Prace nad korpusem *Słownika Frekwencyjnego Języka Polskiego*

Maciej Ogrodniczuk

2003

Spis treści

Wstęp

Przetwarzanie danych *Korpusu frekwencyjnego polszczyzny współczesnej* miało na celu automatyczne uzupełnienie i rozszerzenie dostępnej informacji morfologicznej z uwzględnieniem oryginalnych kodów korpusowych.

Materiałem źródłowym dla prac był zbiór 500 000 słów zawartych w równych transzach odpowiadających pięciu stylom współczesnej polszczyzny pisanej (styl popularnonaukowy, wiadomości prasowych, publicystyczny, prozy artystycznej i dramatu artystycznego). Teksty stylów pochodzą z lat 1963-1967 i zostały dobrane metodą losowania. Każda transza zawiera 2000 próbek ciągłego tekstu o długości ok. 50 słów. Wykorzystane zbiory korpusowe (pięć plików poszczególnych stylów wraz z danymi bibliograficznymi źródeł) zostały przekazane przez K. Szafrana.

Opisy nowego typu tworzono z wykorzystanem analizatora morfologicznego SAM-99[1] Krzysztofa Szafrana [6] — opartego na schematycznym indeksie a tergo Jana Tokarskiego [7] narzędzia udostępniającego dla pojedynczych słów tekstowych charakterystykę morfologiczną form oraz identyfikującego ich postać podstawową.

Wyniki analizy zostały za pomocą dostępnych na licencji GNU skryptów w języku Perl poddane dodatkowym przekształceniom w celu dostosowania reprezentacji polskiej informacji morfologicznej do postaci zgodnej z taksonomią pozycyjną K. Głowińskiej i M. Wolińskiego [2].

Materiałem wynikowym jest pięć plików odpowiadających zawartości każdej z transz zawierających dane próbek w następującym formacie:

- każda próbka rozpoczyna się wierszem informacji bibliograficznej,
- w kolejnych wierszach znajdują się poszczególne formy próbki, po jednej w każdym wierszu,
- wiodące znaki interpunkcyjne (cudzysłów i nawias ótwierający) podawane są bezpośrednio przed formą analizowaną,

[1] Jest to prawdopodobnie jedyny analizator morfologiczny języka polskiego dostępny bezpłatnie w Internecie: `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/`.

- za formą znajduje się wynik analizy ujęty w nawiasy kwadratowe, którego poszczególne człony oddzielone są przecinkami:
  - forma podstawowa,
  - oryginalne (tj. pochodzące ze źródłowej wersji korpusu) dodatkowe oznaczenie skrótowca lub nazwy własnej (napis pusty w przypadku nazwy pospolitej),
  - oryginalne trzycyfrowe oznaczenie kodowe,
  - kod morfologiczny wg taksonomii docelowej.
- końcowe znaki interpunkcyjne (cudzysłów i nawias zamykający, przecinek, kropka itp.) podawane są bezpośrednio za wynikiem analizy,
- formy wielowyrazowe, którym odpowiada pojedyncza analiza, podawane są w jednym wierszu,
- próbki oddziela pusty wiersz,

Rozdział 1

Automatyczna analiza morfologiczna z wykorzystaniem analizatora SAM

1.1 Przetwarzanie list rangowych

Pierwsza część prac nad uzupełnianiem tekstu słownika o oznaczenia gramatyczne polegała na przetworzeniu list rangowych[1] słownika w celu uzyskania zbiorczego opisu znajdujących się na nich form.

Listy mają postać 32 plików — zbiorów form rozpoczynających się kolejnymi literami alfabetu polskiego. Każdy z plików podzielony jest na bloki form o wspólnym haśle, wydzielone parametrami częstości względnej i równomierności rozkładu w stylach (początek bloku, patrz opis w artykule [5]) oraz numerem bloku w pliku (koniec bloku). Zapisy na liście form umieszczone są w osobnych liniach i składają się z sześciu wartości określających liczby wystąpień danej formy w poszczególnych stylach i w całym korpusie, po których następuje zapis formy, opcjonalne oznaczenie dodatkowe (plus, jeśli forma jest nazwą własną lub kropka, jeśli jest skrótowcem) oraz ewentualny kod gramatyczny zgodny z opisem podanym w pracy Marty Nazarczuk [4]. Pierwsza linia bloku stanowi ponadto zbiorczy opis częstości wszystkich

[1] Listy rangowe zostały zatem potraktowane jako wiarygodne źródło informacji o formach występujących w całym słowniku — podejście takie nie ogranicza zakresu prac do form pojedynczego stylu (inny wariant przetwarzania mógłby polegać na tworzeniu osobnych, zawężonych list kodów dla poszczególnych stylów), ale może przenieść do tekstu słownika błędy występujące na listach rangowych.

wystąpień form wchodzących w skład paradygmatu formy podstawowej; wartości odpowiadające częstościom są wówczas sumami liczb z poszczególnych kolumn bloku, zaś w polu przewidzianym dla formy umieszczona jest forma podstawowa, oznaczana zwykle wyłącznie kodem części mowy.

Oto przykład pojedynczego bloku form[2]:

```
65.67 143.16
  48   81   70   14    5   218 walka 1
   0    2    0    0    0     2 walce 131
   8   16   17    2    0    43 walce 161
   6    8    2    2    1    19 walk
   8   10   15    4    2    39 walka
   0    6    1    0    0     7 walkach
   1    2    2    0    0     5 walk@a
   5   11    4    2    1    23 walk@e
   4    7    1    1    0    13 walki 112
  16   17   26    3    1    63 walki 121
   0    2    2    0    0     4 walki 142
 106
```

Warto zauważyć, że *nagłówek bloku* (nazwa przyjęta na oznaczenie jego pierwszej linii zawierającej opis formy podstawowej) pełni tu jedynie funkcję grupującą; wystąpieniom formy podstawowej w poszczególnych stylach słownika odpowiada osobna linia zapisu.

Wynikowa postać zapisu jest plikiem tekstowym, w którego kolejnych wierszach znajdują się, oddzielone przecinkami, następujące parametry:

- forma z listy,
- oznaczenia dodatkowe (nazw własnych i skrótowców, dla nazw pospolitych to pole jest puste),
- kod pierwotny (z listy rangowej),
- forma podstawowa dla danej formy,
- nowy kod zgodny z przyjętą taksonomią[3].

[2]Na oznaczenie polskich liter została w materiałach źródłowych przyjęta konwencja prefiksowa `@znak`.

[3]Na tym etapie prac użyty został kod nie uwzględniający jeszcze ostatniej pozycji zapisu taksonomicznego — informacji o własności nazwy (zobacz [2]).

Odpowiadający przedstawionemu wyżej fragmentowi listy zapis wynikowy uzyskuje postać:

```
walce,,131,walka,SSDF---------
walce,,161,walka,SSLF---------
walk,,,walka,SPGF---------
walka,,,walka,SSNF---------
walkach,,,walka,SPLF---------
walką,,,walka,SSIF---------
walkę,,,walka,SSAF---------
walki,,112,walka,SPNF---------
walki,,121,walka,SSGF---------
walki,,142,walka,SPAF---------
```

Przetwarzanie plików zostało dokonane za pomocą skryptów zapisanych w języku Perl i dla celów kontrolnych rozbite na kilka etapów. Oznaczam je kolejnymi numerami odpowiadającymi nazwom plików zawierających treść skryptów.

Pierwszym etapem prac nad uzupełnieniem opisów o dokładne kody gramatyczne było przekształcenie dostępnej informacji w postać łatwą do dalszej obróbki. Zakres przetwarzania dokonany w pliku `01` przedstawia się następująco:

- połączenie 32 plików z listami rangowymi,
- usunięcie informacji o częstości wystąpień i numerów porządkowych bloków,
- przekodowanie zapisu polskich liter (oznaczenie typu `@znak` zostało zamienione na kod zgodny z ISO-Latin 2),
- specjalna interpretacja form nieodmiennych, oznaczonych kodami cyfrowymi 7 (wykrzykniki), 8 (przysłówki), 9 (spójniki) i 10 (partykuły): w przypadku wystąpienia takiego oznaczenia w nagłówku bloku jest ono dopisywane wszystkim hasłom bloku, gdyż tak oznaczone formy nieodmienne nie będą dalej przetwarzane,
- zapis wyników przetwarzania w opisanej wyżej postaci wyjściowej (bez pola nowego kodu).

Zebranie informacji zawartej w listach rangowych i jej dostosowanie do dalszego przetwarzania jest o tyle ważne, że tworzy pewien punkt odniesienia do dalszych prac. Przedstawienie całości użytecznej informacji w prostej postaci pozwala uniknąć kłopotów z obróbką niewygodnych bloków, a ponadto przyczynia się do wykrycia i wyeliminowania zaburzeń w zapisie list rangowych, mogących stać się przyczyną wielu późniejszych błędów przetwarzania[4].

1.2 Przygotowanie danych dla analizatora SAM

Następnym etapem stało się przygotowanie listy haseł do przetworzenia analizatorem morfologicznym w celu uściślenia zapisów gramatycznych[5]. Lista ta jest zbiorem początkowych pól otrzymanego wcześniej pliku, z pewnymi zastrzeżeniami[6] (wszystkie przykłady oryginalne):

- analizator nie jest w stanie przeprowadzić analizy haseł z apostrofem (*college'u*) ani form wielowyrazowych w ogólnej postaci (*czym prędzej*, *przede wszystkim*) — są one pomijane przy tworzeniu listy,
- analiza form oznaczonych jako nazwy własne i skrótowce jest w większości przypadków bezużyteczna, czasem jednak (przede wszystkim wtedy, gdy zachodzi synkretyzm formy z odpowiadającą jej nazwą pospolitą) może zakończyć się pomyślnie (forma *abelard* — brak analizy, ale *akropol* — analiza poprawna; podobnie *pzpr* — brak analizy, ale *cekaem* — analiza poprawna),
- po niewielkich zabiegach można poddać analizie
  - formy przymiotnikowe z łącznikiem (*afro-azjatycki*); do analizy przedstawiamy drugi człon,
  - formy rzeczownikowe z łącznikiem (*chłop-robotnik*, ale i *decha-ochraniacz*); do analizy przedstawiamy pierwszy człon,

[4] Wykaz znalezionych na etapie przetwarzania list rangowych błędów w ich strukturze wymienia Dodatek B.

[5] Dla wielu form, mimo dość dokładnego opisu gramatycznego w poprzedniej postaci kodu, analizator pozwolił na jego rozszerzenie — np. o informację o rodzaju rzeczownika i przymiotnika.

[6] Ustalenia te, oprócz ograniczenia przetwarzania, mają na celu wyeliminowanie potencjalnie błędnych interpretacji (np. osobna analiza składników formy złożonej jest zwykle całkowicie błędna).

  - opisowe przysłówki i przymiotniki, rozpoczynające się członami *co* (*co najmniej*), *jak* (*jak największy*), *na* (*na gorąco*) i *za* (*za duży*); analizujemy drugi człon,
  - imiesłowy przymiotnikowe i rzeczowniki odczasownikowe zakończone zaimkiem *się* (*domagający się*, *bogacenie się*); analizujemy pierwszy człon.

- mimo pewnych prawidłowości (niestety, nieistotnych lub mało istotnych w procesie automatycznej analizy w stosunku do koniecznych do poniesienia nakładów pracy) nie jest możliwa analiza pozostałych powszechnych form regularnych rozpoczynających się członami
  - *bez* (*bez mała*, *bez ustanku*),
  - *co* (*co gorsza*, *co najmniej*),
  - *do* (*do cna*, *do wczoraj*),
  - *na* (*na bieżąco*, ale i *na bosaka*, *na powrót*),
  - *nie* (*nie lada*, *nie sposób*),
  - *o* (*o ile*, *o tyle*),
  - *od* (*od dawna*, *od wewnątrz*),
  - *po* (*po bohatersku*, *po omacku*),
  - *w* (*w dwójnasób*, *w zamian*),
  - *z* (*z dala*, *z nagła*).

Działanie skryptu `02` polega zatem na sczytaniu form do osobnego pliku zgodnie z powyższymi regułami. Program oznaczony jako `03`, będący skryptem systemu operacyjnego, przetwarza natomiast powstały plik analizatorem SAM-99. W wyniku tej operacji otrzymujemy ciąg analiz o ustalonej strukturze: każda jest blokiem tekstu rozpoczynającym się formą analizowaną, umieszczoną w osobnej linii i oddzieloną od wyników znakiem procenta. Po niej następuje ujęty w nawiasy klamrowe ciąg linii zawierających poszczególne analizy (każda z nich ujęta jest również w nawiasy klamrowe), identyfikujące rozpoznaną formę podstawową i kategorię fleksyjną[7].

[7] Dokładny opis postaci analiz i znaczenie kodów wynikowych znajduje się w pracy autora analizatora SAM-99, Krzysztofa Szafrana [6].

Oto przykładowy blok analiz wieloznacznej formy *kurze*:

```
kurze
%
{{(lN) < kurz(mII::m3)+ }%
{(LV) < kur(mIV::m2 mIV m3)+ }%
{(D) < kura(żIV::)+ }%
{(5) < kurzy(A::[p])+ } }%
```

Analizator rozpoznał ją jako formę mianownika liczby mnogiej rzeczownika *kurz*, miejscownika lub wołacza liczby pojedynczej rzeczownika *kur*, celownika liczby pojedynczej rzeczownika *kura*, a także jako przymiotnik *kurzy* w mianowniku lub bierniku liczby pojedynczej rodzaju nijakiego (*kurze pisklę*) albo mnogiej rodzaju niemęskoosobowego (*kurze pisklęta*). Jak widzimy, otrzymywana analiza może być dość złożona.

1.3 Interpretacja wyników analizy morfologicznej

Następnym etapem jest rozpoznanie kodów analizatora i ich zamiana na kody nowego zestawu po uwzględnieniu dodatkowych warunków (istnienia synkretyzmów, braku kategorii morfologicznych dla poszczególnych części mowy itd.[8]) Służy do tego skrypt w Perlu oznaczony jako `04`, dokonujący dekompozycji analiz i odpowiednich przekodowań. Wyniki zapisywane są w pliku pośrednim w następującym formacie:

- analiza każdego hasła w osobnej linii,
- forma analizowana oddzielona od wyników analiz dwukropkiem,
- warianty analiz oddzielone od siebie przecinkiem,
- forma podstawowa oddzielona od kodu morfologicznego ukośnikiem.

W przypadku braku analiz (nieznajomość formy lub nierozpoznany wynik analizy morfologicznej) po dwukropku występuje napis pusty. Gdy dla jednej formy podstawowej istnieje wiele analiz, forma podstawowa powtarzana

[8] Dokładniejszy opis rodzaju dokonanych przekodowań i interpretacji opisuje Dodatek D.

jest przy każdym oznaczeniu kodowym, by ułatwić późniejsze przetwarzanie. Oto wynik działania pliku 04 dla podanego wyżej bloku (podziału na kilka linii dokonano dla zwiększenia przejrzystości zapisu, w rzeczywistości jest to pojedyncza linia analiz):

```
kurze:kurz/SPNI----------,kurz/SPAI----------,
      kurz/SPVI----------,kur/SSLA----------,
      kur/SSVA----------,kura/SSDF----------,
      kura/SSLF----------,kurzy/ASNNP---------,
      kurzy/ASANP---------,kurzy/APNOP---------,
      kurzy/APNRP---------,kurzy/APARP---------
```

W zasadzie nic nie stoi na przeszkodzie, by w jednym kroku dokonywać od razu przekodowania wyników automatycznej analizy morfologicznej i ich ograniczenia w zależności od istniejących kodów korpusowych. Czynność ta zostanie jednak wykonana w następnym etapie przetwarzania — ma to na celu stworzenie skryptu uniwersalnego, zdolnego w rozsądny sposób interpretować wyniki analizy morfologicznej niezależnie od przetwarzania plików Korpusu. Skrypt ten (04) może zostać użyty do przedstawiania wyników analizy dowolnych danych w nieco przejrzystszej (i przede wszystkim łatwiejszej w obróbce) formie niż ta nadawana wynikom SAM-a. Zaletą skryptu jest też możliwość zmiany wynikowego zestawu kodowego, co zostało osiągnięte poprzez związanie kategorii morfologicznych z wartościami zmiennych definiowanych globalnie i używanie tych zmiennych w kodzie procedur w miejscu oznaczeń wpisywanych bezpośrednio. Każda zmiana zestawu wiąże się zatem z pojedynczą modyfikacją zapisu wartości zmiennej w początkowej części skryptu i jest niezależna od przetwarzania wyników analizy.

Ujednoznacznianie wyników analizy morfologicznej na podstawie pierwotnych kodów słownika dokonywane jest przez skrypt 05. Po dekompozycji obu rodzajów kodów są one porównywane i w przypadku zgodności interpretacja zapisywana jest do końcowego pliku wynikowego — kategorie morfologiczne są w ten sposób zawężane do ustalonych wcześniej wartości.

Rozbicie przekodowania wyników analizy morfologicznej na dwa etapy wiąże się z pewną dodatkową czynnością (wykonywaną przez skrypt 05) — dostosowaniem wyników analiz do konwencji słownikowej[9]. Dzieje się tak, gdy

[9]Niektóre kategorie morfologiczne mogą być interpretowane na wiele sposobów — przykładem jest kategoria imiesłowów przymiotnikowych, przypisywanych przez analizator SAM-99 do kategorii czasowników, a przez konwencję oznaczeń słownika — do kategorii przymiotników. Rozbieżność ta wiązała się z dodatkowym przetwarzaniem; w formacie wynikowym została zastosowana konwencja przymiotnikowa (oryginalna).

analiza morfologiczna formy odbiega od sposobu zapisu jej kategorii w plikach słownika — chcąc uzyskać interpretację wyników analizy niezależną od jej późniejszych zastosowań (skrypt `04`), musimy dodatkowo obsłużyć konwencję słownikową w następnym kroku przetwarzania.

Dodatek A
Taksonomia morfologiczna

Poniżej przedstawiam taksonomię morfologiczną — stosowany w wynikowej wersji Korpusu schemat oznaczeń fleksyjnych do dokładnego opisu tekstów polskich oparty o projekt pod nazwą *Zestaw Kodowy do Reprezentacji Polskiej Informacji Lingwistycznej* autorstwa Katarzyny Głowińskiej i Marcina Wolińskiego.

Pozycyjny charakter kodów, opisany w poniższej tabeli, zapewnia łatwość ich przetwarzania. Dla uproszczenia pominięto oznaczenia nieadekwatności kategorii, określane znakiem minusa na odpowiedniej pozycji kodowej.

Pozycja	**Znaczenie**	**Kod**	**Objaśnienie**
1	typ jednostki zdania (część mowy)	V	czasownik
		S	rzeczownik
		A	przymiotnik
		N	liczebnik
		Z	zaimek
		D	przysłówek
		P	przyimek
		C	spójnik
		I	wykrzyknik
		T	partykuła
		X	kod nieznany
2	liczba	S	pojedyncza
		P	mnoga
3	przypadek[1]	N	mianownik
		G	dopełniacz
		D	celownik
		A	biernik
		I	narzędnik
		L	miejscownik
		V	wołacz

[1]Pozycja została także wykorzystana do przechowania informacji o wartości przypad-

Pozycja	Znaczenie	Kod	Objaśnienie
4	rodzaj	M	męski
		P	męskoosobowy (l. poj.)
		A	męskozwierzęcy
		I	męskorzeczowy
		F	żeński
		N	nijaki
		O	męskoosobowy (l. mn.)
		R	niemęskoosobowy
		T	plurale tantum
5	stopień	P	równy
		C	wyższy
		S	najwyższy
6	osoba	1	pierwsza osoba
	lub oznaczenie	2	druga osoba
	formy	3	trzecia osoba
	bezosobowej	I	bezokolicznik
	czasownika[2]	B	forma bezosobowa (-no, -to)
		U	imiesłów przysłówkowy uprzedni
		W	imiesłów przysłówkowy współczesny
7	czas	T	teraźniejszy
		P	przeszły
		F	przyszły złożony
8	tryb	O	oznajmujący
		P	przypuszczający
		R	rozkazujący
9	aspekt	D	dokonany
		N	niedokonany
10	strona	C	czynna
		B	bierna
		Z	zwrotna
11	akcentowość	T	forma akcentowana
		N	forma nieakcentowana
12	poprzyimkowość	T	forma poprzyimkowa
		N	forma niepoprzyimkowa

kowej formy dla liczebników zbiorowych oraz kategorii przypadka form, z którymi łączą się przyimki.

[2]Połączenie kategorii osoby z identyfikatorami form bezosobowych stało się możliwe wobec rozłączności obu grup oznaczeń.

Pozycja	Znaczenie	Kod	Objaśnienie
13	oznaczenie dodatkowe form czasownikowych[3]	I	bezokolicznik jako forma składowa czasu przyszłego (będzie *pisać*)
		S	forma na -ł jako składowa form czasu przyszłego (będzie *pisał*)
		P	forma na -ł jako czas przeszły z ruchomą końcówką (skoroś *zjadł*)
		W	forma trybu przypuszczającego z ruchomą partykułą (bym *napisał*)
		R	opisowa forma trybu rozkazującego trzeciej osoby (niech *pisze*)
		B	czasownik *być* jako składowa form czasów złożonych (*będzie* pisał)
		O	czasownik *być*, *bywać*, *zostać* jako składowe form strony biernej (*jest* czytany, *bywał* sporządzany, *zostanie* zapisany)
14	oznaczenie nazw własnych	P	nazwa pospolita
		W	nazwa własna
		S	skrótowiec

[3]Pole z oznaczeniami dodatkowych własności form czasownikowych zostało wprowadzone w celu reprezentacji obecnych na listach rangowych Słownika Frekwencyjnego opisów funkcji czasowników złożonych. W przypadku ogólnym uzyskanie tego rodzaju informacji drogą analizy automatycznej może nie być zadaniem łatwym.

Dodatek B
Błędy wykryte na listach rangowych słownika

Rozdział ten zawiera wykaz nieprawidłowości odnalezionych na listach rangowych słownika podczas ich przetwarzania w celu dopisania nowych kodów fleksyjnych. Stworzenie ich spisu ma na celu zasygnalizowanie istnienia błędów w wielu kopiach list rangowych.

Pierwszym rodzajem błędów było zazębianie się dwóch bloków haseł polegające na niecelowym połączeniu opisów, co dawało w wyniku błędne przyporządkowanie formy podstawowej pewnej grupie haseł. Wprowadzone poprawki polegały na wstawieniu ograniczników bloku zgodnie ze strukturą pliku z listą, co jednak jest posunięciem doraźnym, gdyż naruszona zostaje w ten sposób numeracja bloków, mogą też wystąpić problemy z zachowaniem właściwych zapisów informacji o częstości. Oto wykaz połączeń:

- *betancourt – betlej*,
- *biec – bieda*,
- *centrala – centralizacja*,
- *dość – doświadczać*,
- *grzeczny – grzegórzki*,
- *istnienie – istota*,
- *keeler – kelner*,
- *kolorysta – kolos*,
- *kopalniany – koparka*,

- *lada – labuda*,
- *lato – latyfundium*,
- *ładnie – ładowacz*,
- *oparcie się – oparty*,
- *organy – orgia*,
- *organy – organiczny*,
- *pp – praca*,
- *rozdysponować – rozdział*,
- *sekretarz – seksualny*.

Pozostałe błędy (także poprawione):

- zaburzona struktura hasła *mieszkanie*,
- hasło *parlamentaryzm* z przecinkiem w środku (*parla,entaryzm*),
- błędny nagłówek dla hasła *żychliński* (cyfra zamiast formy),
- literówki w nagłówkach (poprawne formy w dalszej części bloku): *afrykananin*, *franciszkananin*, *galilei*, *gratulacj*, *konstrukcji*, *kbściół*, *małżeństw*, *samopoctucie*, *uniwersaliźm*, *zakłopotąie*, *zrewidowaąie*.

Dodatek C
Interpretacja kodów analizatora morfologicznego SAM-99 zgodnie z nową taksonomią morfologiczną

Poniżej zamieszczam wykaz interpretacji kodów analizatora morfologicznego SAM-99 wykorzystanych do tworzenia nowych kodów słownika (skrypt `04`, patrz Rozdział 1).

Opis ten może także służyć za uproszczone przedstawienie kodów wynikowych analizatora SAM-99, opisanych dokładnie w pracy [6]. Z niej też pochodzą wszystkie stosowane schematy oznaczeń.

C.1 Formy rzeczownikowe

Badane są kategorie rodzaju, przypadka i liczby. Dla ułatwienia podaję objaśnienia oznaczeń.

Opis rodzaju:

Oznaczenie SAM-a	**Kod nowej taksonomii**	**Rodzaj gramatyczny**
`m1`	`P`	męskoosobowy
`m2`	`A`	męskozwierzęcy
`m3`	`I`	męskorzeczowy
`ż`	`F`	żeński
`n`	`N`	nijaki
`blp`	`T`	plurale tantum

Dla rzeczowników kategoria rodzaju badana jest niezależnie od kategorii przypadka i liczby — oznacza to zachowanie rodzajów liczby pojedynczej także dla form liczby mnogiej.

Łączna analiza oznaczeń przypadka i liczby:

Oznaczenie SAM-a	**Kod nowej taksonomii**	**Wynik analizy**
N	N	mianownik (forma deprecjatywna)
H	N	mianownik (forma niedeprecjatywna)
G	G	dopełniacz
G'	G	dopełniacz (forma wariantowa)
D	D	celownik
T	A	biernik
I	I	narzędnik
L	L	miejscownik
V	V	wołacz
l	n. d.	zmiana liczby na mnogą

Nowa taksonomia nie wprowadza dodatkowej litery na oznaczenie kategorii deprecjatywności — informacja pochodząca z analizy automatycznej zostanie niewykorzystana. Interpretowane sa natomiast niejawne wyniki analizy rzeczowników, uzyskiwane na drodze uwzględnienia synkretyzmów:

- dla leksemów o odmianie nijakiej biernik i wołacz obu liczb jest równy mianownikowi (*dziecko*, *dzieci*),
- dla leksemów o odmianie żeńskiej
  - miejscownik liczby pojedynczej jest równy celownikowi (*zatoce*),
  - biernik i wołacz liczby mnogiej jest równy mianownikowi (*kule*),
- dla leksemów o odmianie męskiej
  - dla rodzaju męskoosobowego i męskozwierzęcego biernik liczby pojedynczej jest równy dopełniaczowi (*kowala*, *konia*),
  - dla rodzaju męskorzeczowego biernik liczby pojedynczej jest równy mianownikowi (*kołnierz*),
  - dla rodzaju męskoosobowego biernik liczby mnogiej jest równy dopełniaczowi (*kowali*, *królów*),
  - dla rodzaju męskozwierzęcego i męskorzeczowego biernik liczby mnogiej jest równy mianownikowi (*ptaki*, *stołki*),

– dla wszystkich rodzajów męskich wołacz liczby mnogiej jest równy mianownikowi (*kowale*, *konie*, *kołnierze*).

Pewną niejednoznaczność stwarza istnienie form o odmianie niezgodnej z wzorcem przypisanej formie kategorii (np. form rodzaju męskiego o odmianie żeńskiej — *artysta*, *wykładowca* itp.) Klasyfikuję je wg kategorii odmiany (podanym formom przypisywane jest oznaczenie rodzaju żeńskiego).

C.2 Formy przymiotnikowe

Klasyfikacja form przymiotnikowych następuje według tabeli pochodzącej z indeksu a tergo J. Tokarskiego [7] i wykorzystanej także w opisie analizatora SAM-99 [6]. Poniżej przytaczam prosty schemat przekodowań oznaczeń analizatora na kody nowego zestawu opracowany na jej podstawie (dla uproszczenia podano oznaczenia czterech pierwszych kategorii – część mowy, liczba, przypadek, rodzaj – uzupełniane o uzyskiwane niezależnie oznaczenie stopnia i przypisanie wartości nieokreślonych pozostałym kategoriom):

Oznaczenie SAM-a	**Kod nowej taksonomii**
1, com1	ASNP/ASNA/ASNI/ASAI
2, com2	ASGP/ASGA/ASGI/ASGN/ASAP/ASAI
3, com3	ASDP/ASDA/ASDI/ASDN
4, com4	ASIP/ASIA/ASII/ASIN/ASLP/ASLA/ASLI/ASLN/APDO/APDR
5, com5	ASNN/ASAN/APNO/APNR/APAR
6, com6	ASNF
7, com7	ASGF/ASDF/ASLF
8, com8	ASTF/ASIF
9, com9	APNO
10, com10	APGO/APGR/APAO/APLO/APLR
11, com11	APIO/APIR

Powyższe oznaczenia mogą być opatrzone dodatkowymi kwalifikatorami ', '', + oraz literą V, pochodzącymi w prostej linii z pracy J. Tokarskiego [?] i identyfikującymi formy wyjątków. Jako nieistotne w analizie form słownika są one zaniedbywane w dalszym ciągu przetwarzania.

Oznaczenie stopnia przymiotników i przysłówków:

- S (najwyższy) — analiza formy zawiera oznaczenie stopnia wyższego (przedrostek com), zaś sama forma rozpoczyna się członem *naj*,

- `C` (wyższy) — analiza zawiera przedrostek `com`, zaś forma nie rozpoczyna się członem *naj*,
- `P` (równy) — brak oznaczenia `com`.

Uwaga: Imiesłowy przymiotnikowe są w pierwszym etapie przetwarzania zaliczane do form czasownikowych (taki opis przewiduje też nowy zestaw kodowy), zostaną więc omówione niżej, w rozdziale C.3.

C.3 Formy czasownikowe

Pierwszą czynnością jest wyodrębnienie informacji o aspekcie formy z opisu analizy:

Oznaczenie SAM-a	**Kod nowej taksonomii**	**Aspekt czasownika**
`dk`	`D`	dokonany
`ndk`	`N`	niedokonany

Następnie zostaje dokonane sprawdzenie, czy forma jest specjalną, nieosobową formą czasownika:

Oznaczenie SAM-a	**Kod nowej taksonomii**	**Rodzaj nieosobowej formy czasownika**
`B`	`I`	bezokolicznik
`b`	`N`	bezosobnik (formy zakończone na -no, -to)
`u`	`U`	imiesłów przysłówkowy uprzedni
`w`	`W`	imiesłów przysłówkowy współczesny

Analiza oznaczeń rodzaju, liczby, osoby i trybu form osobowych, zapisanych w dość złożony sposób, odbywa się poprzez bezpośrednie sprawdzenie opisu analizy (seria wyrażeń warunkowych). Przytaczam skrót kodów wynikowych, składający się z zapisu kodu części mowy, liczby, rodzaju, osoby, czasu i trybu; skrypt wzbogaca go ponadto o ustalone wcześniej oznaczenie aspektu oraz symbole nieokreśloności dla pozostałych kategorii. Z racji mnogości form podaję odpowiednie przykłady.

Opis nie obejmuje nie rozpoznawanych przez analizator (używanych niezmiernie rzadko, lecz poprawnych) form rodzaju nijakiego liczby pojedynczej w pierwszej i drugiej osobie trybu oznajmującego (*czytałom*, *czytałoś*) i przypuszczającego (*czytałobym*, *czytałobyś*).

Uwaga: Formy szkolnego „czasu przyszłego prostego” (*przeczytam*, *przeczytałoby*) należą do paradygmatu czasownika dokonanego *przeczytać*, a nie do analizowanego niedokonanego *czytać* — stąd ich brak w powyższym opisie.

Analiza form czasu przyszłego złożonego nie jest możliwa bez znajomości tekstu — w plikach słownikowych formy takie są oznaczane dodatkowo.

Oznaczenie SAM-a	Kod nowej taksonomii	Kategoria osobowej formy czasownika	Przykład leksemu
1	VS-1TO	tryb oznajmujący, czas teraźniejszy	czytam
2	VS-2TO		czytasz
3	VS-3TO		czyta
l1	VP-1TO		czytamy
l2	VP-2TO		czytacie
l3	VP-3TO		czytają
m1	VSM1PO	tryb oznajmujący, czas przeszły	czytałem
ż1	VSF1PO		czytałam
mo1	VPO1PO		czytaliśmy
rz1	VPR1PO		czytałyśmy
m2	VSM2PO		czytałeś
ż2	VSF2PO		czytałaś
mo2	VPO2PO		czytaliście
rz2	VPR2PO		czytałyście
m	VSM3PO		czytał
ż	VSF3PO		czytała
n	VSN3PO		czytało
mo	VPO3PO		czytali
rz	VPR3PO		czytały
mC1	VSM1-P	tryb przypuszczający	czytałbym
żC1	VSF1-P		czytałabym
mC2	VSM2-P		czytałbyś
żC2	VSF2-P		czytałabyś
mC	VSM3-P		czytałby
żC	VSF3-P		czytałaby
nC	VSN3-P		czytałoby
moC1	VPO1-P		czytalibyśmy
rzC1	VPR1-P		czytałybyśmy
moC2	VPO2-P		czytalibyście
rzC2	VPR2-P		czytałybyście
moC	VPO3-P		czytaliby
rzC	VPR3-P		czytałyby
i	VS-2-R	tryb rozkazujący	czytaj!
i1	VP-1-R		czytajmy!
i2	VP-2-R		czytajcie!

Formy imiesłowów przymiotnikowych identyfikowane są poprzez zbadanie początkowej części oznaczenia

Oznaczenie SAM-a	Kod nowej taksonomii	Rodzaj imiesłowu przymiotnikowego
w	I	czynny (współczesny)
A	N	bierny
ł	U	przeszły

po której następuje łączny identyfikator liczbowy kategorii przypadka, liczby i rodzaju formy z zakresu 1-11 opisany w tabeli w rozdziale dokumentującym interpretacje form przymiotnikowych C.2.

Uwaga: W przypadku imiesłowów przymiotnikowych czynnych analiza jest szczątkowa, gdyż analizator rozpoznaje jedynie formy mianownika obu liczb (oznaczenia 1 i 9).

C.4 Formy liczebnikowe

Pierwszą czynnością wykonywaną przy rozpoznawaniu analizy liczebników jest ustalenie ich liczby na podstawie formy hasłowej. Dla formy *jeden* przyjowane jest oznaczenie liczby pojedynczej, dla pozostałych form — liczby mnogiej.

Analizator dzieli liczebniki na jedną z czterech kategorii (Ka – Kd). Każda z nich jest badana osobno:

- *Ka* — charakterystyka gramatyczna uzyskiwana jest na podstawie tabeli zamieszczonej w opisie analizatora [6]; oto schemat przekodowania form (podaję oznaczenia części mowy, liczby, przypadka i rodzaju — pozostałe kategorie nie są określone[10]):

Oznaczenie SAM-a	Kody nowej taksonomii
1	NPNR/NPAR
2	NPNO/NPGO/NPGR/NPDO/NPDR/ NPTO/NPIO/NPIR/NPLO/NPLR
3	NPIO/NPIR

[10] Wszystkie oznaczenia w tabeli dotyczą liczby mnogiej, określonej dla większości form liczebnikowych. Z tego względu rodzaj gramatyczny ograniczony został do męsko- i niemęskoosobowego; dla liczby pojedynczej formy rodzajów niemęskoosobowych musiałyby zostać opisane dokładniej.

- *Kb* — formy nie są przetwarzane z powodu braku dodatkowych informacji analizatora (napis `por.`),
- *Kc* — formy liczebników zbiorowych tworzone są na podstawie podawanego przez analizator oznaczenia przypadka (NGDAIL[11]), dołączanego do oznaczeń liczebnika i liczby (mnogiej),
- *Kd* — formom liczebników ułamkowych oprócz oznaczenia części mowy nadawana jest tylko kategoria rodzaju (`m`, `ż`, `n`), uzyskiwana bezpośrednio z analizy.

C.5 Formy zaimków

Analizator opisuje zaimki dzieląc je na trzy kategorie:

- `Za` — zaimki o jednej kategorii fleksyjnej (*kto*, *my*), którym przypisywana jest tylko kategoria przypadka,
- `Zb` — zaimki o dwóch kategoriach fleksyjnych (*ja*, *ty*, *się*), identyfikowane oznaczeniem formy przypadka i kategorii akcentowości (formy akcentowane oznaczane znakiem `T`, nieakcentowane — `N`),
- `Zc` — zaimek *on* (o pięciu kategoriach fleksyjnych), o nieregularnej odmianie, opisany specjalną tabelą; z powodu potencjalnie dużego nakładu pracy związanego z jego automatyczną analizą został on opisany ręcznie.

C.6 Formy przysłówkowe

Analizator przypisuje przysłówkom jedynie oznaczenie stopnia wyższego — podobnie jak w przypadku przymiotników (patrz Rozdział C.2) nie wyróżniono specjalnie form stopnia najwyższego. Są one identyfikowane po zbadaniu początkowej części formy: stopień najwyższy przyjmują formy rozpoczynające się przedrostkiem *naj*.

[11] Oznaczenie biernika podawane przez analizator (`T`) jest różne od nowego oznaczenia kodowego (`A`).

C.7 Nieodmienne części mowy

Pozostałe części mowy (partykuły, wykrzykniki, przyimki), zapewne z racji trudności klasyfikacyjnych, nie są przez analizator rozpoznawane (przypisywane jest im oznaczenie *form nieodmiennych*. Tymczasowo jest im przypisywany specjalny kod `Y`, pozostawiany do ręcznego oznaczenia kategorii gramatycznej, zaś dla form posiadających klasyfikację słownikową części mowy została ona zachowana.

Literatura cytowana

[1] 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`.

[2] Głowińska, Katarzyna. *Taksonomia morfologiczna dla Słownika frekwencyjnego*. W: [1], `Dokumentacja\taksonomia.pdf`.

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

[4] 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.

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

[6] Szafran, Krzysztof. *Analizator morfologiczny SAM-99 — opis użytkowy*. Maj 2000. Instytut Informatyki Uniwersytetu Warszawskiego.

[7] Tokarski, J., Schematyczny indeks a tergo polskich form wyrazowych. Opracowanie i redakcja: Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa 1993.

[8] Wall, Larry; Schwartz, Randal L. *Programming Perl*. 1991. O'Reilly & Associates, Inc.

[9] Wall, Larry; Christiansen, Tom; Potter, Stephen; Schwartz, Randal L. *Perl. Programowanie*. ReadMe, Warszawa, 1998.