym rozszerzonym zestawem oznaczeń (opisuje je szczegółowo zamieszczony na płycie artykuł [12], sam zaś przebieg prac przedstawiony jest w artykule [17]). Ze względów technicznych w wersji wzbogaconej każda forma (lub zestaw form analizowanych jako całość) została zapisana w osobnej linii pliku wraz z postacią hasłową, kodem oryginalnym i rozszerzonym. Oto fragment cytowanej wyżej próbki w wersji wzbogaconej:

```
[...]
w[w,66,P-L----------P]
auli[aula,161,SSLF---------P]
Auditorium Maximum[auditorium maximum,,X------------P],
rektor[rektor,,SSNP---------P]
uczelni[uczelnia,121,SSGF---------P],
profesor[profesor,,SSNP---------P]
Stanisław[stanisław,,SSNP---------W]
Turski[turski,,X------------W]
powitał[powitać,,VS-M-3PON----P]
wicepremiera[wicepremier,141,SSAP---------P]
[...]
```

3 Fiszki i konkordancje

Oprócz opisanych powyżej wersji korpusu na płycie w katalogu `Fiszki` zamieszczono dane wersji podstawowej w formacie naśladującym oryginalne

[2]Jak wspominają autorzy korpusu w oryginalnej instrukcji redakcyjnej (patrz np. [8], str. 177), na postać kodów wpłynęły ograniczenia komputera, na którym dane były oryginalnie przetwarzane. Zainteresowanych historią obróbki zbiorów korpusowych odsyłam do literatury cytowanej w zamieszczonym na płycie artykule wprowadzającym [9], np. [18].

fiszki stosowane przez redaktorów słownika w postaci plików PDF (patrz A.2). Pliki te występują w dwóch wariantach: bez kodów gramatycznych oraz z kodami zapisanymi w tekście, bezpośrednio przy formach, do których się odnoszą. W pierwszej edycji płyty w obu postaciach ze względów technicznych pominięte zostały dodatkowe oznaczenia klasyfikacyjne (np. w przypadku form wielowyrazowych opisywanych pojedynczym kodem gramatycznym znajduje się on po ostatnim członie formy i mimo że odnosi się do formy jako całości, nie jest w żaden sposób powiązany z poprzedzającymi członami). W przygotowywanej edycji planuje się uwzględnienie oznaczeń łączących formy wieloczłonowe.

Jako uzupełnienie danych przygotowana została również hasłowana konkordancja wersji wzbogaconej korpusu. Katalog `Konkordancje` zawiera jej dwie postacie — wersję bez lokalizacji form analizowanych (czyli bez identyfikacji stylu i numeru próbki), podzieloną na przeszło 250 plików zawierających poszczególne części zbioru analiz pogrupowanych alfabetycznie oraz konkordancję z lokalizacjami, zawartą w pojedynczym pliku (o rozmiarze przekraczającym 60 MB). W konkordancjach uwzględniono wszystkie wystąpienia form, zarówno te, którym udało się dopisać jednoznaczne analizy w postaci rozszerzonej automatycznie, jak i te, dla których operacja ta nie była możliwa.

4 Inne informacje na płycie

Katalog `Dokumentacja` zawiera teksty w postaci plików PDF opisujące stosowane w oryginalnej wersji korpusu numeryczne kody gramatyczne, techniczny opis prac nad wzbogacaniem tekstów korpusowych w opisy morfologiczne zgodne z nowym, rozszerzonym zestawem oznaczeń oraz sam indeks nowych oznaczeń morfologicznych. Płyta zawiera ponadto katalog `Varia` z tekstami źródłowymi zamieszczonych artykułów, dotyczącymi korpusu pracami magisterskimi Marty Nazarczuk [15] i moją [16], dokumentem umożliwiającym wydruk okładki do płyty i pewnymi dodatkowymi narzędziami.

5 Edytor Emacs jako narzędzie użytkownika płyty

Ważnym elementem nowej edycji płyty będzie gotowa do uruchomienia, zaadaptowana z płyty CD [1] wersja dostępnego również na licencji GNU darmowego edytora tekstów Emacs. To zaawansowane narzędzie edycyjne może

być przez użytkownika wykorzystane co najmniej do wygodnego przeglądania i przeszukiwania plików korpusowych.

5.1 Wyszukiwanie przyrostowe

Jedną z zalet edytora Emacs jest tryb tzw. wyszukiwania przyrostowego, W którym Emacs rozpoczyna przeszukiwanie już po wpisaniu pierwszych znaków szukanego tekstu. Wyszukiwanie przyrostowe w treści otwartego pliku dostępne jest po naciśnięciu kombinacji klawiszy[3] `CTRL+S` (od ang. *search* – przeszukiwanie) lub `CTRL+R` (ang. *reverse search* – przeszukiwanie w tył), a następnie wpisaniu szukanego hasła. Ponowne naciśnięcie `CTRL+S` po znalezieniu ciągu powoduje przejście do jego następnego wystąpienia. W przypadku jego braku Emacs wyświetli komunikat `Failing search` (niepowodzenie wyszukiwania).

Więcej informacji o trybie przyrostowym można znaleźć m. in. w obszernej dokumentacji (w jęz. angielskim) dostępnej razem z edytorem.

5.2 Wyszukiwanie z wykorzystaniem wyrażeń regularnych

Wyrażenia regularne to wzorce wspomagające wyszukiwanie ciągów znaków przy użyciu ustalonych symboli, np. reprezentujących dowolną literę, cyfrę lub inną grupę znaków. Używając wyrażeń regularnych możemy przeszukiwać korpus na podstawie złożonych kryteriów, znacznie rozszerzających możliwości wyszukiwania „dosłownego".

Oto kilka przykładów zastosowania tego formalizmu (podane wyrażenia nie są jedynymi właściwymi dla osiągnięcia żądanych wyników, tego rodzaju kwerendy możemy zazwyczaj formułować na wiele sposobów):

`\<droga\[2`

> Polecenie wyszukania wszystkich przymiotnikowych wystąpień formy *droga* wraz z odpowiadającym im kodami.
>
> Symbol `\<` oznacza początek słowa, co pozwala na uniknięcie wystąpień w rodzaju *niedroga*; znak `\` sygnalizuje dosłowne użycie znaku następnego – tu: nawiasu kwadratowego, a `2` jest korpusowym oznaczeniem przymiotnika (tworzenie podobnych wzorców wymaga zapoznania się z zestawem używanych oznaczeń morfologicznych, np. na podstawie artykułu [10]).

[3] `CTRL+znak` i `CTRL+ALT+znak` uzyskujemy poprzez przytrzymywanie klawisza `CTRL` lub jednocześnie obu `CTRL` i `ALT` podczas wciskania klawisza `znak`.

`\<miejsc\w*\[1`

Polecenie wyszukania wszystkich form gramatycznych słowa *miejsce*. Symbol `\w` oznacza dowolną literę wchodzącą w skład słowa, znak gwiazdki – powtórzenie poprzedzającej gwiazdkę litery dowolną liczbę razy (bądź brak tej litery, czyli „powtórzenie 0 razy"), zaś `1` jest oznaczeniem rzeczownika.

Warto zwrócić uwagę, że użycie powyższego wyrażenia spowoduje wyszukanie także form rzeczownika *miejscowość* (o ile wystąpiły w tekście) — aby tego uniknąć, należy doprecyzować wyrażenie używając symbolu alternatywy `\|`, tworząc wyrażenie w rodzaju `\<miejsce\|\<miejsca`.

`\w+\[1\([0-9]\)\([0-9]\)\]\W+\w+\[2\1\2`

Polecenie wyszukania przymiotników występujących po rzeczownikach i zgadzających się z poprzedzającym rzeczownikiem w liczbie i rodzaju. Znak plusa oznacza powtórzenie poprzedzającej litery 1 lub więcej razy, wyrażenie `[`*lista znaków*`]` oznacza wystąpienie jednego ze znaków z listy (tu: [0-9] to oznaczenie dowolnej cyfry), ciąg `\(`*wyrażenie*`\)` umożliwia zapamiętanie wyrażenia do późniejszego użycia, znak `\W` zastępuje znak nie wchodzący w skład słowa, zaś oznaczenia `\1` i `\2` umożliwiają przywołanie odpowiednio pierwszego i drugiego zapamiętanego wyrażenia, w celu zapewnienia zgodności przymiotnika w liczbie (druga pozycja w nawiasie kwadratowym) i rodzaju (trzecia pozycja) z zapamiętanymi liczbą i rodzajem występującego bezpośrednio wcześniej rzeczownika.

Emacs przełącza się w tryb wyszukiwania wyrażeń regularnych po naciśnięciu kombinacji `CTRL+ALT+S`, po której to kombinacji należy podać szukany ciąg znaków.

Informacje w języku polskim dotyczące pełnych możliwości tego formalizmu można znaleźć w [11].

5.3 Tryb *occur*

Oprócz możliwości przeglądania wyników wyszukiwania wyrażeń regularnych bezpośrednio w tekście, może okazać się użyteczne sporządzenie wykazu zawierającego wyłącznie te linie pliku korpusowego, które zawierają poszukiwane wyrażenie. Ten specjalny tryb jest szczególnie użyteczny w przypadku wersji wzbogaconej korpusu, w której każda forma z analizą zapisana jest w osobnym wierszu. Wywołujemy go wciskając `ALT+X`, następnie wpisując słowo `occur`, naciskając `ENTER` i podając szukany wzorzec. Lista linii z jego wyróżnionymi wystąpieniami zostanie wyświetlona w osobnym buforze.

Oto fragment wyniku wyszukiwania dwuczłonowych przyimkowych wyrażeń przysłówkowych (specyfikowanych wyrażeniem `\[\w+ \w+,.*,D`):

```
 4954:od razu[od razu,,D----------P]
 5234:na raz[na raz,,D----------P]
 7022:od dawna[od dawna,,D----------P]
11556:na pewno[na pewno,,D----------P]
18428:raz po raz[raz po raz,,D----------P]
21471:jak najdalej[jak najdalej,,D----------P]
23264:na razie[na razie,,D----------P]
37343:co dzień[co dzień,,D----------P]
39067:na ogół[na ogół,,D----------P]
42813:na krótko[na krótko,,D----------P]
[...]
```

Liczby poprzedzające treść trafień to numery linii w pliku zawierających kolejne wystąpienia poszukiwanego wzorca.

Bufor z wynikami wyszukiwania pozwala na łatwe obejrzenie kontekstu każdego wystąpienia — po jego wskazaniu kursorem lub myszą na liście wyników i naciśnięciu klawisza `ENTER`, ukaże się kontekst żądanego wystąpienia.

6 Dostępność danych

Jak wspomniałem we wstępie, korpus słownika frekwencyjnego jest dostępny na warunkach licencji GNU; wyboru tej podstawy prawnej autorzy dokonali z inicjatywy Janusza S. Bienia. Właściwy korpus został udostępniony na zasadach Powszechnej Licencji Publicznej GNU (ang. GNU General Public License — GPL [4], [5]), zaś fragmenty słownika frekwencyjnego stanowiące jego dokumentację na zasadach Licencji GNU Swobodnej Dokumentacji (ang. GNU Free Documentation License — GFDL [6], [7]), wydanych przez Fundację Oprogramowania Swobodnego (ang. Free Software Foundation)[4].

Licencja GPL posługuje się pojęciem *oprogramowania swobodnego* (ang. free software), przez co rozumie się przyznanie wszystkim zainteresowanym na mocy umowy prawnej pełnego prawa do uruchamiania programu, jego kopiowania, rozpowszechniania i modyfikacji (co stanowi istotną nowość w procesie licencyjnym — tego rodzaju umowy wprowadzają zwykle ograniczenia na używanie i udostępnianie programów). Na co warto zwrócić uwagę, oprogramowanie zostaje „wyzwolone" na różnych poziomach użycia — powstaje nie tylko możliwość samego uruchamiania programu czy

[4]Termin *free software* jest często tłumaczony jako *oprogramowanie wolne*, przymiotnik *swobodny* wydaje się jednak lepiej oddawać jego charakter.

dystrybucji jego kopii, ale także sprawdzenia użytych w nim konstrukcji aż po prawo do jego modyfikacji i dystrybucji stworzonej przez siebie wersji.

Jedynym ograniczeniem wprowadzanym przez licencję jest konieczność udzielenia odbiorcy redystrybuowanego oprogramowania oryginalnego lub pochodnego wszystkich praw zapewnianych przez autora, co w praktyce oznacza, że programy powstałe w wyniku modyfikacji źródeł dostępnych na licencji GPL muszą być także dostępne na tych samych warunkach (z pełnym prawem do udostępniania, modyfikacji i redystrybucji).

Licencja GFDL, oparta na GPL, umożliwia zapewnienie użytkownikom praw do kopiowania i redystrybucji (z ewentualnymi zmianami) podręczników i pozostałych rodzajów dokumentacji. Podobnie jak w przypadku GPL, dokumenty pochodne muszą zostać udostępnione na prawach licencji oryginalnej. Dokumentacja podległa tej licencji może zostać, przy zachowaniu określonych warunków, modyfikowana (w tym tłumaczona na inne języki), kopiowana, wypożyczana i udostępniania do użytku publicznego.

A Dodatek: Praca z płytą CD

A.1 Pobieranie obrazu i nagrywanie płyty

Aktualna edycja wzbogaconego korpusu słownika frekwencyjnego dostępna jest w formie skompresowanego obrazu ISO płyty CD pod adresem `http://www.mimuw.edu.pl/polszczyzna/`. Po pobraniu pliku (jego rozmiar wynosi ok. 50 MB) należy go zdekompresować i zapisać na płycie CD używając dowolnego programu nagrywającego. Tak powstała płyta będzie mogła zostać odczytana w dowolnym napędzie CD.

A.2 Format danych i uruchamianie płyty

Wszystkie zawarte na płycie pliki zapisane są w uniwersalnych, łatwych do przetwarzania formatach. Pliki w formacie PDF można przeglądać i drukować korzystając z darmowego programu Adobe Acrobat Reader, natomiast pliki tekstowe — używając edytora Emacs.

Nowa edycja płyty zostanie skonfigurowana w taki sposób, by po włożeniu jej do napędu CD-ROM Emacs uruchomił się automatycznie.

A.3 Informacje dla korzystających z poprzedniej edycji płyty

Użytkownicy poprzedniej edycji płyty pragnący wykorzystać możliwości Emacsa w systemie MS Windows mogą pobrać go z Internetu, korzystając z oryginalnej dystrybucji (`http://www.gnu.org/software/emacs/windows/ntemacs.html`), bądź też wykorzystać Emacsa będącego częścią dystrybucji TEXLive [3] lub przygotowanej przez Janusza S. Bienia płyty CD [1].

Literatura

[1] Bień, Janusz S. (red.). *Wybrane narzędzia przetwarzania tekstów wielojęzycznych dla Windows 95/98/NT/2000 i komputerów PC*. Wersja 0.93P — grudzień 2001. Płyta CD-ROM. Skompresowany obraz płyty dostępny pod adresem `http://www.orient.uw.edu.pl/wnptw/wnptw093p.iso.bz2`.

[2] Bień, Janusz S.; Woliński, Marcin (red.). *Wzbogacony korpus Słownika frekwencyjnego polszczyzny współczesnej*. Warszawa 2001. Płyta CD-ROM. Skompresowany obraz płyty dostępny pod adresem `http://www.mimuw.edu.pl/polszczyzna/wksf/wksf.iso.bz2`.

[3] *TeX Live — TeX on CD-ROM*. Płyta CD-ROM. Najnowsza wersja dystrybucji dostępna pod adresem `http://tug.ctan.org/tex-archive/systems/texlive/Images`.

[4] *GNU General Public License*, wersja 2, czerwiec 1991, `http://www.gnu.org/copyleft/gpl.html`.

[5] *Powszechna Licencja Publiczna GNU*, wersja 2, czerwiec 1991. Nieoficjalne tłumaczenie [4] dostępne pod adresem `http://gnu.org.pl/text/licencja-gnu.html`.

[6] *GNU Free Documentation License*, version 1.2, listopad 2002, `http://www.fsf.org/copyleft/fdl.html`.

[7] *Licencja GNU Wolnej Dokumentacji*, wersja 1.1, marzec 2000, Nieoficjalne tłumaczenie poprzedniej wersji [6] dostępne pod adresem `http://gnu.org.pl/text/GFDL-pl.html`.

[8] Bień, Janusz S.; Szafran, Krzysztof. Analiza morfologiczna języka polskiego w praktyce. *Biuletyn Polskiego Towarzystwa Językoznawczego*, zeszyt LVII (2001), s. 171-184. Autoryzowana wersja tekstu jest udostępniona w Internecie w formacie PDF (`http://www.orient.uw.edu.pl/publikacje/JSB-KS-PTJ01.pdf`) i Postscript (`http://www.orient.uw.edu.pl/publikacje/JSB-KS-PTJ01.ps`).

[9] Bień, Janusz S.; Woliński, Marcin. *Wzbogacony korpus Słownika frekwencyjnego polszczyzny współczesnej*. W: [2], `wksf.pdf`.

[10] Bień, Janusz S.; Woliński, Marcin (red.). *Numeryczne kody gramatyczne we wzbogaconym korpusie Słownika frekwencyjnego polszczyzny współczesnej*. W: [2], `Dokumentacja\kodynum.pdf`.

[11] Friedl, Jeffrey E.F. *Wyrażenia regularne*, Warszawa 2001, Wydawnictwo Helion.

[12] Głowińska, Katarzyna. *Taksonomia morfologiczna dla Słownika frekwencyjnego*. W: [2], `Dokumentacja\taksonomia.pdf`.

[13] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne:*
tom 1, części 1-2: Teksty popularnonaukowe;
tom 2, części 1-2: Drobne wiadomości prasowe;
tom 3, części 1-2: Publicystyka;
tom 4, części 1-3: Proza artystyczna;
tom 5, części 1-25: Dramat artystyczny. Warszawa 1974-1977, PAN.

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

[15] 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.

[16] Ogrodniczuk, Maciej. *Wykorzystanie SGML i TEI do zapisu polskich danych lingwistycznych.* Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa, wrzesień 2000. Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego.

[17] Ogrodniczuk, Maciej. *Wzbogacenie korpusu słownika frekwencyjnego o nowe kody gramatyczne.* W: [2], `Dokumentacja\taksonomia.pdf`.

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

Uwagi o wersji elektronicznej

Artykuł

> Maciej Ogrodniczuk.
> Nowa edycja wzbogaconego korpusu słownika frekwencyjnego.
> Stanisław Gajda (red.),
> *Językoznawstwo w Polsce. Stan i perspektywy.*
> Polska Akademia Nauk — Komitet Językoznawstwa,
> Uniwersytet Opolski — Instytut Filologii Polskiej.
> Opole 2003, s. 181–190.
> `http://www.mimuw.edu.pl/~jsbien/MO/JwP03/`
> ISBN 83-86881-36-4

jest dostępny w formacie PDF (około 200 Kb)

> `http://www.mimuw.edu.pl/~jsbien/MO/JwP03/MO-JwP03.pdf`

i PostScript (około 170 Kb)

> `http://www.mimuw.edu.pl/~jsbien/MO/JwP03/MO-JwP03.ps`

Rekomendowany sposób cytowania niniejszego artykułu w formacie LaTeX2e:

```
\bibcite{MO-JwP03}
Maciej Ogrodniczuk.
Nowa edycja wzbogaconego korpusu słownika frekwencyjnego.
Stanisław Gajda (red.),
\textit{Językoznawstwo w Polsce. Stan i perspektywy}.
% Polska Akademia Nauk --- Komitet Językoznawstwa,
% Uniwersytet Opolski --- Instytut Filologii Polskiej.
Opole 2003, s. 181--190.
\url{http://www.mimuw.edu.pl/~jsbien/MO/JwP03/}.
% ISBN 83-86881-36-4
```

Data przygotowania wersji elektronicznej:

> 12 grudnia 2003 r.