

Maciej Ogrodniczuk, Katarzyna Głowińska, Marcin Woliński

Elektroniczna wersja korpusu *Słownika frekwencyjnego polszczyzny współczesnej*

24 września 2001 roku

Obecna postać elektronicznej wersji korpusu powstała w wyniku prac ... Jeden ze stylów został uporządkowany przez Martę Nazarczuk w ramach pracy magisterskiej [3]. Korekta pozostałych wykonana przez ... została sfinansowana przez projekt KBN

◁ Potrzebne
sugestie do
rozbiegówki

Korpus w tej postaci miał jawnie oznakowane jedynie formy, które osoby tworzące fiszki uznały za niejednoznaczne. Pracę polegającą na dodaniu znaczników morfosyntaktycznych do wszystkich form w korpusie wykonał Maciek Ogrodniczuk. Nowy system znaczników został zaprojektowany przez Katarzynę Głowińską i Marcina Wolińskiego. Projekt ten zastosował do korpusu (ze zmianami) Maciek Ogrodniczuk. Do automatycznego dodania brakujących znaczników (i po części do weryfikacji istniejących kodów) wykorzystano analizator morfologiczny SAM autorstwa Krzysztofa Szafrana ([5]). Wykorzystano także informacje obecne na listach rangowych *Słownika*.

◁ Jak to było?

1. Struktura korpusu

Korpus składa się z 5 podkorpusów, reprezentujących różne style współczesnej polszczyzny:

◁ Czy są jakieś
oficjalne nazwy
stylów?

A	publicystyka
B	drobne notatki prasowe
C	popularno-naukowy
D	proza
E	dramat

Dla każdego ze stylów zgromadzono 2000 próbek¹ po ok. 50 słów. W sumie korpus liczy więc ok. 10 000 próbek i 500 000 słowoform.

◁ Co z tymi
brakami?

Elektroniczna wersja korpusu ma postać pięciu plików tekstowych w stronie kodowej CP 1250 (MS Windows). Próbkę są w plikach oddzielone od siebie pustym wierszem (czyli dwoma następującymi bezpośrednio znakami nowego wiersza).

Pierwszy wiersz próbki zawiera jej numer i opis bibliograficzny. Poszczególne jego elementy są oddzielone znakiem tyldy ~. Numer próbki składa się z litery oznaczającej podkorpus (A, B, C, D lub E) i czterech cyfr wyrażających numer kolejny w obrębie podkorpusu.

¹ Wyjątkiem jest podkorpus B, w którym próbki są ponumerowane od 1 do 2080, ale 19 fiszek brakuje (numer 353, 368, 369, 1256–1267, 1349–1352).

Opis bibliograficzny obejmuje nazwisko autora, tytuł dzieła, nazwę wydawnictwa, rok wydania, określenie miejsca z którego pochodzi próbka, a w przypadku periodyków ich nazwę i numer.

◄ Niestety
informacja jest
różna
w różnych
stylach

Pozostałe wiersze próbki zawierają kolejne słowoformy tekstu, po jednej w wierszu. Bezpośrednio po słowoformie znajduje się umieszczona w nawiasach kwadratowych jej interpretacja jako wykładnika jakiegoś leksemu. Znaki interpunkcyjne są umieszczane w tym samym wierszu, jeżeli nie są oddzielone odstępami od formy. Tak więc znaki „doklejone” do początku formy (otwierające cudzysłowy i nawiasy) umieszczane są na początku wiersza. Znaki stojące na końcu formy (zamykające cudzysłowy i nawiasy, przecinek, średnik, kropka, wykrzyknik, pytajnik, itd.) umieszczane są na końcu wiersza, po interpretacji morfologicznej. Pauzy typowo stanowią osobny wiersz w pliku.

Nie ma w plikach informacji o podziale na wersy w oryginalnych dziełach, ponieważ informacja ta nie została wprowadzona na fiszki. Nie zostały też specjalnie oznaczone końce zdań.

Interpretacja słowoformy składa się z trzech części rozdzielonych przecinkami. Są to kolejno: nazwa leksemu przypisanego danej formie, kod cyfrowy pochodzący z pierwotnego ręcznego znakowania (jeżeli występował), nowy znacznik morfosyntaktyczny. Jeżeli danej formie nie udało się jednoznacznie przyporządkować nowego znacznika, na tej pozycji stoją wariantywne znaczniki oddzielone ukośnikami (/).

Oto przykładowy fragment oznakowanego korpusu:

```
B0001~Dziennik Bałtycki~05.01.1963~str. 5~kol. 1
Wychodząc[wychodzić,,V----W--N----P]
z[z,62,P-G-----P]
założenia[założenie,121,SSGN-----P] ,
że[że,,D-----P/C-----P]
różne[różny,212,APNOP-----P/APNRP-----P]
są[być,,VP---3TON----P]
właściwości[właściwość,112,SPNF-----P]
i[i,,C-----P]
różne[różny,212,APNOP-----P/APNRP-----P]
możliwości[możliwość,112,SPNF-----P]
zakładów[zakład,,SPGI-----P]
pracy[praca,121,SSGF-----P] ,
organizatorzy[organizator,,SPNP-----P/SPVP-----P]
konkursu[konkurs,,SSGI-----P]
przedstawili[przedstawić,,VP-0-3POD----P]
```

Ponadto w korpusie występują następujące symbole specjalne:

◄ Oznaczenia wg
Marty.
Nie
sprawdzono
na ile
konsekwentnie
stosowane

- [&] urwany tekst ostatniego zdania próbki; zdanie nie kończy się kropką, znakiem zapytania lub wykrzyknikiem, na końcu tekstu próbki występuje wielokropek, lub nie ma żadnego znaku interpunkcyjnego
- [#] oznaczenie końca próbki, z oryginału której wykreślono więcej niż jeden znak; pojawia się na końcu próbki
- [~] w treści próbki brak fragmentu tekstu; występuje w miejscu wykreślonych formuł, np. wzorów, symboli etc.
- [|] pojawia się przy końcu próbki, w miejscu, w którym na papierowej fiszce między słowoformami zaznaczono pionową kreskę; prawdopodobnie zaznaczenie granicy po 50 wyrazie

- [>] występuje na początku próbki, jeżeli pierwsze zdanie na papierowej fiszce nie rozpoczynało się dużą literą (możliwe, że nie przepisano początku zdania z pozycji źródłowej, gdyż występował na poprzedniej stronie); oznaczenie pojawia się na początku fiszki
- [^] ???

2. Kody cyfrowe pochodzące z ręcznego znakowania form niejednoznacznych

- 111 — rzeczownik w mianowniku, liczba pojedyncza
112 — rzeczownik w mianowniku, liczba mnoga
121 — rzeczownik w dopełniaczu, liczba pojedyncza
122 — rzeczownik w dopełniaczu, liczba mnoga
131 — rzeczownik w celowniku, liczba pojedyncza
132 — rzeczownik w celowniku, liczba mnoga
141 — rzeczownik w bierniku, liczba pojedyncza
142 — rzeczownik w bierniku, liczba mnoga
151 — rzeczownik w narzędniku, liczba pojedyncza
152 — rzeczownik w narzędniku, liczba mnoga
161 — rzeczownik w miejscowniku, liczba pojedyncza
162 — rzeczownik w miejscowniku, liczba mnoga
171 — rzeczownik w wołacz, liczba pojedyncza
172 — rzeczownik w wołacz, liczba mnoga
- 211 — przymiotnik w mianowniku, liczba pojedyncza
212 — przymiotnik w mianowniku, liczba mnoga
221 — przymiotnik w dopełniaczu, liczba pojedyncza
222 — przymiotnik w dopełniaczu, liczba mnoga
231 — przymiotnik w celowniku, liczba pojedyncza
232 — przymiotnik w celowniku, liczba mnoga
241 — przymiotnik w bierniku, liczba pojedyncza
242 — przymiotnik w bierniku, liczba mnoga
251 — przymiotnik w narzędniku, liczba pojedyncza
252 — przymiotnik w narzędniku, liczba mnoga
261 — przymiotnik w miejscowniku, liczba pojedyncza
262 — przymiotnik w miejscowniku, liczba mnoga
271 — przymiotnik w wołacz, liczba pojedyncza
- 31 — liczebnik w mianowniku
32 — liczebnik w dopełniaczu
33 — liczebnik w celowniku
34 — liczebnik w bierniku
35 — liczebnik w narzędniku
36 — liczebnik w miejscowniku
- 41 — zaimek w mianowniku
42 — zaimek w dopełniaczu
43 — zaimek w celowniku
44 — zaimek w bierniku
45 — zaimek w narzędniku
46 — zaimek w miejscowniku
- 501 — Nic z poniższych

- 5 —
- 50 — błąd, powinno być [5]
- 511 — czasownik zwrotny w bezokoliczniku w funkcji formy czasu przyszłego, będą[56] odbywać[511] się
- 51 — czasownik niezwrotny w bezokoliczniku w funkcji formy czasu przyszłego, podróżować[51] będą[56]
- 521 — czasownik zwrotny jako składnik formy czasu przyszłego zakończonych na -t będę[56] się zachowywała[521]
- 52 — czasownik niezwrotny jako składnik formy czasu przyszłego zakończonych na -t będziesz[56] ty siedział[52]
- 531 — czasownik zwrotny jako składnik formy czasu przeszłego z ruchomą końcówką bardzośmy się kłócili[531]
- 53 — czasownik niezwrotny jako składnik formy czasu przeszłego z ruchomą końcówką myśmy gadali[53]
- 541 — czasownik zwrotny jako składnik formy trybu warunkowego z ruchomą partykułą -by by[8] się ucieszyli[541]
- 54 — czasownik niezwrotny jako składnik formy trybu warunkowego z ruchomą partykułą -by można[54] by[8] porozmawiać
- 551 — czasownik zwrotny użyty w czasie teraźniejszym w funkcji trybu rozkazującego niech pan[111] się napije[551]
- 55 — czasownik niezwrotny użyty w czasie teraźniejszym w funkcji trybu rozkazującego niech będzie[55] jasno
- 56 — czasownik być, bywać, zostać jako człon składowy w czasach złożonych odwiedzać[51] będą[56]
- 57 — czasownik być, bywać, zostać jako człon składowy strony biernej byłem[57] przygotowany[211]
- 61 — zaimek łączący się z mianownikiem
- 62 — zaimek łączący się z dopełniaczem
- 63 — zaimek łączący się z celownikiem
- 64 — zaimek łączący się z biernikiem
- 65 — zaimek łączący się z narzędnikiem
- 66 — zaimek łączący się z miejscownikiem
- 7 — wykrzyknik
- 8 — partykuła
- 9 — spójnik

3. Nowe znaczniki morfosyntaktyczne

Każdy znacznik morfosyntaktyczny ma postać napisu złożonego z 14 znaków (tak samo dla wszystkich klas gramatycznych). Kategorie gramatyczne odpowia-

dające poszczególnym pozycjom zestawiono w poniższej tabeli. Jeżeli dana kategoria nie jest adekwatna dla danej klasy gramatycznej, na odpowiedniej pozycji stoi znak minusa.

Pozycja	Znaczenie	Kod	Objaśnienie
1	klasa gramatyczna (część mowy)	V S A N Z D P C I T X	czasownik rzeczownik przymiotnik liczebnik zaimek przysłówek przyimek spójnik wykrzyknik partykuła kod nieznany
2	liczba	S P	pojedyncza mnoga
3	przypadek ²	N G D A I L V	mianownik dopełniacz celownik biernik narzędnik miejscownik wołacz
4	rodzaj	M P A I F N O R T	męski męskoosobowy (l. poj.) męskozwierzęcy męskorzeczowy żeński nijaki męskoosobowy (l. mn.) niemęskoosobowy plurale tantum
5	stopień	P C S	równy wyższy najwyższy
6	osoba lub oznaczenie formy bezosobowej czasownika	1 2 3 I B U W	pierwsza osoba druga osoba trzecia osoba bezokolicznik bezosobnik (forma na <i>-no</i> , <i>-to</i>) imiesłów przysłówkowy uprzedni imiesłów przysłówkowy współczesny
7	czas	T P F	teraźniejszy przeszły przyszły złożony

² Pozycja została także wykorzystana do przechowania informacji o wartości przypadkowej formy dla liczebników zbiorowych oraz kategorii przypadku form, z którymi łączą się przyimki.

Pozycja	Znaczenie	Kod	Objaśnienie
8	tryb	O P R	oznajmujący przypuszczający rozkazujący
9	aspekt	D N	dokonany niedokonany
10	strona	C B Z	czynna bierna zwrotna
11	akcentowość	T N	forma akcentowana forma nieakcentowana
12	poprzyimkowość	T N	forma poprzyimkowa forma niepoprzyimkowa
13	oznaczenie dodatkowe form czasow- nikowych ³	I S P W R B O	bezokolicznik jako forma składowa czasu przyszłego (będzie <i>pisać</i>) forma na -ł jako składowa form czasu przyszłego (będzie <i>pisał</i>) forma na -ł jako czas przeszły z ruchomą końcówką (sukroś <i>zjadł</i>) forma trybu przypuszczającego z ruchomą partykułą (bym <i>napisał</i>) opisowa forma trybu rozkazującego trzeciej osoby (niech <i>pisze</i>) czasownik <i>być</i> jako składowa form czasów złożonych (będzie <i>pisał</i>) czasownik <i>być</i> , <i>bywać</i> , <i>zostać</i> jako składowe form strony biernej (<i>jest</i> czytany, <i>bywał</i> sporządzany, <i>zostanie</i> zapisany)
14	oznaczenie nazw własnych	P W S	nazwa pospolita nazwa własna skrótowiec

Literatura

- [1] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Warszawa, 1974. Uniwersytet Warszawski.
- [2] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Szafran, Krzysztof. *Słownik frekwencyjny polszczyzny współczesnej*, Kraków, 1990. Instytut Języka Polskiego PAN.
- [3] 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.
- [4] Saloni, Zygmunt. *Słownik frekwencyjny polszczyzny współczesnej*. W: Computer-World. S. 16-17. 4 listopada 1991.
- [5] Szafran, Krzysztof. *Analizator morfologiczny SAM-95 — opis użytkowy*. Maj 1996. Instytut Informatyki Uniwersytetu Warszawskiego.

³ Pole z oznaczeniami dodatkowych własności form czasownikowych zostało wprowadzone w celu reprezentacji obecnych na listach rangowych Słownika Frekwencyjnego opisów funkcji czasownikowych form złożonych.