Elżbieta Hajnicz i Anna Kupść

Przegląd analizatorów morfologicznych dla języka polskiego

Nr 937

Warszawa, grudzień 2001

Streszczenie

Niniejsze opracowanie zawiera przegląd sześciu analizatorów morfologicznych języka polskiego: Gram, PoMor, SAM, LEM, XeLDA, AMOR. Raport rozpoczyna się od prezentacji testów opracowanych w celu porównania omawianych analizatorów oraz współczynników służących do oceny statystycznej skuteczności ich działania. W następnych rozdziałach po kolei omawiane są poszczególne analizatory wraz z efektami ich działania na przygotowanych testach. Porównanie wyników działania analizatorów przedstawione zostało w podsumowaniu.

Słowa kluczowe: analizator morfologiczny, etykiety części mowy, dokładność analizatora, formy podstawowe

Abstract

A survey of morphological analysers for the Polish language

This report contains a survey of six morphological analysers for the polish language: Gram, PoMor, SAM, LEM, XeLDA, AMOR. The report begins with a presentation of tests we adopted for a comparison of the analysers. We also introduce definitions of factors which serve for statistical evaluation of the tests. Subsequent sections present particular analysers and discuss test results. The summary provides a comparison of the tested analysers.

Keywords: morphological analyser, PoS tags, precision, base forms

1 Wstęp

Celem niniejszego raportu jest przedstawienie oraz porównanie działania niektórych analizatorów morfologicznych języka polskiego. Przegląd ten został dokonany celem wyboru analizatora najlepiej przygotowanego do wykorzystania w automatycznej anotacji morfosyntaktycznej dużych zbiorów tekstowych polszczyzny pisanej (projekt KBN nr 7 T11C 043 20).

Wstępny etap analizy morfologicznej (bądź też całego procesu anotacji tekstu) stanowi podział tekstu na słowa, i już tutaj mogą pojawić się pewne problemy, związane z doborem znaków mających stanowić separatory słów. Narzucającym się w sposób oczywisty założeniem jest uznanie spacji za bezwzględny separator słów. Ale już karetka (znak końca linii) może sprawiać pewne trudności w przypadku słów przenoszonych do nowej linii, choć jest oczywiste, że w większości wypadków spełnia ona tę funkcję. Kolejny problem stanowią znaki przestankowe. Na przykład potraktowanie łącznika jako separatora powoduje, że formy typu *kujawsko-pomorskie* rozbijane są na dwie składowe, co może utrudniać ich dalszą analizę. Jednak rezygnacja z traktowania znaków przestankowych jako separatorów słów uniemożliwia analizę w przypadku braku spacji w tekście (*kandydatów.Łatwiej*). Banalnym przykładem wskazującym, że podział tekstu na słowa nie jest oczywisty, są edytory tekstowe wyposażone w funkcje *przenoszenia kursora do następnego słowa* i/lub *kasowania słów*: różne edytory w odmienny sposób realizują te polecenia.

Za wyrazy uznajemy słowa należące do konkretnego języka (np. polskiego). Bień (2001) definiuje pojęcie wyrazu morfologicznego jako abstrakcyjną konstrukcję służącą do klasyfikacji wyrazów ze względu na ich kształt. Tak więc analiza morfologiczna powinna opisywać formę wyrazową w taki sposób, by jednoznacznie identyfikować jej kształt. Na przykład, skoro biernik dla rodzaju męskiego l.poj. nigdy nie tworzy odrębnej formy, lecz jest tożsamy albo z mianownikiem (rodz. męskorzeczowy), albo z dopełniaczem (rodz. męskoosobowy i męskożywotny), uwzględnianie go w opisie formy można uznać za nadmiarowe. Natomiast wyraz, któremu w sposób jednoznaczny przypisany zostanie zestaw *indykatorów morfologicznych* (tj., część mowy, wartości cech fleksyjnych, np. osoba, rodzaj, liczba, przypadek itd.), nazwany został przez Bienia *wyrazem morfosyntaktycznym*, gdyż indykatory takie opisują w istocie funkcję syntaktyczną danego wyrazu.

Analiza morfologiczna operująca na poszczególnych wyrazach w izolacji od kontekstu nie może rzecz jasna przypisać poszczególnym formom jednoznacznych zestawów indykatorów; istnieje jednak możliwość przypisania im wszystkich, jakie mogą charakteryzować daną formę we wszystkich możliwych kontekstach. W takim przypadku możemy mówić o analizie morfosyntaktycznej.

Reasumując, zadaniem analizatora morfologicznego jest automatyczne określenie cech morfosyntaktycznych danego wyrazu na podstawie jego kształtu graficznego oraz podanie jego maksymalnie wyczerpującej analizy. Chodzi tu o rozpoznanie wszystkich możliwych leksemów reprezentowanych przez dane słowo oraz podanie wszystkich interpretacji gramatycznych w obrębie każdego rozpoznanego leksemu. Przykładowo, słowo *klucz* może być formą rzeczownika *klucz* (liczba pojedyncza, rodzaj męski, mianownik, biernik lub wołacz), ale może to być także forma czasownika *kluczyć* (tryb rozkazujący, druga osoba, liczba pojedyncza). Zadaniem analizatora morfologicznego byłoby zatem automatyczne rozpoznanie dwóch leksemów (rzeczownik *klucz* oraz czasownik *kluczyć*) oraz podanie odpowiednich informacji dotyczących tych leksemów (trzy interpretacje dla rzeczownika: mianownik, biernik lub wołacz liczby pojedynczej oraz jednej interpretacji dla czasownika: rozkaźnik).

Jednak wiele tekstów zawiera, poza wyrazami danego języka (do których zgodnie z powyższą definicją należy zaliczyć także skróty), również wiele innych rodzajów słów, takich jak znaki przestankowe i specjalne jako takie, liczby, daty, godziny, słowa obce, adresy elektroniczne i inne. Rozpoznawanie takich słów przez analizator w maksymalnym stopniu zdecydowanie ułatwia dalszy proces anotacji tekstu. Jednak zadanie identyfikacji takich napisów uniemożliwia traktowanie znaków przestankowych, np. kropki (*24.12.2001*), przecinka (*7,104*) i dwukropka (*13:15*) jako znaków przestanowych.[1] Natomiast potraktowanie niektórych znaków jako warunkowych separatorów słów zwiększa już nie tyle liczbę analiz, co samych podziałów tekstu na słowa (przy założeniu, że kropka jest warunkowym separatorem słów, napis *24.12.2001* może zostać podzielony na słowa na cztery sposoby).

Powyższe rozważania wyraźnie wskazują, jak wiele założeń wstępnych musi zostać przyjętych przy konstruowaniu każdego analizatora morfologicznego. Zdając sobie sprawę z tego faktu, jako podstawowe kryterium oceny działania analizatorów przyjęłyśmy ich przydatność w procesie automatycznej anotacji korpusu. Tak więc skoncentrowałyśmy się na lingwistycznej poprawności i pełności algorytmów analizujących, dużą wagę przykładając także do zakresu rozpoznawania przez nie pozajęzykowych składników tekstu, mniejszą wagę przywiązując do ich efektywności.

Z wielu narzędzi do analizy morfologicznej dostępnych dla języka polskiego, w raporcie dokładniej omówione zostało sześć: SAM, LEM (firmy LEX), Gram (firmy Neurosoft), AMOR, PoMor (firmy MorphoLogic) oraz

[1] Dr Wołosz zwrócił nam uwagę, że niektórzy piszą daty ze spacjami, np. *24. 12. 2001*, co mogłoby stawiać pod znakiem zapytania nawet uznanie spacji za bezwzględny separator słów.

XeLDA (firmy Xerox). Wybór ten został uwarunkowany czynnikami technicznymi, takimi jak dostępność analizatora, stopień jego funkcjonalności itd.

Raport składa się z następujących rozdziałów. Rozdział 2 zawiera krótkie omówienie przeprowadzonych testów lingwistycznych oraz współczynników określających poprawność analizatorów. Rozdział 3 zawiera listę narzędzi do analizy morfologicznej dostępnych dla polskiego, na podstawie której dokonano wyboru narzędzi do prezentacji. Rozdziały 4–9 zawierają opisy poszczególnych analizatorów. Dodatek A zawiera porównanie działania analizatorów na podstawie testów przeprowadzonych na listach słów, dodatek B zawiera pliki tekstowe, na których przeprowadzano testy, zaś w dodatku C przedstawione są próbki analizy generowanej przez poszczególne analizatory (dla początkowego fragmentu tekstu z pliku PUBLICYSTYKA).

Raport jest częściowo finansowany przez KBN w ramach grantu nr 7 T11C 043 20.

Autorki chciałyby w tym miejscu podziekować twórcom poszczególnych analizatorów, to jest dr. Krzysztofowi Szafranowi, dr. Robertowi Wołoszowi, Cezaremu Dołędze, Joannie Rabiedze, a także prof. Januszowi Bieniowi i dr Dorocie Kopcińskiej za wiele cennych uwag dotyczących wstępnych wersji raportu.

2 Testy

Przeprowadzono dwa rodzaje testów lingwistycznych: na listach słów (słowa poprawne i niepoprawne) oraz na czterech krótkich tekstach pochodzących z wydania dziennika „Rzeczpospolita" z dnia 27 czerwca 2001. Teksty dotyczyły wiadomości krajowych (KRAJ), kulturalnych (KULTURA), doniesień ze świata (ŚWIAT), a jeden tekst stanowił reportaż (PUBLICYSTYKA). Konkretne pliki tekstowe są umieszczone w dodatku B.

2.1 Listy słów

Punktem wyjścia do stworzenia list słów poprawnych i niepoprawnych była praca McShane (2001), lecz w dużej mierze listy te tworzone były nieco przypadkowo, zgodnie z intuicjami autorek. W szczególności, nie zostały uwzględnione wszystkie koniugacje, jedynie III, IV, V, VI, VII, IX, X, XI, XII (zgodnie z systematyzacją Tokarskiego (1993)). Także imiesłowy — przymiotnikowe i przysłówkowe — są słabo przebadane. Tabelaryczne zestawienie wyników działania poszczególnych analizatorów na listach słów jest przedstawione w dodatku A.

Liczba słów umieszczonych na tych listach, pogrupowanych według przynależności do poszczególnych części mowy oraz wszystkich słów łącznie, jest podana w tabeli na str. 4.

Konstruując listy słów świadomie wprowadziłyśmy do nich pewne wyrazy, których obecność w tego typu teście może budzić kontrowersje. Chodzi o wyrazy lub formy archaiczne, takie jak *mełty, awaryj, stacyj, książęciu.* Dotyczy to także form powszechnie używanego obecnie czasownika *mielić*, przez tradycyjne słowniki uznawanych za niepoprawne, lecz akceptowanych zarówno w Markowski (1999), jak i Podracki (2001). Z drugiej strony wiele rzadkich form pojawiło się w tekście przez przypadek, jako homonimy wyrazów, których przebadanie było naszą rzeczywistą intencją. Szczególną konfuzję mogły spowodować formy *źrebiu, lasi, żołądzie*, które w założeniu miały sprawdzać niepoprawne odmiany wyrazów *źrebię, las, żołądź*, a które zostały rozpoznane jako formy rzadkich wyrazów *źreb, laszy, żołąd* przez PoMor'a. Oczywiście ani brak, ani wystąpienie takich form nie może być uważane za błąd czy też nawet słabość danego analizatora.

W poniższej tabeli przedstawione są dane liczbowe dotyczące plików poddanych analizie, określające liczbę wyrazów należących do poszczególnych kategorii gramatycznych (części mowy) w każdym z rozważanych plików. W tabeli tej uwzględniono zarówno listy słów, jak i teksty wybrane do analizy. Kategoria *niejednoznaczne* utworzona dla listy słów poprawnych dotyczy przypadków homonimii morfologicznej, gdy odpowiednie formy podstawowe analizowanych wyrazów należą do różnych części mowy. Jeśli takie zjawisko miało miejsce w analizowanych plikach tekstowych, wyrazowi przypisano kategorię gramatyczną zgodną z kontekstem, w którym wystąpiło rozważane słowo. Ponadto uwzględniono jedynie typowe skróty, takie jak *mld, tzw.*, skróty nazw własnych (*USA, NATO, FSC, MSWiA*) zaliczając do nazw własnych. Natomiast przymiotniki utworzone od nazw własnych (*polski, lubelski*) nie zostały w żaden szczególny sposób wyróżnione, i zaliczono je do przymiotników. Założenie to wydaje się zgodne z podejściem twórców niektórych analizatorów, umieszczających w słownikach szerszą reprezentację takich przymiotników niż nazw własnych, od których one pochodzą.

	poprawne	niepoprawne	kraj	świat	kultura	publicystyka	RAZEM
czasowniki	72	36	38	38	13	112	201
rzeczowniki	126	65	131	78	24	221	454
gerundia	1	0	4	7	1	2	39
zaimki rzeczowne	5	0	11	8	3	29	51
przymiotniki	40	21	38	13	11	95	157
imiesłowy przym.czyn.	0	0	1	3	0	3	7
imiesłowy przym.bier.	2	0	2	10	4	12	28
zaimki przymiotne	0	0	9	4	4	17	34
przysłówki	19	2	21	3	2	27	53
imiesłowy przysł.współcz.	0	0	0	0	0	1	1
imiesłowy przysł.uprz.	1	0	0	0	0	0	0
zaimki przysłowne	0	0	1	2	0	9	12
liczebniki główne	4	0	11	3	2	11	27
liczebniki porządkowe	0	0	0	0	0	4	4
liczebniki zbiorowe	2	0	0	0	0	1	1
liczby	4	0	21	4	1	12	38
przyimki	0	0	72	40	18	104	230
spójniki i partykuły	4	0	28	21	6	72	127
nazwy własne	16	0	20	32	17	54	123
przymiotniki własne	0	0	17	5	2	4	28
skróty	5	0	0	2	0	7	9
niejednoznaczne	83	0	–	–	–	–	–
inne	58	0	–	–	–	–	–
odmienne	340	–	304	202	83	615	1204
RAZEM	398	124	425	273	108	820	1626

2.2 Statystyczna analiza tekstów

Dobór wartości i współczynników statystycznych mających służyć do oceny działania analizatorów morfologicznych na plikach tekstowych nie jest zadaniem łatwym. Generowane przez dane narzędzie analizy oceniane są na dwóch płaszczyznach: uwzględnienie poprawnej analizy wyrazu zgodnie z kontekstem jej wystąpienia oraz niezależnie od takiego kontekstu. W tym drugim przypadku kontroli podlegają wszystkie zaprezentowane analizy, a w szczególności formy podstawowe wyrazów, względem których analiza została przeprowadzona. Ich dobór i

liczba są w zasadniczy sposób zależne od bogactwa słownika związanego z danym analizatorem. Jednak bogactwo słownika może stanowić zarówno zaletę, jak i wadę analizatora. Z jednej strony większa liczba słów gwarantuje, że większa liczba słów zostanie poprawnie rozpatrzona, ale powoduje także zwiększenie niejednoznaczności przeprowadzanej analizy, co utrudnia proces dezambiguacji podczas przeprowadzania anotacji tekstów. Zdarza się, że słownik zawiera wiele słów rzadkich, archaicznych bądź gwarowych, lecz brak w nim nowych słów, pojawiających się z dużą częstotliwością we współczesnym języku. Analiza statystyczna nie jest w stanie tego uchwycić.

Ostatecznie zdecydowałyśmy się na wybór następujących wartości, naszym zdaniem najlepiej charakteryzujących proces analizy tekstów:

- $|S|$ — liczba słów w tekście,
- $|S_F|$ — liczba słów odmiennych,
- unknown(S) — liczba nierozpoznanych słów, tj. liczba słów, dla których analizator nie podał żadnej analizy,
- errors(S) — liczba słów bez poprawnej analizy w kontekście (łącznie z nierozpoznanymi słowami),
- generated — liczba wygenerowanych form podstawowych,
- F-generated — liczba wygenerowanych form podstawowych dla słów odmiennych,
- less — liczba form podstawowych zanalizowanych bez podania wszystkich form (bez unknown),
- non-exist — liczba nieadekwatnych form podstawowych (w tym nieistniejących),
- more — liczba form podstawowych zanalizowanych z podaniem zbyt wielu form (w tym non-exist),
- correct — liczba wszystkich poprawnie zanalizowanych form podstawowych, niezależnie od kontekstu,
- F-correct — liczba wszystkich poprawnie zanalizowanych form podstawowych dla słów odmiennych.

Wyróżnienie słów odmiennych wynika z faktu, że dla wyrazów nieodmiennych (takich jak spójniki czy przyimki) analizatory działają zazwyczaj prawidłowo: jest to niewielka liczba słów, które łatwo jest umieścić w słowniku, a wtedy analiza odbywa się poprawnie. Z drugiej strony słowa te występują w tekście z dużą częstotliwością, co może mieć wpływ na wyniki analizy statystycznej tekstu. Jeśli chodzi o analizę niezależną od kontekstu, każdy analizator traktowany był osobno, i oceniano jedynie formy wygenerowane przez dany program. Brak pewnej formy podstwowej związanej z danym słowem może być spowodowany dwoma czynnikami: błędami algortymu lub niewystępowaniem danego słowa w słowniku, jednak czynniki te nie były sprawdzane.

Powyższe wartości posłużyły do określenia następujących współczynników: pomiar błędu (c-error-rate), procent form podstawowych bez podania wszystkich analiz (less-error-rate), procent form podstawowych z podaniem zbyt wielu analiz (more-error-rate), dokładność (precision), dokładność dla słów odmiennych (F-precision) oraz średnia ilość **poprawnych** form podstawowych zaproponowanych przez analizator dla danego (przeanalizowanego) slowa (average-#correct-analyses). Sposób określania poszczególnych współczynników jest podany poniżej.

- c-error-rate = errors(S)/$|S|$,
- less-error-rate = less/F-generated,
- more-error-rate = more/F-generated,
- precision = correct/generated,
- F-precision = F-correct/F-generated,
- average-#correct-analyses = correct/($|S|$ − unknown).

Wartości współczynników dla testowanych analizatorów zostały podane w rozdziałach opisujących poszczególne analizatory. Uproszczone analizy przeprowadzono także dla list słów.

W przypadku plików tekstowych, podstawę do obliczeń stanowiła za każdym razem liczba form podstawowych wygenerowanych przez dany analizator, natomiast w przypadku listy słów punkt odniesienia stanowiły wszystkie leksemy wygenerowane przez którykolwiek z analizatorów.

Pliki tekstowe wybrane zostały w sposób przypadkowy, natomiast listy słów z założenia miały zawierać wyrazy mogące sprawiać trudność w procesie analizy. Ponieważ próbka ta nie była konstruowana na podstawie jakichkolwiek własności frekwencyjnych badanych wyrazów, wyniki statystyczne mogą być obciążone sporym błędem i powinny być traktowane jedynie jako wskazówka przy ocenie tych konkretnych testów.

3 Analizatory morfologiczne

Poniższa lista przedstawia w miarę pełen zestaw istniejących narzędzi do analizy morfologicznej języka polskiego, wraz z podaniem adresu kontaktowego i/lub sposobu udostępniania narzędzi.

- Gram, analizator firmy Neurosoft, Wrocław; Cezary Dołęga, e-mail: `cezar@neurosoft.pl`, strona internetowa `http://www.neurosoft.pl`; w szczególności wersja demo dostępna pod adresem `http://www.neurosoft.pl/gram/demo`
- PoMor, analizator firmy MorphoLogic, Węgry; do celów badawczych firma umożliwiła nam nieodpłatny dostęp do analizatora; informacje można uzyskać na stronie `www.morphologic.hu` oraz od dra Roberta Wołosza, e-mail: `wolosz@btk.jpte.hu`
- SAM, analizator dra Krzysztofa Szafrana; dostępny bezpłatnie do celów niekomercyjnych pod adresem: `http://www.mimuw.edu.pl/~kszafran/SAM-dists/`
- LEM, produkt firmy LEX, Poznań; prof. Zygmunt Vetulani, e-mail: `vetulani@amu.edu.pl`
- XeLDA, produkt firmy XEROX: tokenizer i analizator morfologiczny; wersja demo dostępna pod adresem: `http://www.xrce.xerox.com/research/mltt/demos/polish`, dodatkowe infromacje można znaleźć pod adresem `http://www.xrce.xerox.com/ats/xelda`
- narzędzia dra Jana Daciuka do analizy morfologicznej dla niemieckiego, francuskiego, angielskiego i polskiego; dostępne bezpłatnie do celów niekomercyjnych pod adresem: `http://www.pg.gda.pl/~jandac/fsa.html`
- analizator firmy TiP, działający m.in. w przeglądarce na serwerze `www.onet.pl`; ogólne informacje dostępne pod adresem: `http://szukaj.onet.pl/pomoc/slownik.html`
- baza fleksyjna języka polskiego, Kraków, AGH; prof. Wiesław Lubaszewski: `http://www.icsr.agh.edu.pl/fleksbaz/`
- Filip Graliński, Poznań, e-mail: `filipos@venus.wmid.amu.edu.pl`
- analizator AMOR, Joanna Rabiega, e-mail: `jrabiega@priv7.onet.pl` oraz Michał Rudolf, e-mail: `mrudolf@mercury.ci.uw.edu.pl`
- baza fleksyjna języka polskiego; Marjorie McShane, e-mail: `marge@crl.nmsu.edu`
- POLLEX, analizator prof. Krzysztofa Bogackiego, opracowany w ramach projektu BILEDITA (więcej informacji w pracy Ch. Bogacki, „POLLEX — un dictionnaire électronique morphologique du polonais", *Bulletin de Linguistique Appliquée et Générale, Université de Franche Comté*, Numéro Spécial Actes FRACTAL '97, pp. 55–63)

Po przeprowadzeniu wstępnego rozeznania, lista narzędzi do bezpośredniego lub pośredniego przetestowania okazała się znacznie mniejsza. Część wymienionych wyżej narzędzi znajduje się nadal w przygotowaniu (analizator Filipa Gralińskiego; narzędzia Jana Daciuka nie zawierają jeszcze słownika dla języka polskiego i jest on obecnie doraźnie tworzony przez autora), niektóre narzędzia nie są dostępne (baza fleksyjna Marjorie McShane) lub nie spełniają wymogów narzucanych przez przewidziane przez nas zastosowanie (firma TiP dysponuje bardzo skutecznym narzędziem do sprawdzania pisowni, lecz nie jest to analizator morfologiczny: formy jednego leksemu mogą być zanalizowane jako formy różnych leksemów). Nie doszło do bliższych ustaleń co do udostępnienia do testów bazy fleksyjnej z AGH. Informacja o analizatorze prof. Bogackiego dotarła do nas zbyt późno.

W rezultacie, przeprowadzono testy dla następujących sześciu analizatorów: Gram, SAM, LEM, PoMor, XeLDA oraz AMOR.

4 Gram

Analizator w testowanej wersji (Gram 2.2) ma architekturę typu klient-serwer, jest dostępny w wersji dla Windows NT 4.0/2000 (DCOM) lub Linuxa (demon TCP/IP). Możliwa jest bezpośrednia integracja z aplikacjami WWW, dostępny w postaci bibliotek DLL. Program odczytuje litery polskie w różnych kodach, w szczególności posiada opcję automatycznego wykrywania strony kodowej tekstu.

Ponieważ miałyśmy dostęp jedynie do uproszczonej wersji Gram'a działającego w serwisie WWW, który w elegancki graficzny, jednak uproszczony sposób podaje informacje na temat przeanalizowanych form wyrazowych.

Poniżej podajemy informacje dotyczące zarówno tej uproszczonej, jak i pełnej wersji programu; jednak wszelkie wyniki testów dotyczą wyłącznie tego pierwszego przypadku. Tak więc jestesmy jedynie w stanie stwierdzic, ze analizator odrzuca poprawną formę lub akceptuje błędną, bez możliwości zbadania poprawności przypisywanych opisów morfosyntaktycznych.

4.1 Oznaczenia PoS

W testowanej wersji analizatora stosowane są tylko bardzo ogólne oznaczenia, określające części mowy (ang. PoS, part of speech), przy użyciu następujących etykiet: rzeczownik, czasownik, przymiotnik, przysłówek, przyimek, zaimek, liczebnik, imiesłów, inny. Dodatkowe informacje, takie jak przypadek, liczba czy rodzaj, nie są podane jawnie. Istnieje jednak możliwość odkodowania ich w wersji analizatora bez interfejsu graficznego.

I tak, rzeczowniki i przymiotniki są etykietowane za pomocą rodzaju (dla rzeczowników wyłącznie *męski, żeński i nijaki*), liczby i przypadka; leksemom czasownikowym przypisywane są etykiety boolowskie *Osobowość, Aspekt, Przechodniość, Zwrotność*, a ich formy oznaczane są za pomocą osoby wraz z liczbą, czasu, trybu i formy (*osobowa, bezosobowa, bezokolicznik*). Zauważmy, że bezokoliczniki czasowników posiadających taką samą koniugację (np. *ugrząść – ugrzęznąć, wyląc – wylęgnąc, zaprząc – zaprzęgnąć*) traktowane są jako formy alternatywne tego samego leksemu. Ponadto wyróżniane jest wszystkie pięć rodzajów imiesłowów, zaimki dzielą się na podklasy *niezdefiniowany, nieoznaczony, rzeczownikowy, przymiotnikowy, przysłówkowy, liczebny, dzierżawczy, osobowy, pytajny, wskazujący, względny, zwrotny*, zaś wyrazy nieodmienne dzielą się na podklasy *niezdefiniowany, spójnik, partykuła, wykrzyknik, onomatopeja*.

Analizator współpracuje z kilkoma słownikami, m.in. nazw własnych, imion i nazwisk, co umożliwia rozpoznawanie niektórych nazwisk, imion czy nazw geograficznych. Grupy te wyróżniane są za pomocą odrębnych etykiet.

Stosowane są także oznaczenia do określenia innych rozpoznanych form: data, godzina, sekwencja, liczba, URL, skrót, exp (w wypadku rozpoznania liczb rzymskich, np. XIV), wulgaryzm.

4.1.1 Formy podstawowe

Dla każdej formy, analizator określa zarówno formę podstawową (forma odmienianego leksemu) oraz formę kanoniczną (forma, od której pochodzi dany leksem). Związek między formą kanoniczną i podstawową jest opcjonalny. W pełnej wersji Gram'a forma kanoniczna definiowana jest na podstawie tzw. *relacji pochodzenia*, umożliwiającej wiązanie przysłówka z odpowiednim przymiotnikiem, czasowników niedokonanych z ich dokonanymi odpowiednikami, odsłowników i imiesłowów z czasownikami (dokonanymi bądź nie), od których pochodzą, zdrobnień i zgrubień z odpowiednim rzeczownikiem/przymiotnikiem (Np.*grubiutki, grubachny* z *gruby*, itd. Zależność ta może być w różny sposób formułowana zgodnie z wolą użytkownika. Uwzględnienie takich pozafleksyjnych związków morfologicznych jest niewątpliwie istotną zaletą programu.

Poniżej przedstawiono wartości form podstawowych dla poszczególnych PoS w testowanej przez nas wersji programu. W wersji tej forma kanoniczna jest w zasadzie równa formie podstawowej, z dokładnością do informacji dodatkowych zawartych w pierwszej z nich.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów	przysłówek
forma kanoniczna	mianownik liczba poj.	bezokolicznik	ten sam stopień, co badanej formy mianownik, r. męski, liczba poj.	bezokolicznik	ten sam stopień, co badanej formy

4.2 Wyniki testów dla list słów

4.2.1 Analiza form poprawnych

Słownik języka polskiego współpracujący z analizatorem nie obejmuje wszystkich testowanych słów. Dotyczy to słów dość pospolitych, np. *pawiu* czy *codzienne* nie zostały w ogóle rozpoznane.

W kilku wypadkach nie zostały znalezione wszystkie możliwe interpretacje danej formy. Przykładowo, *wolno* jest rozpoznane tylko jako przysłówek, ale nie jako czasownik; słowo *pal* zostało rozpoznane jako forma czasownika *palić*, ale nie skojarzono jej z rzeczownikiem; podobnie dla *płaszcz* (rozkaźnik czasownika *płaszczyć*, ale nie rzeczownik), *szalej* (tylko *szaleć*) i *pieprz* (tylko *pieprzyć*), *klucz* (tylko *kluczyć*, nie rozpoznano formy rzeczownika). Ponieważ została znaleziona forma rzeczownika *klucz* przy analizie słowa *kluczy* (jak również czasownik *kluczyć*), fakt ten wskazuje na błąd w algorytmie znajdowania form podstawowych, a nie brak słowa *klucz* w słowniku. Inne słowa, dla których nie znaleziono wszystkich form podstawowych:

- *bez*: przyimek i rzeczownik *beza*, brak rzeczownika *bez*;
- *uczony*: przymiotnik i imiesłów, brak rzeczownika;
- *wieść*: czasownik *wieścić* i rzeczownik *wieść*, brak czasownika *wieść*;
- *wiedźmy*: rzeczownik *wiedźma*, brak czasownika *wieść*;
- *dziczej*: czasownik *dziczeć* i przymiotnik *dziczy*, lecz brak przysłówka/przymiotnika *dziko*/*dziki*;
- *bogaciej*: tylko przysłówek, brak czasownika *bogacieć*;
- *srożej*: tylko przysłówek, brak czasownika *srożeć*;
- *dwunogi*: tylko przymiotnik, brak rzeczownika *dwunóg*;
- *wypełzły*: tylko imiesłów przymiotnikowy, ale brak formy osobowej czasownika *wypełznąć* (czasownik jest w słowniku, bo jest podany jako forma kanoniczna imiesłowu);
- *rośnie*: czasownik *rosnąć* i (rzadko używana forma) *róść* (chociaż dla innych czasowników alternatywne formy bezokolicznika nie są podawane), brak natomiast archaicznego przymiotnika *rośny*;
- *gzowi*: przymiotnik, ale nie rzeczownik *giez*;
- *ćmi*: czasownik, ale nie przymiotnik.

Nierozpoznawanie form od czasownika *wieść* (*wiodła, wiódł, wiodę*) oraz brak form czasownikowych *wieść* i *wiedźmy* powyżej wskazują na brak tego czasownika w słowniku.

Analizator ma problemy z odmianą bardzo nieregularnych wyrazów takich jak: *pełznąć* (formy *pełźli, pełzli, pełznie* nie zostały rozpoznane, lecz *pełźnie* zostało rozpoznane jako forma *pełznąć*; w słowniku jest także czasownik *pełzać*, gdyż został podany jako forma podstawowa dla *pełza*), *wlec* (analizowane są formy rzadziej spotykane *wlekę, wlekł, wlekła*, lecz formy *wlokę, wlókł, wlokła* są nierozpoznane).

Nierozpoznanie form *mielić, mielił, mieli* jako form czasownika *mielić* wskazuje na brak tego czasownika w słowniku (dla *mieli* została podana tylko forma podstawowa *mieć*, zaś pozostałe wcale nie zostały rozpoznane). Natomiast formy czasownika *mleć* (*mleć, mełł, miele*) zostały poprawnie rozpoznane.

Rzadziej spotykane formy odmiany są na ogół odrzucane: *ząb* (analizowane *zęba, zębie*, lecz nie *zębu*), *dąb* (analizowane *dębu*, ale nie *dęba* [w *stanąć dęba*]), *łazęga* (rozpoznane tylko *łazęg*, ale nie *łazęgów*), *książę* (nierozpoznana forma *książęciu*), *wstęga* (akceptowane *wstęg*, ale nie *wstąg*).

4.2.2 Analiza form niepoprawnych

Zauważono błędy m.in. w odmianie następujących wyrazów:

- *koń*: *końmi* odrzucone, zaś niepoprawne *koniami* uznane za poprawne;
- *dureń*: poprawnie rozpoznane *durniów*, lecz nierozpoznane *durniu*; zaakceptowane niepoprawne *dureniu*;
- *ząb*: zaakceptowane niepoprawne formy *ząbie, ząby*;
- *ugrzęznąć*: poprawnie rozpoznane *ugrzązł*, forma *ugrzęzła* rozpoznana tylko jako imiesłów, zaś niepoprawne *ugrzązła* zaakceptowane jako forma czasownika;
- *żołądź*: *żołądź, żołędzi* poprawnie rozpoznane, ale zaakceptowane też niepoprawne *żołądzi, żołądzie*;
- *dom*: zaakceptowane niepoprawne *domie*;
- *jadł*: rozpoznane poprawnie jako forma czasownika *jeść* oraz niepoprawnie jako forma rzeczownika *jadło*;
- *gród*: forma *grodzi* rozpoznana m.in. jako forma rzeczownika *gród*.

4.2.3 Formy nierozpoznane

Analizator nie rozpoznaje ruchomych klityk czasownikowych dołączonych nie do czasownika, tj. *-(e)m, -(e)ś, -(e)śmy* itd. Zatem nie została rozpoznana forma *alem*. Są natomiast analizowane formy przyimków ze skróconą formą zaimka, jak np. *doń*. Forma *zeń* została rozpoznana, lecz zaklasyfikowana błędnie jako zaimek, a nie przyimek.

Nie zostały rozpoznane formy typu *prostu, polsku*, które występują tylko po przyimku *po*. Jednak zgodnie z twierdzeniem twórcy programu, przyczyną odrzucenia tych form jest założenie akceptowania takich form jedynie wewnątrz fraz *po polsku, po prostu*, gdy tymczasem w teście zostały one umieszczone w izolacji. Następnie, nie została rozpoznana część skrótów (głównie akronimy pisane wielkimi literami, nie zawierające kropki, np. *MHz, USA, PTTK, PCK*) oraz nazw własnych (np. nazwisko *Grzymały*). Odrzucone zostały także utworzone regularnie (tzn. poprzez przekształcenie istniejących słów) neologizmy, np. *metnik, komputeruje, kluczówka, pracowstręt, piotruje, wiertopięt*. Funkcja taka w Gram'ie istnieje, lecz na razie została zablokowana.

4.3 Statystyczna analiza tekstów

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach.

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
|S|	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	17	4	27	24	72
errors(S)	37	10	34	43	124
generated	431	110	277	852	1670
F-generated	329	86	213	647	1275
less	329	86	211	646	1272
more	0	0	0	0	0
non-exist	0	0	0	0	0
correct	427	101	267	833	1628
F-correct	325	77	203	628	1233
error-rate	0.087	0.0926	0.124	0.0525	0.0763
less-error-rate	1	1	1	1	1
more-error-rate	0	0	0	0	0
precision	0.991	0.918	0.964	0.978	0.975
F-precision	0.988	0.895	0.953	0.971	0.967
average-#correct	1.047	0.971	1.085	1.046	1.048

4.4 Wnioski

Zdecydowaną zaletą analizatora jest rozpoznawanie adresów e-mailowych, sieciowych (URL), liczb (rzymskich i arabskich), sekwencji znaków (np. ciąg '**********' jest rozpoznany jako 10[*]), skrótów zawierających kropkę, np. *inż.*, *hab.*. Słowniki imion, nazwisk i nazw geograficznych pozwalają na identyfikację, przynajmniej części z nich, w tekście.

Wadą analizatora są wyraźne braki słownika języka polskiego (brakuje dość powszechnych form) oraz niepełne słowniki nazw geograficznych. Widoczne są też błędy algorytmu rozpoznawania form, o czym świadczą przykłady omówione w rozdziale 4.2.2. W związku z tym można podejrzewać, że w przypadku analizy zawierającej kompletną informację morfosyntaktyczną, pojawiłyby się kolejne błędy, nie dające się wykryć na testowanym poziomie.

Na koniec warto zauważyć, że chociaż testowana graficzna wersja Gram'a nie nadaje się do obróbki dużych zbiorów tekstów, system jako taki wydaje się do tego celu nieźle przystosowany ze względu na możliwość sterowania jego działaniem poprzez ustawianie różnych opcji.

5 PoMor

Analizator PoMor jest produktem komercyjnym, sprzedawanym przez firmę MorphoLogic. Na potrzeby projektu, w ramach którego powstało obecne opracowanie, firma zgodziła się nieodpłatnie analizować nadsyłane jej teksty. Analizator jest dostępny jako wykonywalny plik `.exe`, działający w środowisku Windows. Program działa także w środowisku DOS i Macintosh, potrafi również analizować pliki zapisane w HTML i XML (nie tylko tekstowe). Litery polskie zapisywane są w kodzie Windows 1250.

Analizator rozpoznaje słowa dwóch języków polskich: język określany kodem 2069, to możliwie bogaty zasób polskich słów, zawierający także dawne formy oraz regionalizmy, zaś język określany kodem 1045 jest pomocny przy sprawdzaniu ortografii, a więc zawiera formy współczesne, z pominięciem niektórych rzadziej spotykanych

słów. Ponadto pierwszy z nich rozpoznaje formy dopełniacza l.mn. zakończone na **-ij, -yj**, deprecjatywne formy rzeczowników liczby mnogiej rodzaj męskososobowego (*głupie chłopy*) oraz formy 1. i 2. os. l.poj. rodzaju nijakiego czasu przeszłego (*śpiewałom, byłoś*). Wszystkie testy zostały przeprowadzone dla obydwu języków. W raporcie przeanalizowane zostały rozbieżności pomiędzy obydwoma językami dla testowych list słów, lecz dla plików tekstowych podane zostały tylko rezultaty analizy przy użyciu pełniejszej wersji słownika.

System wyposażony jest też w moduł syntezy morfologicznej, który nie był jednak przez nas testowany.

5.1 Oznaczenia PoS

Analizator stosuje bogate oznaczenia PoS, pokrywające się w dużej mierze z oznaczeniami z pracy Tokarski (1993) oraz częściowo Kurcz i in. (1990). Oznaczenia obejmują m.in.: liczbę, rodzaj i przypadek dla rzeczowników, przypadek i rodzaj dla przymiotników (w tym imiesłowów przymiotnikowych), formę, tryb, czas, aspekt, osobę i liczbę dla czasowników, stopień dla przymiotników i przysłówków. Dla liczebników zastosowany jest podział na liczebniki główne, zbiorowe i ułamkowe, wprowadzając także typ odmiany liczebników (typ *b*: liczebniki główne **dwa, oba, obydwa, trzy, cztery**; typ *a*: pozostałe liczebniki główne; typ *c*: liczebniki zbiorowe; typ *d*: liczebniki ułamkowe). Wartości przypadków liczebników są określane na podstawie tabeli, w zależności od rodzaju gramatycznego, (Tokarski, 1993, str. 24).

Analizator rozpoznaje ponadto skróty, liczby (rzymskie i arabskie), znaki interpunkcyjne. Analizator rozpoznaje także wiele nazw własnych, lecz nie ma oddzielnej etykiety identyfikującej je.

Wykaz oznaczeń, udostępniony przez dra Roberta Wołosza, znajduje się pod adresem:
`http://www.ipipan.waw.pl/~aniak/Wolosz/Skroty.pdf.gz`.

Ponadto program wyposażony jest w opcje, umożliwiające przypisywanie wyrazowi właściwości *dawny, przestarzały. indywidualizm, gwarowy* itp., a także określanie łączliwości czasownika z *się*. Opcje te nie były testowane.

Formy podstawowe Analizator określa formę podstawową, podając szczegółowe informacje dotyczące kategorii morfosyntaktycznych (np. przypadek, liczba, osoba, rodzaj, aspekt itd.).

Tabela poniżej podaje formy podstawowe podawane przez analizator dla poszczególnych części mowy.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów	przysłówek
forma podstawowa	mianownik liczba poj.	bezokolicznik	mianownik, l. poj., r. męski, st. równy	bezokolicznik	przymiotnik st. równy

Dla gerundiów (tj. rzeczowników odczasownikowych zakończonych na *-nie, -cie*) formę podstawową stamowi czasownik. Jeśli dany wyraz występuje także jako rzeczownik, to posiada także taką interpretację. Jest to konsekwencją rozróżnienia czynności (*podanie* czegoś komuś) od „zwykłego" rzeczownika (*napisać podanie*).

5.2 Wyniki testów dla list słów

5.2.1 Analiza form poprawnych

Wśród testowanych form poprawnych, oprócz regularnie utworzonych neologizmów, nie zostały rozpoznane słowa (w żadnym z dwóch języków): *książęciu, wielkanoc, mielić, mielił*, nazwiska *Grzymały, Madalińskiego* oraz adresy e-mail oraz URL. Są to świadome decyzje autora, przy czym *wielkanoc* nie została rozpoznana jako błędnie zapisana z małej litery. Natomiast brak nieodmiennego słowa *niczym*.

Zasadnicza różnica między językiem 2069 a 1045 polega na podawaniu większej liczby rozwiązań w wypadku analizy językiem 2069. Słowa, dla których zanotowano rozbieżności (podane tylko analizowane słowo i dodatkowa interpretacja zaproponowana przy analizie językiem 2069):

- *woźnicy*: *woźnik* (rzeczownik męskoos.);
- *cudzoziemcze*: *cudzoziemczy* (przymiotnik);
- *księży*: *księża* (rzeczownik, rodz. żeński);
- *imienia*: *imienie* (rzeczownik, rodz. nijaki);
- *przęśle*: *przęśl* (rzeczownik, rodz. żeński);
- *wyląg*ł, *wylęgła*: *wylec* (czasownik dokonany);

- *rośnie*: *rośnia* (rzeczownik, rodz. żeński);
- *pleć*: *pleć* (rzeczownik, rodz. żeński);
- *poręczy*: *poręcze* (rzeczownik, rodz. nijaki);
- *zboczy*: *zbocz* (rzeczownik, rodz. żeński);
- *pierze*: *pierz* (rzeczownik, rodz. męskorzeczowy);
- *brać*: *bracić* (czasownik dokonany);
- *wrogi*: *wróg* (rzeczownik męskoos., liczba mnoga).

Większość z tych słów nie występuje w słowniku Markowski (1999), ani w Szymczak (1975), a nawet w bardzo bogatym słowniku Podracki (2001). Są to wyrazy archaiczne, uwzględnione w słowniku Doroszewski (1997). Dotyczy to także form wyrazowych *awaryj, stacyj* stanowiących archaiczną postać dopełniacza liczby mnogiej od słów *awaria, stacja*. W przypadku interpretacji *wyląg̨ł, wylęgła* jako form czasownika *wylec* wydaje się, że jest to błąd.

W pozostałych przypadkach, wyniki analizatora były jednakowe dla obu języków, a zatem dalej będą one omawiane łącznie.

W kilku przypadkach podawane są dość zaskakujące interpretacje. Dotyczy to następujących słów:

- *liścia*: oprócz dopełniacza *liść*, dopełniacz l. poj. rzeczownika nijakiego *liście*;
- *dwojacy*: poza przymiotnikiem, rozważany jest rzeczownik rodzaju męskoosobowego *dwojak* (synonim *bliźniak*), gdy tymczasem naszym zdaniem *dwojak* jest rzeczownikiem rodzaju męskorzeczowego, oznaczającym podwójne naczynie (por. Markowski (1999); Doroszewski i Kurkowska (1973));
- nadmiarowe interpretacje gerundiów, np. *uczenia* jako l. mnoga od *uczenie*
- *wyląg̨ł*: dwukrotnie zinterpretowane jako forma czasownika dokonanego *wyląc* (dwa różne podziały słowa dają tę samą interpretację); podobnie dla rzeczownika męskorzeczowego *dziób*
- *pal*: oprócz rzeczownika *pal* i rozkaźnika dla *palić*, zinterpretowane jako rzeczownik żeński *pala*
- *grodź*: rzadko używany rzeczownik żeński *grodź* oraz rozkaźnik czasownika *grodzić*;
- przyimki ze ściągniętymi formami zaimków *zeń, doń* zinterpretowane jako formy nieodmienne, a nie przyimek+zaimek.

Słowa *liście, pala* są to wyrazy przestarzałe, obecnie niespotykane, i podobnie jak dla słów rozpoznawanych wyłącznie językiem 2069, pojawiają się jedynie w słowniku Doroszewski (1997).

Specyficzny problem powoduje czasownik *ciąć*. W zasadzie ma on aspekt niedokonany, jednak słowniki poprawnej polszczyzny Markowski (1999); Doroszewski i Kurkowska (1973) dopuszczają także aspekt dokonany, jednak jedynie w czasie przeszłym *ciął go szablą)*.

Interpretacje niektórych form są niesymetryczne. Forma *gzowi* jest zanalizowana jako forma przymiotnikowa (poza rzeczownikową), ale brak przymiotnikowej interpretacji dla analogicznej formy *biesowi* (tylko rzeczownik). Jest to świadoma decyzja twórców słownika, oparta na materiale leksykograficznym.

Analizator dysponuje bardzo bogatym słownikiem, który umożliwia rozpoznawanie i interpretację form z klitykami czasownikowymi, np. *-(e)ś, -(e)śmy* itd. Analizator podaje zatem możliwie wyczerpujące interpretacje słów. Jedynym dostrzeżonym przykładem niekompletnej analizy jest brak interpretacji formy *moście* jako liczby mnogiej od rzeczownika *mość* (w obu językach).

Analizator w sposób niejednolity traktuje znaki przestankowe. Rozpoznaje jedynie kropkę i myślnik/łącznik, gdyż mogą one pełnic specjalną rolę: kropka stanowi zakończenie wielu skrótów, łącznik występuje w złożonych przymiotnikach (*biało-czerwony*). Dla skrótów zakończonych kropką (*tzw.*) jest ona uznawana za część analizowanego wyrazu (wynik analizy zapisywany jest jako `tzw.[skr]`), w przeciwnym wypadku kropka uznana zostaje za „elemnt dodatkowy" (koniec zdania): ⟨*analiza wyrazu*⟩`+.[.]`. Jednak kropka występująca np. po nawiasie zamykającym jest ignorowana. Pozostałe dwa znaki kończące zdania ('!','?') nie były testowane. Inne znaki przestankowe zapisywane są w ramach analizowanej formy, brak ich jednak w samej analizie. Zgodnie z twierdzeniem twórcy analizatora, wersja programu działająca jako korektor pisowni wyłapuje błędy typu spacja przed przecinkiem czy po nawiasie otwierającym.

5.2.2 Analiza form niepoprawnych

Jedyne formy, które zostały zanalizowane to:

- *źrebiu, żołądzie*: zanalizowane jako formy rzadkich rzeczowników *źreb, żołąd*;
- *domie*: rzadki, kontrowersyjny wołacz rzeczownika *dom*;
- *odprząż*: kontrowersyjny rozkaźnik czasownika *odprząc*;
- *weźmij*: zanalizowane jako rozkaźnik czasownika *wziąć*;
- *lasi*: forma rzadkiego przymiotnika *laszy* pochodzącego od *Lach*;
- *dniowy*: forma przestarzała, występująca obecnie jedynie w złożeniach (*siedmio-dniowy*).

Spośród powyższych form, słownik Markowski (1999) podaje jawnie jako niepoprawną formę jedynie *weźmij*.

5.3 Statystyczna analiza tekstów

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach. W pliku `kraj` nie rozpoznano 2 form wyrazowych, w pliku `kultura` nie rozpoznano 2 nazw własnych, w pliku `świat` nie rozpoznano 14 form, w tym 12 nazw własnych i 2 skrótów nazw własnych. W pliku `publicystyka` nie rozpoznano 13 form, w tym 3 nazw własnych i 5 skrótów nazw własnych. Ponadto skrót *ul.* uznany został jedynie za rzeczownik (niezgodnie z kontekstem), gdy tymczasem *proc.* potraktowane został zarówno jako froma rzeczownikowa, jak i skrót.

Etykietą `[fraz]` opatrzone zostały następujące formy: *angielska, części, dęba, dobre, drodze, głąb, grand, kolei, miejsca, nowo, piśmie, polska, pro, prostu, przykład, warunkiem.*

W procesie analizy pojawiły się następujące słowa rzadkie bądź archaiczne: *gościa, imienie, jaka, koleja, kilka, mima, miotło, niża, posłanek, rówień, sąsiada, schrona, schroń, zbrodzień, musić, niżyć, swoić, kwietny.* Nie zostały potraktowane jako błędne.

Trzem czasownikom *powodzić, pozwalać, przypadać* przypisany został zarówno aspekt dokonany, jak i niedokonany. Są to przypadki homonimii morfologicznej: ewidentnie dla drugiego wyrazu (*pozwalać drzewa*); dwa pozostałe są rzadsze i bardziej kontrowersyjne (*powodzić kogoś za nos; przypadać lód śniegiem*).

Kwestia występowania odsłowników w liczbie mnogiej jest rzecz jasna dyskusyjna. Jednak rozważanie ich wyłącznie w liczbie pojedynczej pozwala uniknąć traktowania takch wyrazów jak *być, pięć* jako form odsłowników *bycie, pięcie.* Ponadto naszym zdaniem nadmiarowo za rzeczowniki zostały uznane wyrazy *przeprowadzenie, zachowanie*, zaś brak takiego oznakowania dla wyrazu *ograniczenie*; wątpliwości budzi też *utrzymanie.*

Ponadto za wyrazy występujące wyłącznie w liczbie pojedynczej uważamy (zgodnie z Markowski (1999)) *biotechnologia, gotowość, ludność, obrona, ochrona, psychologia, sprawiedliwość.* Liczba mnoga dla nazw własnych jest kwestą bardziej kontrowersyjną; PoMor podaje taką interpretację dla wyrazów *Gdańsk(i), Śląsk(i)*, ale już nie *Hag-a/i, Finlandi-a/i, Szwecj-a/i, Szwajcari-a/i* (w trzech ostatnich przypadkach ujawnia się to poprzez brak dopełniacza l.mn.).

Spacja i koniec linii są jedynymi, a zarazem bezwględnymi separatorami słów. W rezultacie, z jednej strony napisy *kandydatów.Łatwiej* oraz *).W* potraktowane zostały jako pojedyncze wyrazy, których nie udało się zanalizować. Z drugiej strony, pojawienie się karetki wewnątrz pojedynczej formy uniemożliwia jej identyfikację, a więc i poprawną analzię; w ramach przeprowadzonych testów dotyczy to złożonych przymiotników *kujawsko-*⟨`CR`⟩*-pomorski* i *filtro-*⟨`CR`⟩*-wentylacyjne.* Dotyczy to prawdopodobnie także słów dzielonych w wyniku przenoszenia do następnej linii ([...]-⟨`CR`⟩[...]); zjawisko to nie wystąpiło w badanych tekstach, więc nie zostało przebadane.

Wyrazom *niektóre, nietechniczne* przypisano analizy `nie[NEG]`⟨*wyraz*⟩`Adj`, gdy tymczasem wyrazowi *niejeden* — `niejeden[Adj]`, co można uznać za brak konsekwencji. Była to świadoma decyzja twórcy analizatora, wynikająca z przyczyn technicznych.

Dane poniżej są podane dla analizy językiem 2069. Wyrazom uznanym za niepoprawne w Markowski (1999) przypisano taką właśnie cechę. Nasze wątpliwości budzi także uznanie za formy nieodmienne wyrazów *jedno, prawo, nikt*; być może powinny być opatrzone etykietą `fraz`. Natomiast nieodmienne *a, da, do, na, o* to zapewne wykrzykniki. Za brak konsekwencji można uznać także przypisanie oznacznia `a[lit]` spójnikowi *a*, gdy tymczasem inne wyrazy jednoliterowe takiego oznaczenia nie mają.

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
|S|	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	2	2	14	13	31
errors(S)	4	2	15	16	37
generated	570	131	328	1056	2085
F-generated	404	99	249	803	1555
less	0	0	0	0	0
more	4	1	9	15	29
non-exist	1	1	0	3	5
correct	566	130	319	1041	2056
F-correct	400	98	240	788	1526
error-rate	0.0094	0.0185	0.0548	0.0195	0.0228
less-error-rate	0	0	0	0	0
more-error-rate	0.0099	0.0101	0.0361	0.0187	0.0186
precision	0.993	0.9924	0.9726	0.9858	0.9861
F-precision	0.990	0.964	0.981	0.981	0.982
average-#correct	1.338	1.226	1.232	1.290	1.288

5.4 Wnioski

Analizator dysponuje bardzo bogatym słownikiem języka polskiego, uwzględniającym formy dawne czy niektóre regionalizmy, a także nazwy własne (imiona, nazwiska, nazwy geograficzne itd.). Istotną zaletą jest podawanie bogatej informacji morfosyntaktycznej, np. uwzględniającej wielość rodzajów w języku polskim (w liczbie pojedynczej wyróżnione są trzy rodzaje męskie, rodzaj żeński, nijaki). Pewnym utrudnieniem przy ujednoznacznianiu opisów form w kontekście może być fakt, że przy przymiotnikach i liczebnikach podawane jest niekiedy tylko jedno oznaczenie dla kilku wartości, np. przy przymiotnikach '01' oznacza mianownik rodzaju męskiego (rodzaju m1, m2, m3) oraz biernik rodzaju m3, jednak bez trudu można dopisać algorytm „rozkodowujący" takie oznaczenia.

Analizator bardzo dobrze spełnia swoje podstawowe zadanie, tj. analizę morfologiczną słów. Nie ma natomiast dodatkowych funkcji umożliwiających rozpoznawanie adresów e-mailowych czy URL. Dodanie takich funkcji nie jest jednakże rzeczą trudną. Istotną wadą programu jest także fakt, że formy nieropoznane zapisywane są na oddzielny plik wynikowy, przez co gubiony jest kontekst występowania tych słów, a zarazem tracona jest ciągłość przetwarzanego tekstu. Także tę niedogodność można jednak w prosty sposób naprawić.

Natomiast za ważną zaletę można uznać możliwość opcjonalnego uzyskiwania dodatkowych informacji.

6 SAM

Kolejnym omawianym analizatorem jest SAM. Testowana wersja (SAM99:v3.4a, May 18 2001) dostępna jest spod DOS-u, litery polskie są zapisywane w kodzie latin2. W siecie dostępna jest także wersja dla systemu Linux, jednak w momencie rozpoczynania testów nie miałyśmy dostępu do tego systemu. Opis użytkowy programu znajduje się w pracy Szafran (1996).

Program może przyjmować zarówno słowa z klawiatury, jak i z pliku tekstowego. W tym drugim przypadku wyniki pojawiają się na pliku tekstowym, o nazwie `wyniki.sam`. Z góry określony zbiór wynikowy utrudnia poddanie analizie danych umieszczonych na kilku plikach. Program udostępnia również rozszerzoną analizę, w której sugerowane są także formy podstawowe nie występujące w słowniku. Leksemy takie mogą być do słownika dodawane, lecz funkcja ta nie była testowana, a w późniejszych wersjach systemu autorzy zrezygnowali z tej opcji. Istnienie możliwości rozbudowy słownika może być przydatne, w szczególności ze względu na zmiany zachodzące w języku polskim oraz specyfikę różnych podjęzyków branżowych.

6.1 Oznaczenia PoS

Analizator SAM rozróżnia następujące części mowy: rzeczownik, czasownik, przymiotnik, liczebnik, zaimek, przysłówek, imiesłów (przymiotnikowy czynny, bierny i przeszły, przysłówkowy współczesny i uprzedni). Pozostałe leksemy nieodmienne nie są w żaden sposób oznaczane (mają puste etykiety). Zarówno zaimki przymiotne, jak i liczebniki porządkowe, traktowane są jako przymiotniki. Dokładnie te same części mowy rozróżniane są przez analizator PoMor (por. rozdz. 5.1); jest to zgodne z indeksem morfologicznym Tokarskiego (1993).

Formy podstawowe rzeczowników (mian.l.poj.) etykietowane są za pomocą oznaczenia rodzaju (`m, ż, n, blp` — *plurale tantum*), po czym następuje liczba rzymska oznaczająca deklinację. Rodzaj ten wskazuje typ odmiany, ponadto dla rodzaju męskiego jest podawna bardziej szczegółowa informacja o rodzaju (`m1, m2, m3`, np. `koń(mI::m2)`), dotyczy to także wyrazów rodzaju męskoosobowego o odmianie żeńskiej (np. `wydawca(żII::m1)`). Poszczególne formy oznaczane są przez ciąg dużych liter opisujących przypadki (`N,H` — mian., `G,G'` — dop., `D` — cel., `T` — bier., `I` —narz., `L` — miejsc., `V` — wołacz; litera `l` poprzedza przypadki liczby mnogiej. W przypadku rzeczowników istnieją (ściśle określone i wymienione w opisie użytkowym programu) sytuacje, w których informacja o przypadkach, w jakich występuje dana forma rzeczownika, nie jest kompletna. I tak:

- Dla leksemów rodzaju nijakiego pomija się symb. `T` i `V`;
- Dla leksemów rodzaju żeńskiego, przed literą `l` pomija się symb. `L`;
- Dla leksemów rodzaju męskiego pomija się symb. `T`.
- Po literze `l` pomija się symb. `T` i `V`.

Może to utrudniać wykorzystanie programu do dalszej analizy (np. składniowej). W szczególności, ponieważ dla rzeczowników nie wyróżnia się rodzaju męskoosobowego, nie istnieje możliwość automatycznego stwierdzenia, czy biernik ma postać tożsamą z mianownikiem czy dopełniaczem.

Formy podstawowe czasowników są etykietowane za pomocą oznaczenia koniugacji (liczby rzymskie I–XIII), dodawana jest także informacja dotycząca aspektu (`dk, ndk`). Poszczególne formy czasu teraźniejszego/przyszłego prostego (domyślny) oznaczane są liczbami 1,2,3 określającymi osobę. Natomiast formy czasu przeszłego opatrzone są oznaczeniami rodzaju liczby: poza `m, ż, n` dla liczby pojedynczej, także `mo` — męskoosobowy oraz `rz` — niemęskoosobowy dla liczby mnogiej. Ponadto formy bezokolicznika oznaczone są literą `B`, formy bezosbnika — `b`, trybu rozkazującego — `i` (dodatkowo liczby 1, 2 określają osobę liczby mnogiej), formy imiesłowu przysłówkowego uprzedniego — `u`, formy imiesłowu przysłówkowego współczesnego — `w`, formy imiesłowu przymiotnikowego biernego — `A`, zaś formy imiesłowu przymiotnikowego czynnego — `w` (po czym następuje liczba determinująca formę przymiotnika, co odróżnia go od imiesłowu przysłówkowego współczesnego, oznaczanego tą samą literą), symbol `ł` służy do oznaczania form imiesłowu przymiotnikowego przeszłego. Formy odsłownika (gerundium) oznaczane są przez `g`, po czym następuje ciąg liter determinujących przypadek.

Formy podstawowe przymiotników etykietowane są za pomocą litery `A`. Przyjęto założenie, że wszystkie leksemy przymiotnikowe mają kategorie fleksyjne liczby (`sg, pl`), rodzaju (`m1, m2, m3, ż, n`), przypadka i deprecjatywności (*h, –h*). Jednak informacje te nie są podawane bezpośrednio, lecz za pomocą liczb zakodowanych w odpowiedniej tabeli, przedstawionej w opisie użytkowym analizatora. Ponadto przymiotniki w stopniu wyższym i najwyższym oznaczane są symbolem `com`; stopnie te nie są rozróżniane. Podobnie rzecz się ma w przypadku przysłówków (których formy podstawowe oznaczane są literą `J`). Wszystkie przymiotniki i przysłówki podlegają stopniowaniu (np. *calszy, finansowszy, rysunkowszy; dodatkowiej, obecniej*).

Formy podstawowe liczebników etykietowane są literą `K`. Liczebniki główne mają kategorie fleksyjne rodzaju (`m1, -m1`) i przypadka. Rownież w tym wypadku nie są one bezpośrednio podawane, lecz za pomocą liczb zakodowanych w odpowiedniej tabeli. Wyjątek stanowią liczebniki *dwa, oba, obydwa, trzy, cztery*, dla których jedyną podawaną informacją dotyczącą konkretnej formy jest `por.`, i użytkownik jest odsyłany do odpowiedniej tabeli. Ponadto wyraz *jeden* traktowany jest wyłącznie jako przymiotnik, choć jest to także liczebnik główny. Natomiast *jedno* potraktowane zostało jako wyraz nieodmienny nie powiązany z *jeden*.

Ponieważ jedyną kategorią fleksyjną liczebników zbiorowych jest przypadek, jest on określany dla każdej analizowanej formy. Z kolei dla liczebników ułamkowych (*półtora, półtrzecia*) podawana jest uproszczona informacja o rodzaju: `mn` — męski lub nijaki oraz `ż` — żeński.

Jedyny wyróżniany typ zaimków to zaimki rzeczowne, etykietowane literą `Z`. Formy tych zaimków mają oznaczoną kategorię fleksyjną przypadka, dla części z nich podawana jest kategoria akcentowości `akc, -akc`. Ponadto dla zaimka *on*, posiadającego poza przypadkiem także kategorie liczby i rodzaju, podawana jest jedynie informacja `por.`, odsyłająca do tabeli z opisu użytkowego analizatora.

Na koniec trzeba zaznaczyć, że przyimki posiadające „dłuższą" formę zakończoną literą '*e*' (np. *pod – pode, w – we*) oznaczone są literą `E`, co przy braku rozróżnienia pozostałych przyimków od spójników i partykuł można uznać za kuriozum, jednak jest to konsekwencją faktu, że to właśnie te przyimki są odmienne, a analizator oznacza wyłącznie wyrazy odmienne.

W poniższej tabeli przedstawiamy formy podstawowe podawane przez analizator dla poszczególnych części mowy.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów	odsłownik	przysłówek
forma podstawowa	mianownik liczba poj.	bezokolicznik	mianownik, l.poj., męski, st. równy	bezokolicznik	bezokolicznik	st.równy

Niektóre słowa mają także dodatkową informację o charakterze frazeologicznym. Przypadki wykryte w przebadanych plikach to `dobre (0:: na)`, `tylko (0:: nie)`, `tym (0:: poza)`, `wszystkim (0:: przed)`.

6.2 Wyniki analizy dla listy słów

6.2.1 Analiza form poprawnych

Analizator SAM akceptuje większość spośród przebadanych poprawnych słów j. polskiego. Napotkane wyjątki to:

- *łazęgów* — dopełniacz liczby mnogiej dla rodzaju męskoosobowego;
- przymiotnik *dwojaki* uznany dodatkowo za l.mn. od rzeczownika *dwojak*;
- *dwunogi* — potraktowany tylko jako forma rzeczownika *dwunóg*, a nie jako przymiotnik.

Powyższe przykłady dotyczą raczej zawartości słownikowej bazy danych, a nie algorytmu analizy. Istnieją jednak przykłady na braki (błędy) w programie.

- *susi* — traktowany jest jako mianownik liczby mnogiej rzeczownika *sus*, a nie przymiotnika *suchy*. Podobnie *cisi* od *cis* oraz *głusi* od *głuch*. Wynika to z odmiany *mnich — mnisi* i *lowelas — lowelasi*, co jednak nie jest dla powyższych słów poprawne, a tym bardziej nie tłumaczy braku odmian przymiotnikowych.
- *najciemniej* — traktowany jako rozkaźnik od *ciemnieć*. Podobnie *najjaśniej*, ale w tym wypadku stopień wyższy przysłówka jest też uwzględniony.
- W przypadku, gdy istnieją dwie formy bezokolicznika danego czasownika (np. *wyląc, wylęgnąć* czy *ugrząść, ugrzęznąć*), pewne formy wiązane są z jednym, a inne z drugim. Np. *ugrzązł — ugrzęznąć*, ale *ugrzęźnij — ugrząść*.
- *tnę* — w ogóle nie jest akceptowane (chociaż *tnie* — tak). Podobnie *klnę, gnę*. Formy *ciągnę, wezmę, rozpocznę* analizowane są prawidłowo.
- *wylęgła* — traktowane jest jako imiesłów przymiotnikowy od *wylęgnąć*. Podobnie *pełźli*. Odpowiada to frazom typu *świeżo wylęgłe pisklęta*, co nie tłumaczy jednak braku form osobowych czasownika. Taką, nieuzasadnioną interpretację mają też słowa *zaprzęgła* oraz pochodzące z plików tekstowych *zdezaktualizowały, nakładały*, niezależnie od poprawnie rozpatrzonych form osobowych.
- *lepiej* — przysłówki stopniowane nieregularnie miewają dwie formy podstawowe — *lepiej, dobrze*;
- niepoprawna interpretacja *działa, woła, boi* jako dopełniacza od *dział, wół, boje*.

Analizator SAM nie rozpoznaje liczb (w tym dat). Ponadto ponieważ znaki przestankowe, takie jak kropka, przecinek, myślnik/łącznik, nie są traktowane jako separatory słów, brak spacji powoduje „zlepienie" słów i w konsekwencji brak interpretacji. Dotyczy to także form dwuwyrazowych, takich jak *biało-czerwony*, opatrzonych pustym opisem. Ponadto omawiany analizator nie rozpoznaje ani skróconych form zaimka (*doń, zeń*), ani ruchomych klityk czasownikowych dołączonych nie do czasownika (*alem*). Zignorowane zostały również testowane skróty, zarówno wyrazów pospolitych (*mgr, MHz*), jak i nazw własnych (*USA*). Dotyczy to także większości nazw własnych, z wyjątkiem tych, które pokrywają się z wyrazami pospolitymi. Teoretycznie analizator wyróżnia nazwy własne, ale jedynym zarejestrowanym przypadkiem rozpoznania nazwy własnej była forma *Marii* (`Maria (żI:: ⟨iw⟩)`). Natomiast formy typu *prostu, polsku* zostały potraktowane jak inne leksemy nieodmienne. Neologizmy, nawet „regularnie" tworzone, nie zostały rozpoznane przez podstawowy analizator, oparty na słowniku, zaś wersja „sugerująca" rozwiązania jest niezmiernie nadmiarowa. Na przykład forma podstawowa dla słowa *metnik* to rzeczowniki *metnik, metnika, metniko*; zaś dla słowa *piotruje* to rzeczowniki *piotruj, piotruja*, przymiotnik *piotrui* oraz czasowniki *piotrujeść, piotrować, piotrywać, piotrać, piotruć*. Należy zauważyć, że w każdym przypadku analiza zgodna z intuicją została uwzględniona. Na koniec, adresy elektroniczne nie zostały rozpoznane.

6.2.2 Analiza form niepoprawnych

Natomiast omawiany analizator akceptuje także wiele form nieporawnych, zwłaszcza w wypadku rzeczowników, co uwidoczniło się już po części w poprzednim rozdziale. Wynika to prawdopodobnie z faktu, że uznawane są wszystkie możliwe formy właściwe dla np. danej deklinacji. Przykłady:

- *domek – domku, *domka*, ale *kwiatek – kwiatka* (akceptowane są także formy **domeku, *kwiateka*);
- podobnie *skuter, skutera, *skutra* i *kuter, kutra, *kutera*;

- *sypialnia – sypialni, *sypialnii*, ale *agonia – agonii, *agoni*;
- *mądry – mądrzejszy*, ale *dobry – *dobrzejszy*.

Takie przykłady można mnożyć. Są one skutkiem przyjętego w SAM-ie założenia, że analizator akceptuje nie tylko formy zgodne z rzeczywistą odmianą danego wyrazu, ale wszystkie formy, które mogą zostać utworzone dla tego wyrazu zgodnie z indeksem Tokarskiego. Wynika z tego jednak wyraźnie, że analizator ten może mieć zastosowanie jedynie w sytuacjach, w których akceptowanie form niepoprawnych nie przynosi zbyt wiele szkody. Niewątpliwie oznacza to także, że rozbudowanie programu o moduł syntezy morfologicznej mogłoby być utrudnione (lub wiązałoby się z generowaniem wielu niepoprawnych form).

6.3 Statystyka analizy plików tekstowych

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach. W pliku `kraj` nie rozpoznano 59 form wyrazowych, w tym 25 nazw własnych i 21 liczb. W pliku `kultura` nie rozpoznano 13 form, w tym 10 nazw własnych i 2 liczb. W pliku `świat` nie rozpoznano 40 form, w tym 33 nazw własnych, 4 liczb i 2(6) skrótów. W pliku `publicystyka` nie rozpoznano 159 form, w tym 33 nazw własnych, 7(14) skrótów i 12 liczb (dwa skróty, *ul.* i *proc.* potraktowane zostały jako rzeczowniki). Liczby w nawiasach oznaczają ogół skrótów, w tym skróty nazw własnych. Ponadto zdarzały się przypadki (2, 5, 1, 8 odpowiednio), w których nazwa własna została rozpoznana jako inny wyraz, np. *Polska* jako żeńska forma przymiotnika *Polski* czy *Poznaniu* jako forma odsłownika *poznanie*. Fakt, że analizator zawiera słowo *Warszawa*, a nie zawiera słowa *Polska*, można uznać za dziwny, choć podobnie zachowuje się LEM. Brak także analiz dla następujących wyrazów: *informatyka, biotechnologia, iberystyka, politologia, wiceszef, dezaktywacji, powinniśmy, nietechniczne, niepublikowane, pochodzących, łatwiej, zamierzające, przypadających*, gdy tymczasem słownik zawiera wiele rzadkich, archaicznych słów, takich jak *imienie, niża, posłanek, poręcze, prześl, rówień, sąsiada, schrona, straża, zbrodzień, bracić, ciesać, niżyć, swoić, odpowiedny, wcześny, bezpieczno, trudnie*. Wyrazy archaiczne etykietowane są literą `d`, ale oznakowanie nie jest konsekwentne: brak go dla słów *rośnia, kwietni, mima, schroń*, gdy tymczasem z niewiadomych przyczyn widnieje przy wyrazie *powodzić (się)*.

Ponadto brak analizy dla formy *tego*, mimo że *tym* jest poprawnie wiązane z *ten*, oraz *bunkry*, chociaż formę *bunkrów* zinterpretowano prawidłowo.

Chociaż słowa takie jak *drzwi* mają rodzaj `blp.`, to analizator nie przewiduje sytuacji odwrotnej — braku liczby mnogiej, dla takich słów jak *bezpieczeństwo, gotowość, ludność, obrona, ochrona, sprawiedliwość*. Także wiele odsłowników niepotrzebnie traktowanych jest dodatkowo jako rzeczowniki (*budowanie, bycie, przeprowadzenie, wyznaczenie, wznoszenie, zainstalowanie, zarządzanie, zgłębianie*) i bardziej kontrowersyjne *utrzymanie*, co prowadzi do błędnego traktowania dopełniacza tych słów także jako mianownika liczby mnogiej. Na podobnej zasadzie liczebnik *pięć* został zinterpretowany jako forma rzeczownika *pięcie*.

Ponadto słowo *były* traktowane jest wyłącznie jako przymiotnik, a nie imiesłów przymiotnikowy przeszły od *być*; *więcej, najwięcej* opisywane są jako pozbawione stopnia wyrazy nieodmienne o formie podstawowej *więcej*, bez wiązania z *dużo, wiele*; *równie* stanowi formę rzeczowników *równia, rówień*, a nie przysłówek (podobnie *zwłaszcza*); *widać* uznane jest za słowo nieodmienne, a nie czasownik niewłaściwy; *bunkrów, filtrów, remontów* traktowane są jako formy skrócone przmiotników *bunkrowy, filtrowy, remontowy* (poza poprawną analizą), zaś czasowniki *próbować, wlec* mają przypisany aspekt dokonany.

Czasowniki mogące występować z zaimkiem *się* opatrywane są taką etykietą (ujętą w nawias, gdy zaimek ten nie jest wymagany). Oznakowanie takie stosowane jest bardzo niekonsekwentnie. Po pierwsze, brak takiej etykiety dla słów *dostać, parać, wziąć*, chociaż *parać* ewidentnie wręcz wymaga wystąpienia tego zaimka. Po drugie, etykieta `(się)` towarzyszy słowom *bóść, obowiązywać, oskarżyć, oznajmić, pleść, przesłuchać, próbować, przewidywać, reprezentować, rozstrzygnąć, uregulować, ustalać, wyeliminować, wyznaczyć, zlokalizować*, które wręcz przeciwnie nie występują z *się*, pomijając banalne przypadki fraz bezosobowych *przewiduje się, oskarży się*, które dotyczą także czasowników bez takiej etykiety (*ugrzęźnie się, odrzuca się*). Jedynymi (testowanymi) czasownikami nie posiadającymi takiej formy są *boleć* i *obowiązywać* (!). Z drugiej strony, wystąpienie *się* po wyrazach *płaszczyć, wylęgnąć* jest obowiązkowe. Na koniec oznakowanie `się` przypisywane jest słowom *odrzucić, zarządzać*, choć powinny one mieć etykietę pustą.

Niedeprecjatywna forma mianownika l.mn. występuje tylko dla rzeczowników rodzaju męskoosobowego. Niewyróżnianie tego rodzaju powoduje proponowanie takiej analizy dla wspomnianych już form *susi, cisi* (od *sus, cis*), a także *czeka, musi, odrzuci, stolicy, tworzy, ludzi* od *czek, mus, odrzut, stolik, twór, lud*. Dotyczy to także form *grodzi, powodzi, uczeni* od *gród, powód, uczeń* z listy słów.

Najbardziej zaskakujący zdaje się sposób odmiany przez przypadki słowa *rok* w liczbie mnogiej (*lata*). I tak mianownik *lata* wiązany jest zarówno ze słowem *rok*, jak i *lato*, gdy tymczasem *lat, latach* stanowi jedynie formę słowa *lato*, bez powiązania z *rok* czy choćby *lata*. Lista błędów jest dłuższa, lecz poprzestaniemy na już wymienionych przypadkach.

Wyniki te przekładają się na następujące wyniki statystyczne:

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
$\|S\|$	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	59	13	40	73	185
errors(S)	108	19	56	109	292
generated	413	104	283	876	1676
F-generated	317	78	216	684	1295
less	88	5	51	179	323
non-exist	14	1	11	22	48
more	14	1	13	40	68
correct	318	88	221	723	1350
F-correct	211	65	160	531	967
error-rate	0.254	0.176	0.205	0.133	0.180
less-error-rate	0.278	0.064	0.236	0.262	0.250
more-error rate	0.044	0.013	0.060	0.058	0.053
precision	0.770	0.846	0.781	0.825	0.805
F-precision	0.666	0.833	0.740	0.776	0.747
average-#correct	0.868	0.926	0.948	0.968	0.937

6.4 Wnioski

Niestety, wobec przyjętego przez nas założenia *poprawne formy, wszystkie poprawne, wyłącznie poprawne*, SAM wypada słabo w porównaniu z innymi analizatorami, niezależnie od kryteriów oceny. Po pierwsze, wyróżnia jedynie podstawowe części mowy, zgodne z indeksem Tokarskiego (co prawda PoMor też), nie rozróżnia też ani liczb, ani skrótów. Istotną wadą jest także świadomie przyjęte założenie opuszczania niektórych przypadków przy opisie rzeczowników, co wystarcza do identyfikacji danej formy, może jednak utrudniać dalsze przetwarzanie informacji ‘, np. anotację morfosyntaktyczną tekstu. Omawiany analizator posiada bogaty słownik, brak w nim jednak wielu współczesnych wyrazów, takich jak *informatyka* czy *biotechnologia*. Biorąc jednak pod uwagę wyraźnie nadmiarowy algorytm analizy, wskazujący wiele nieistniejących opisów danej formy, bogactwo słownika, szczególnie występowanie w nim wielu rzadkich słów, można uznać za wadę (co objawia się niską wartością wskaźnika F-precision). Natomiast istotną zaletą wyróżniającą SAM-a wśród przetestowanych analizatorów jest możliwość proponowania analizy dla słów nieobecnych w słowniku czy też neologizmów (abstrahując od silnej redundancji tego procesu), wraz z umożliwieniem dopisywania ich do słownika, co pozwala na tworzenie własnego słownika, zgodnego z konkretnym zastosowaniem.

7 LEM

Wersja systemu, z jaką mieliśmy szansę zetknąć się bezpośrednio, dostępna jest spod DOS-u. Nie jest ona wyposażona w najprostszy nawet system dialogowy — wywołanie programu polega na podaniu pliku z tekstem źródłowym oraz pliku wynikowego, na który zapisywany jest rezultat analizy. Litery polskie zapisywane są w kodzie Windows 1250. System wyposażony jest też w moduł syntezy morfologicznej, który nie był przez nas testowany.

Opis narzędzi użytych przy tworzeniu analizatora, w szczególności taksonomii cech morfosyntaktycznych pojawiających się na plikach wynikowych tego programu, można znaleźć w pracy Vetulani i in. (1998).

7.1 Oznaczenia PoS

Analizator LEM dostarcza bardziej dokładną informację morfologiczną niż omówiony powyżej SAM. Rozważane są następujące części mowy: rzeczownik — `N`, zaimek rzeczowny — `NPRO`, czasownik — `V` (wyróżnione czasowniki modalne — `VM` i niewłaściwe — `VNI`), odsłownik — `NV`, przymiotnik — `ADJ` (wyróżnione przymiotniki liczebnikowe — `ADJNUM`, takie jak *dwojaki, pojedynczy, podwójny*), zaimek przymiotny — `ADJPRO`, przysłówek — `ADV`, zaimek przysłowny — `ADVPRO`, imiesłów (przymiotnikowy czynny — `ADJPRP`, bierny — `ADJPAP` i przeszły — `ADJPP`, przysłówkowy współczesny — `ADVPRP` i uprzedni — `ADVANP`), liczebnik (główny — `NUMCRD`, porządkowy — `NUMORD`, zbiorowy — `NUMCOL`), przyimek — `P`, spójnik — `CONJ`, partykuła — `PART`, onomatopeja — `ONO`, zawołanie — `APP` (tak oznakowana została jedna z interpretacji *dalej*). Dla tych części mowy

oznaczane są następujące kategorie morfologiczne (w j.angielskim) Aspect (perfect, imperfect); Verbform (b – infinitive, personal, impersonal); Mood (declarative, conditional, imperative); Tense (past, present, future); Person (1,2,3); Degree (positive, comparative, superlative); Number (singular, plural); Gender (personal, animal, inanimal, female, neutral); Case. Rzecz jasna, kategorie morfologiczne wyznaczane są jedynie dla tych części mowy, których dotyczą; dla przyimków podawane są wymagane przez nie przypadki. Kolejność umieszczania etykiet nie jest jednoznacznie ustalona, nawet w ramach tej samej części mowy (przymiotniki).

Zaproponowane rozróżnienie pomiędzy czasem przyszłym i teraźniejszym wydaje się redundantne, skoro wszystkie czasowniki występujące w którymś z tych czasów opatrywane są etykietą Trf, a czasy te rozróżniane są wyłącznie za pomocą aspektu. Co gorsza, dla czasownika *być* aspekt w ogóle nie jest podawany, i w ten sposób formy *jest, będzie* czy *są, będą* mają identyczne opisy.

Kiedy analizowana forma występuje we wszystkich przypadkach, czasem oznaczana jest za pomocą G* (rzeczowniki *plurale tantum*, zaimki rzeczowne *się, siebie*), a czasem w ogóle pomijana (przymiotniki).

W poniższej tabeli przedstawiamy formy podstawowe podawane przez analizator dla poszczególnych odmiennych części mowy.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów przym.	odsłownik	przysłówek
forma podstawowa	mianownik liczba poj.	bezokolicznik	mian., l.poj., męski, st. równy	mian., l.poj., męski, st. równy	mianownik liczba poj.	st.równy

Jak widać, imiesłowy i odsłowniki nie są wiązane z czasownikiem, od którego pochodzą.

7.2 Wyniki analizy dla listy słów

7.2.1 Analiza form poprawnych

W słowniku brak jest najprawdopodobniej leksemów *dąb, dwunóg, okoń, przysłowie, Turek, bóść, nadejść, święcić, wylęgnąć (wyląc), żółcić, wypełzły (*imiesłów*), srodze, codziennie*, w tym także słów archaicznych *imienie, liście, przęśl, rośnia, mędrszy, cudzoziemczy*, co jest przyczyną braku analizy dla różnorakich ich form. Dziwić może natomiast brak czasowników *pójść* (formy *pójdźmy, poszedł*) czy *znaleźć* (forma *znajdź*), i modalnych *wolno, można, powinien* (dwa ostatnie były w poprzedniej wersji programu). Natomiast *widać* zostało potraktowane jako „zwykły" bezokolicznik, a nie czasownik niewłaściwy; jedynym wyrazem opisywanym jako czasownik niewłaściwy (etykieta VNI) jest *czmych*.

Ponadto wykryte zostały następujące błędy.

- *gzowi* — występuje tylko jako forma przymiotnika *gzowy*; analizator przyjmuje za to *gzu*;
- *mąk* — tylko od *mąka*, brak słowa *męka*;
- *koniami* — tak, *końmi* — nie. Jednak obie formy *dłoniami, dłońmi* są uznawane;
- forma *wstęg* została przyjęta, natomiast *wstąg* — nie;
- *woźnych* — wiązane jest z leksemem rzeczownikowym *woźny*, brak leksemu *woźna*;
- *wrogi* — nie jest traktowane jako l.mn. od *wróg*;
- brak rzadkiej odmiany *zębu* (roślina *koński ząb*);
- brak rzeczownika rodzaju męskoosobowego *głąb*, stąd odrzucenie formy *głąba*;
- *wyląg, wylęgła* — nie akceptuje;
- *stop* — brak rozkaźnika od *stopić*;
- *strzyże* — brak analizy, choć forma *strzygą* jest poprawnie wiązana z czasownikiem *strzyc*;
- *ugrząźć* — nie akceptuje (jest *ugrząść* — błąd ortograficzny), literówka także w słowie *wręczyc*;
- *brać* — brak rzeczownika;
- *wypełzły* — brak imiesłowu;
- przymiotnik *dwojaki* — brak formy męskoosobowej *dwojacy*;
- *gródź* — brak rozkaźnika od *grodzić*; natomiast odwrotnie *grodź*: rozkaźnik jest, rzeczownika brak;
- przymiotnikowa interpretacja formy *sypialni*, będąca wynikiem specyfiki algorytmu analizy morfologicznej dla przymiotnika, choć w rzeczywstości wyraz *sypialny* występuje jedynie w rodzaju męskorzeczowym;

- błędne oznakowanie aspektu dla czasowników *wlec, rosnąć*;
- brak partykuły *czyż*.

Analizator nie rozpoznaje ruchomych klityk czasownikowych dołączonych nie do czasownika, (zignorowanie form *alem, żebyś*). Akceptacja formy *biało-czerwona* świadczy raczej o umieszczeniu jej w słowniku niż o rozbudowaniu algorytmu o metodę rozpoznawania takich form, o czym świadczy m.in. odrzucenie form *polsko-francuski, obronno-ochronne*.

Wszystkie słowa oznaczane są przez analizator LEM symbolem ⟨`W`⟩, liczby naturalne — symbolem ⟨`N`⟩, znaki przestankowe ('.',',', ale także $ czy @) — symbolem ⟨`P`⟩, spacja i karetka (ale także * i %) — symbolem ⟨`B`⟩. Ponieważ jednak znaki przestankowe traktowane są jako (bezwarunkwe) separatory słów, liczby rzeczywiste, daty czy adresy elektorniczne nie tylko nie są rozpoznawane, ale zostają rozbite na części. Z drugiej strony myślnik/łącznik ('-') interpretowany jest jako litera, dwa słowa połączone łącznikiem traktowane są jako jedno słowo, ale samotny myślnik oznaczany jest jak słowo symbolem ⟨`W`⟩. Zignorowane zostały także formy typu *prostu, polsku*, które występują tylko po przyimku *po*. Liczby rzeczywiste oraz daty dzielone są „na kawałki" — liczby naturalne pooddzielane znakami przestankowymi. Podobnie rzecz się ma z adresami elektronicznymi. Zignorowane zostały również testowane skróty, zarówno wyrazów pospolitych (*mgr, MHz*), jak i nazw własnych (*USA*). Dotyczy to także większości nazw własnych, z wyjątkiem tych, które pokrywają się z wyrazami pospolitymi. Na koniec, odrzucone zostały także neologizmy.

Stwierdzenie, że niektóre przymiotniki pełnią również rolę rzeczowników, jest truizmem. Kłopot w tym, że trudno jest zdecydować, którym przymiotnikom warto przypisać podwójną interpretację. Twórcy analizatora LEM uczynili to dla następujących przymiotników (spośród przebadanych): *mądry, młody (młoda), polityczny, polski, zły*. Dziwi w tym kontekście brak takiej podwójnej interpretacji dla słów *biegły, święty*, i choćby *chętny*.

7.2.2 Analiza form niepoprawnych

Lem odrzuca większość form niepoprawnych. Poniżej przeedstawiamy kilka wyjątków.

- *męk* — dopełniacz l.mnogiej wyrazu *męka*, brak formy *mąk*;
- *ucznie, uczni* — błędne formy wyrazu *uczeń*;
- *koniami* — zamiast *końmi*, *gzu* zamiast *gzowi*;
- *ugrzęznąłem, odprząg, odprzęż* — błędne formy czasownikowe;
- *dniowy* — forma występująca wyłącznie w złożeniach (*trzydniowy*.

7.3 Statystyka analizy plików tekstowych

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach. W pliku `kraj` nie rozpoznano 18 form wyrazowych, w tym 12 nazw własnych. W pliku `kultura` nie rozpoznano 14 form, w tym 8 nazw własnych i 3 słów obcych. W pliku `świat` nie rozpoznano 37 form, w tym 24 nazw własnych, 1 liczby i 2(8) skrótów. W pliku `publicystyka` nie rozpoznano 65 form, w tym 33 nazw własnych, 7(14) skrótów i 1 liczby (ponadto podobnie jak w innych analizatorach, dwa skróty, *ul.* i *proc.* potraktowane zostały jako rzeczowniki). Jak poprzednio, liczby w nawiasach oznaczają ogół skrótów, w tym skróty nazw własnych. Ponadto zdarzały się przypadki (3, 5, 0, 4 odpowiednio), w których nazwa własna została rozpoznana jako inny wyraz, np. *Polska* jako żeńska forma przymiotnika *polski* czy *Poznaniu* jako forma odsłownika *poznanie*. Brak także analiz dla następujących wyrazów: *dezaktywacji, ograniczeń, wiceszef, nietechniczne, wentylatorowa wyróżniana, dofinansować, podusić, powinno, powinniśmy, wyremontowano, wzmacniając, zdezaktualizowały*. Brak także czasownika modalnego *można* (był w poprzedniej wersji słownika).

Duża liczba słów nieodmiennych oznakowana jest zarówno jako partykuła, jak i spójnik (*a, ale, chociaż, czy, jednak, też, tylko, więc, wprawdzie*), wiele innych ma także podwójne, a nawet potrójne oznakowania. W szczególności, jako wykrzyknik (etykieta `EXCL`) oznakowane zostały wyrazy *a, aha, ale, i, o*, jako zawołanie (etykieta `APP`) — *dalej*, i jako onomatopeje (etykieta `ONO`) — *ciach, miau, tak*. Powoduje to generowanie dużej liczby form dla słów nieodmiennych (średnio 1.28 w plikach tekstowych łącznie, bez liczb i skrótów), w przeciwieństwie do form odmiennych (średnio 1.04).

Rzeczowniki *bezpieczeństwo, ludność, psychologia, sprawiedliwość, ultimatum* mają rozpatrywane formy liczby mnogiej, choć występują wyłącznie w liczbie pojedynczej. Nie dotyczy to już jednak podobnych słów *gotowość, ochrona, obrona*. Podobnie rzecz się ma z wyrazem *wznoszenie*, w ogóle niepotrzebnie interpretowanym jako rzeczownik (wystarczy odsłownik). Z drugiej strony wyrazy *porozumienie, przygotowanie, przyjęcie, uregulowanie, zainteresowanie* potraktowane zostały wyłącznie jako odsłowniki, choć są także rzeczownikami, o czym świadczy

istnienie form liczby mnogiej. W opisie formy *specjalności* zabrakło mianownika, biernika i wołacza liczby mnogiej. Przymiotnik *większy* potraktowany został nie tylko jako stopień wyższy przymiotnika *duży*, ale także jako niezależny przymiotnik w stopniu równym.

Ponadto podczas kompresji opisów przymiotników i imiesłowów przym., dla form rodzaju żeńskiego liczby pojedynczej, oznaczenie mianownika zostało zastąpione przez biernik (np. ⟨`W sprawny, ADJ/DpNsCavGf`⟩`sprawna`). Dotyczy to wszystkich napotkanych form tego typu i znacząco wpłynęło na wyniki statystyczne dla analizatora LEM. Błędem mającym charakter typowej „literówki" jest też interpretowanie słowa *pod* jako formy czasownika podać — ⟨`W podać,V/RnApVpMdTrfNsP3`⟩`pod` — (powinno być *poda*).

Dla analizatora LEM uzyskano następujące wyniki statystyczne charakteryzujące jego działanie na testowaych plikach tekstowych:

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
|S|	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	18	15	37	65	135
errors(S)	24	22	38	76	160
generated	456	110	286	914	1766
F-generated	314	76	203	654	1247
less	5	3	2	13	23
non-exist	1	0	1	6	8
more	6	2	4	18	30
correct	449	106	281	892	1728
F-correct	307	72	198	632	1209
error-rate	0.056	0.204	0.139	0.093	0.083
less-error-rate	0.016	0.039	0.010	0.020	0.018
more-error rate	0.019	0.026	0.020	0.027	0.024
precision	0.985	0.964	0.983	0.976	0.980
F-precision	0.974	0.947	0.975	0.966	0.971
average-#correct	1.103	1.139	1.191	1.182	1.160

7.4 Wnioski

Analizator LEM prezentuje się dobrze na tle innych analizatorów. Przede wszystkim wyróżnia się bogatym zestawem części mowy oraz niezależnym etykietowaniem wszystkich rozważanych parametrów. Wyniki statystyczne psuje brak wielu wyrazów w słowniku, w tym wielu nazw własnych, oraz występujące w nim błędy. Ponieważ jednak słownik jest cały czas aktualizowany, trudno takich błędów uniknąć.

8 XeLDA

XeLDA (**Xe***rox* **L***inguistic* **D***evelopment* **A***rchitecture*) jest pakietem lingwistycznym opracowywanym w firmie Xerox. Program zaprojektowany został dla systemów Windows NT lub Solaris i działa w środowisku Java. Pakiet opracowany został dla wielu języków, nie tylko europejskich. Aby go uzyskać, potrzebna jest specjalna licencja. Moduł języka polskiego wciąż jest w fazie testowania; dzięki temu Zespół Inżynierii Lingwistycznej IPI PAN uzyskał prawa do wykorzystania pakietu do celów naukowych (licencja akademicka). Poza analizatorem morfologicznym, pakiet zawiera także moduł dezambiguacji i identyfikowania fraz rzeczownikowych.

Program działa w systemie okienek, tak więc tekst do analizy może być albo wpisany ręcznie do okienka, albo wciągnięty z pliku, a następnie zaznaczony (jak do kopiowania). Wynik analizy pojawia się w drugim okienku; można go ewentualnie zapisać na plik. Litery polskie zapisywane są w kodzie Windows 1250.

8.1 Oznaczenia PoS

XeLDA rozróżnia następujące części mowy: rzeczownik (`NOUN`), czasownik (`Verb`), przymiotnik (`ADJ`), liczebnik (`NUM`), zaimek (`Pron`), przysłówek (`ADV`), imiesłów (przymiotnikowy — `VPRT` — czynny (`Act`), bierny (`Pass1`) i przeszły (`Pass2`), przysłówkowy — `VGER` — współczesny (`Pres`) i uprzedni (`Past`)), przyimek (`PREP`), spójnik (`CONJ`), partykuła (`PRTCL`).

Rzeczowniki etykietowane są za pomocą przypadka (`Nom, Gen, Dat, Acc, Ins, Loc, Voc`), liczby (`Sg, Pl`) i rodzaju (`M1` — *męskoosobowy*, `M2` — *męskozwierzęcy*, `M3` — *męskorzeczowy*, `Fem, Neut`). Wyrazy nieodmienne (*muzeum, ultimatum*) opatrywane są etykietą `InvCase` (ew. także `InvNum`). Niektóre formy rzeczowników męskoosobowych oznaczane są także etykietą `Depr` mówiącą o deprecjatywności formy (*gospodarze, funkcjonariusze, obrońcy, przywódcy, właściciele, woźnice*); jednak we wszystkich tych przypadkach forma opisywana jest także jako „zwykły" mianownik. Jedyne wyjątki (spośród przetestowanych) to niepoprawne *mieszkaniowe (mieszkaniowiec)* i *Czechy (Czech)*.

Opis poszczególnych form rzeczownikowych wygląda następująco:

⟨*f.podst*⟩+`Noun`+⟨*rodzaj*⟩+⟨*liczba*⟩+⟨*przypadek*⟩+`NOUN`.

Wśród rzeczowników wyróżnione są nazwy własne, dla których etykieta `Noun` zastępowana jest przez `Prop`. Do tej grupy zaliczane są skróty, opisywane następująco:

inż.`+Noun+Abbr+InvNum+InvCase+NOUN`.

Jeśli jest to skrót nazwy własnej (*PCK*), etykieta `Noun` zastępowana jest przez `Prop`.

Czasowniki etykietowane są przez aspekt (`Perf, Imperf`). Formy osobowe oznaczane są ponadto przez tryb (`Ind, Impv, Cond`), czas (`Past, Pres, Fut`, przy czym czas teraźniejszy i przyszły rozważane są osobno), osobę (`1P, 2P, 3P`), liczbę, a dla czasu przeszłego także rodzaj (`Masc, Fem, Neut` dla liczby pojedyńczej oraz `M1, M23FN` dla liczby mnogiej). Pełny opis poszczególnych form osobowych przedstawia się następująco:

`Verb`+⟨*aspekt*⟩+⟨*tryb*⟩+⟨*czas*⟩+⟨*osoba*⟩+⟨*liczba*⟩+{⟨*rodzaj*⟩}+`VFIN`,

zaś form bezosobowych następująco: `Verb`+⟨*aspekt*⟩+`Nonpers+VFIN`.

Poza formami osobowymi, oznaczanymi wspólnie z bezosobowymi jako `VFIN`, i bezokolicznikami (`VINF`), XeLDA rozróżnia także czasowniki modalne (`VMOD` — *chcieć, móc, musieć, powinien*) oraz czasowniki niewłaściwe (`QVRB` – *można, widać, brakować* i z niewiadomych przyczyn *chcieć*), gdy tymczasem brak czasowników *trzeba, wolno*. Czasowniki modalne są etykietowane podobnie jak pozostałe czasowniki, czasowniki niewłaściwe — w sposób skrótowy. Przykładowe opisy (dla form *można, brakuje, chce*; etykieta `Inv` charakteryzująca wyraz nieodmienny pojawia się także przy opisie innych form):

można+`QVerb+Inv+QVRB`,
brakować+`QVerb+Imperf+Ind+Pres+QVRB`,
chcieć+`QVerb+Imperf+Ind+Pres+QVRB`,
chcieć+`Verb+Imperf+Ind+Pres+3P+Sg+VMOD`.

Czasowniki *być, zostać* oznaczane są jako pomocnicze — `VAUX`. Poza tym ich opis nie różni się od pozostałych czasowników. Uznanie *zostać* za czasownik pomocniczy wynika z możliwości tworzenia przy jego pomocy strony biernej *jest/został/zostanie namalowany*, jednak specjalna rola czasownika *być* jest znacznie bardziej wyraźna (czas przyszły złożony).

Dla specjalnych klas czasowników `VAUX, VMOD, QVRB` nie ma rozróżnienia form na `VFIN` i `VINF`; jedyną informacją odróżniającą formy osobowe od bezokoliczników jest tryb dla tych pierwszych i etykieta `Infinit` dla drugich. Tym bardziej dziwi, że w przypadku „zwykłych" czasowników, bezokoliczniki (pełny opis `Verb`+⟨*aspekt*⟩+`Infinit+VINF`) wyróżnione zostały jako oddzielna podgrupa; jest to informacja redundantna.

Imiesłowy traktowane są jako podgrupa leksemów czasownikowych (`Verb`). Pełny opis form imiesłowów przymiotnikowych to: `Verb`+⟨*aspekt*⟩+`Part`+⟨*czas*⟩+⟨*typ*⟩+⟨*rodzaj*⟩+⟨*liczba*⟩+⟨*przypadek*⟩+`VPRT`,
przy czym ⟨*typ*⟩ imiesłowu to: czynny (`Act`), bierny (`Pass1`) i przeszły (`Pass2`), zaś ⟨*rodzaj*⟩, ⟨*liczba*⟩, ⟨*przypadek*⟩ jak dla przymiotników. Nie jest jasne, czemu ma służyć umieszczanie informacji dotyczącej czasu. Pełny opis dla form imiesłowów przysłówkowych to: `Verb`+⟨*aspekt*⟩+`Gerund`+⟨*typ*⟩+`VGER`,
przy czym typ może być `Pres` bądź `Past`. Natomiast nie ma specjalnego oznaczenia dla odsłowników: są one etykietowane jako ⟨*bezokolicznik*⟩+`Noun`+⟨*rodzaj*⟩+⟨*liczba*⟩+⟨*przypadek*⟩+`NOUN`. Podobnie jak w SAM-ie, dla wszystkich tych form podawany jest aspekt czasownika, od którego zostały utworzone.

Ponadto XeLDA posiada specjalną etykietę `PartFut`, którą opatrywane są formy trzeciej osoby czasu przeszłego, mogące współtworzyć czas przyszły złożony: *będę/będziesz/będzie pił/skakała/płakało, będziemy/będziecie/będą czytali/pisały*.

Przymiotniki (`ADJ`) etykietowane są za pomocą przypadka (podobnie jak w LEM-ie, uwzględnany jest wołacz), liczby i rodzaju. Dowolny rodzaj oznaczany jest przez `MFN`, niemęskoosobowy przez `M23FN` itp. Ponadto formy skrócone opatrywane są etykietą *Brev*. Stopień równy jest domyślny, ponadto pojawiają się oznaczenia `Comp` dla stopnia wyższego i `Sup` dla najwyższego. Pełny opis dla przymiotnika w stopniu najwyższym (na przykładzie wyrazu *najlepsze*) jest następujący: `naj+Sup+dobry+Adj+Neut+Sg+NomAccVoc+ADJ`.
Podawanie informacji `naj+Sup` wydaje się pewną redundancją.

Formę podstawową przysłówka (`ADV`) stanowi odpowiadający mu przymiotnik (jeśli taki istnieje); ponadto podawana jest informacja o stopniu. Przykłady (*najjaśniej, oburącz*):

`naj+Sup+jasny+Adv+ADV`,
`oburącz+Adv+ADV`.

Liczebniki (NUM) etykietowane są za pomocą przypadka i rodzaju (XeLDA jako jedyny analizator uwzględnia wołacz dla liczebników). Ponadto liczebniki główne oznaczane są przez liczbę pojedynczą (Sg), zaś zbiorowe — mnogą (Pl), co nie wydaje się zbyt szczęśliwym rozwiązaniem. Dla przykładu podajemy opis formy *pięciorgu*:

pięć+Num+MFN+Pl+Dat+NUM.

Liczebniki porządkowe, posiadające odmianę przymiotnikową, opatrywane są etykietą NumOrd+⟨*rodzaj*⟩+⟨*liczba*⟩+⟨*przypadek*⟩+ADJ. XeLDA wyróżnia także przymiotniki liczebnikowe (*kilka, jeden, wiele*). Przykładowy opis takiej formy to: jeden+Adj+Num+M3+Sg+NomAccVoc+ADJ.

Kategoryzacja zaimków (Pron) ma charakter nieco odmienny niż w innych analizatorach. Klasa ta dzieli się na główne podklasy: zaimki osobowe — PPERS (*ja, wy, on, oni*), względne — PREL (*taki, który, niektóry*) oraz wyróżniony zaimek zwrotny (PREFL) *się*; pozostałe zaimki oznaczane są etykietą PRON. Jednak w tej ostatniej grupie wyróżnione są zaimki dzierżawcze (Poss — *mój, twój, swój ich, jego, jej*), wskazujące (Dem — *ten*) i pytajne (Interrog — *co, który*).

Zaproponowaną klasyfikację zaimków trudno uznać za spójną i uporządkowaną. Zaimki rzeczowne i przymiotne nie są w zasadzie wyróżniane, ale zaimki dzierżawcze „z natury rzeczy" pełnią rolę przymiotnikową, podobnie wyrazy *taki, który, niektóry, ten*, uznawane przez LEM-a za zaimki przymiotne, zaś przez SAM-a za przymiotniki. Ale już słowa *inny, każdy, tamtejszy* XeLDA traktuje jako przymiotniki. Z drugiej strony zaimki przysłowne *gdzie, jak, kiedy* opatrzone zostały etykietą Adv+APron+ADV, bez żadnej informacji o ich charakterze pytajnym. Forma *tu* została oznaczona przez Adv+ALoc+ADV, a więc nie została uznana za zaimek; formy typu *tam, wtedy* nie zostały zbadane. Wyrazy *co, kto, który, niektóry, taki, ony* przypisane zostały do dwóch klas zaimków: PREL i PRON, przy czym w tym drugim przypadku oznaczane są dodatkowo etykietą interrog, co w przypadku trzech ostatnich wydaje się nieporozumieniem (nie są one także zaimkami względnymi). Klasy zaimków względnych i pytajnych w zasadzie się pokrywają; wyjątek stanowi zaimek pytajny *czyj*. Tak więc klasa PREL zdaje się zawierać w klasie PRON, do której należą także zaimki *nikt, nic, wszyscy, wszystko* (bez dodatkowych etykiet).

Zaimek *się* opisywany jest na dwa sposoby. Po pierwsze, jako zaimek zwrotny opatrywany jest etykietą Pron+Refl+Acc/Gen+PREFL (podobnie *sobie, siebie* z odpowiednimi przypadkami), oraz drugą etykietą Pron+General+PRON. Brak natomiast etykietowania sugerowanego przez twórców pakietu się+Prt+Reciproc.

Wśród partykuł (oznaczanych przez Partcl+PRTCL) wyróżniane są partykuły pytajne (*czy, czyż, czyżby*) oznaczane przez Partcl+Interrog+PRTCL. Na koniec, podobnie jak w LEM-ie, dla przyimków (PREP) podawane są wymagane przez nie przypadki (np. przy+Prep+Loc+PREP).

Zauważmy, że powyższa taksonomia nie jest konsekwentna. Informacja o części mowy podawana jest zazwyczaj dwukrotnie, np. Adj⟨...⟩ADJ, czasami jednak któryś z opisów jest bardziej szczegółowy. Dla większości części mowy główną klasę opisuje pisana dużymi literami etykieta kończąca opis. I tak wśród rzeczowników (NOUN) wyróżnia się nazwy własne (Prop) i skróty (Abbr) zaś wśród przymiotników (ADJ) — liczebniki porządkowe (NumOrd) i część liczebników o odmianie przymiotnikowej (Adj+Num). Tymczasem czasowniki poklasyfikowane są całkiem odmiennie: klasa Verb zawiera formy osobowe (VFIN), bezokoliczniki (VINF), imiesłowy (VPRT i VGER) i odsłowniki (NOUN). Podobnie rzecz się ma ze złożoną klasyfikacją zaimków. Przyczyny tych różnic w tworzeniu hierarchii klas należy prawdopodobnie upatrywać w fakcie, że etykieta kończąca opis determinuje typ odmiany.

W poniższej tabeli przedstawiamy formy podstawowe leksemów podawane przez analizator dla poszczególnych części mowy.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów	odsłownik	przysłówek
forma podstawowa	mianownik liczba poj.	bezokolicznik	mianownik, l.poj., r. męski, (st. równy)	bezokolicznik	bezokolicznik	przymiotnik, (st. równy)

8.2 Wyniki analizy dla listy słów

8.2.1 Analiza form poprawnych

Spośród poprawnych słów j. polskiego wybranych do przetestowania, słownik analizatora XeLDA nie zawiera następujących: *gmin, gródź, karp, okoń, przysłowie, śledź, swąd, szerszeń, wielkanoc, żołądź, odprząc, mieść, pleść, srożyć, ugrząźć, ugrzęznąć, wyląc, wylęgnąć, wypełznąć, zaprząc, zaprzęgnąć, dwojaki, srodze, wręcz*. Dziwi obecność w słowniku słowa *odprzęgnąć* przy braku *zaprzęgnąć*. Ponadto program odrzuca następujące formy wyrazowe:

- *dłoniami*, gdy tymczasem forma *dłońmi* została przyjęta;
- podobnie *wstąg*, ale formę *wstęg* zaakceptowano;

- *awaryj* — odrzucona archaiczna forma dopełniacza liczby mnogiej, choć zaakceptowana została analogiczna forma *stacyj*;
- *łazęgów* — dopełniacz liczby mnogiej dla rodzaju męskoosobowego;
- Inne nierozpoznane formy rzeczownikowe to *pereł, powodzi, ciem, drzazdze*, a także *woźnice, pasmie*, co jest o tyle dziwne, że program zaproponował poprawne sugestie analizy, a odpowiednie leksemy znajdują się w słowniku;
- *pasem* — rozpoznany narzędnik słowa *pas* (a także nieodmiennego słowa *paso*); brak dopełniacza l.mnogiej wyrazu *pasmo*;
- *mielić, mielił* — uważane przez specjalistów za niepoprawne, lecz będące w użyciu; natomiast akceptowaną formą 1. os.cz.ter. czasownika *mleć* jest *mieli*, a nie uznawane za poprawne *miele*;
- *buź, buzi, sypialni* — brak dopełniacza liczby mnogiej, natomiast zaakceptowana została forma *sypialń*;
- Analizator ma problemy z bardziej skomplikowanymi (lub rzadszymi) formami czasownikowymi. I tak nie rozpoznano *pójdźmy, pełźli, pełźnie* (choć zaakceptowano *pełzli, pełznie*), *wlekę, wlekł, wlekła* (ale jest *wlokę, wlókł, wlokła*) oraz *odprzągł, odprzężmy*;
- *dwój, trój, krój* — nie rozpoznano rozkaźników od *dwoić, troić, kroić*;
- *mędrszy* odrzucone, jest natomiast *mądrzejszy*;
- zignorowane zostały przymiotniki w st. wyższym *dzikszy, sroższy, weselszy* oraz przysłówki *goręcej, srożej, węziej*;
- *gzowi* — brak dopełniacza rzeczownika *giez* (jest rozpatrywana forma przymiotnikowa);
- brak przymiotników odzwierzęcych *psi, muszy*, ale są *ćmy, gzowy, dziczy, karpi, lwi, muli*. Zaskakuje fakt, że brak rzeczownika *karp*, a jest odpowiadający mu przymiotnik; rzeczownik *pies* oczywiście jest.

Ponadto forma *woźnych* wiązana jest wyłącznie z leksemem rzeczownikowym *woźny*, brak leksemu *woźna*; *moczymorda* powinien być rodzaju męskoosobowego (jak *jak mężczyzna, woźnica*), a nie żeńskiego; wyrazowi *Białystok* została błędnie przypisana liczba mnoga, czasowniki *biegnąć, pełznąć, powodzić, wlec* powinny mieć aspekt niedokonany, w związku z czym formy *pełznie, powodzi, wlecze* występują w czasie teraźniejszym. Ponadto wyrazy *zły, żaden* z niewiadomych względów zostały zaliczone do klasy `Adj+Num`.

Błędnie zostały zinterpretowane liczebniki: *półtora* niepoprawnie potraktowany jako rzeczownik, zaś *wiele, wielu* ze źle opisanymi przypadkami: *wiele* występuje wyłącznie w mianowniku i bierniku dla rodzaju niemęskoosobowego, zaś wielu we wszystkich przypadkach dla rodzaju męskoosobowego (jak to zostało przedstawione), lecz także w dopełniaczu, celowniku i miejscowniku dla rodzaju niemęskoosobowego. Takie same błędy wykryto w odmianie form *kilka, kilku*.

Program wskazał także niepoprawne lub nadmiarowe interpretacje następujących form:

- *woła* — niepoprawny dopełniacz i biernik od wyrazu *wół*;
- *trąd, głąb* — nadmiarowy rodzaj męskozwierzęcy (poza męskorzeczowym);
- *wrogu, kwiatku* — nadmiarowy dopełniacz i biernik l.poj;
- *mózgu, pociągu, domku, kwiatka, wylęgu* — nadmiarowy biernik l.mnogiej, gdy tymczasem dla rodzaju męskorzeczowego (`M3`) biernik jest równy mianownikowi;
- *pawiu, źrebcu* — nadmiarowy biernik l.poj;
- *zimorodek* — nadmiarowy biernik l.poj., gdy tymczasem dla rodzaju męskozwierzęcego (`M2`) biernik jest równy dopełniaczowi;
- *woźnicy* — nadmiarowy mianownik liczby mnogiej (uznany również za formę deprecjatywną), za to odrzucono prawidłową formę *woźnice*;
- *uczenia* — nadmiarowa liczba mnoga dla odsłownika *uczenie*;
- *biegli* — nadmiarowy imiesłów przymiotnikowy przeszły dla czasownika *biegnąć*;
- *dęba* — uznane za formę podst. rzeczownika,
- *głąb* — zinterpretowane (również) jako przysłówek.

Jak wiadomo, czas przyszły złożony formuje się jedynie dla czasowników niedokonanych. XeLDA nie przestrzega tej reguły. I tak przeznaczoną do tego celu etykietą `Verb+Perf+Ind+PartFut+`⟨*rodzaj*⟩`+`⟨*liczba*⟩`+VFIN` zostały opatrzone formy *postawiły, postąpiły, potwierdziły, przekonała, przestały, uruchomiły, wprowadziły, zdezaktulizowały, zdziczały.* Poprawnie etykietowane są natomiast wyrazy *bogaciła, czytali, działa, myli, nakładały, pił, przewidywało, rósł, składali, tył*, brak natomiast takich opisów dla form *biegli, bódł, bodła, ciął, jadł, mełł, miał, mieli, mogli, mogła, wiódł, wiodła, piekło.* Specyficzny przypadek stanowi czasownik *być*: chociaż ma aspekt niedokonany, to jednak frazy *będzie był* i *będą były* są niepoprawne; jednak XeLDA opatruje etykietą `PartFut` wyrazy *był, były.*

XeLDA opatruje etykietą `NUM` zarówno liczebniki główne, jak i liczby. Program rozpoznaje także tak liczby całkowite, jak i ułamkowe (rozdzielone zarówno kropką — 0.75, jak i przecinkiem — 1,6); opatrywane są one etykietą `+Num+Dig+Card+NUM`. Zauważmy, że *27.4.1987* zinterpretowane zostało zarówno jako liczba (`27.4.1987+Num+Dig+Card+NUM`), jak i data (`27.4.1987+Num+Dig+Card+DATE`), gdy tymczasem *17:12* opatrzone zostało jedną etykietą `TIME`. Liczba, po której występuje kropka, etykietowana jest przez `Num+Dig+Ord+ORD`.

Ponadto wybrane słowa obce etykietowane są przez `FRGN` (*pro* tak, ale *Prix* już nie), a niektóre słowa mające wystąpienia idiomatyczne oznaczane są jako `PHRAS` (*kolei, przykład, angielska, polska, dobre, prawo, nowo, razie, wprost*). Także formy typu *prostu, polsku*, które występują tylko po przyimku *po*, opatrzone są taką etykietą. Wykrzyknik *aha* został opatrzony etykietą `Excl+EXCL`.

XeLDA dysponuje dość bogatą klasyfikacją dla symboli. Znaki przestankowe (`Punct`) kończące zdania '.', '?', '!' etykietowane są przez `Punct+Final+SENT`, ';' opatrywany jest oznaczeniem `SENT`, tyle że bez `Final`, ',' etykietowany jest przez `Punct+Comma+CM`, a pozostałe znaki przestankowe — ':', '-', '(', ')', '...' — przez `Punct+Misc+PUNCT`. Niestety, dla wielu symboli, „sklejenie" ich z wyrazem uniemożliwia jego interpretację. Dotyczy to w szczególności braku spacji po kropce kończącej zdanie. Krańcowym przykładem jest napis *.W*, z którym radzi sobie większość analizatorów (poza PoMor'em), nawet te, które nie są w stanie zinterpretować napisu *kandydatów.Łatwiej.* Innymi słowy, znaki przestankowe są rozpoznawane (i przy okazji „odklejane" od słowa) tylko wówczas, gdy z odpowiedniej strony znajduje się spacja bądź karetka. Na przykład myślnik/łącznik nie może „przyklejać się" z żadnej strony. Kuriozalnym tego przykładem jest nierozpoznawanie liczb ujemnych, opatrzonych za to jedną jedyną, i to w zasadzie poprawną, sugestią (`-15+Num+Dig+Expr+guessed+NUM`).

Analizator wyposażony jest w bogaty słownik nazw własnych, zwłaszcza imion, krajów, miast, a także wybranych nazwisk (nie dotyczy to jednak nazwiska *Grzymała*). Jak już wspominaliśmy, opatrzone są one standardową etykietą końcową `NOUN`; jednak początkową etykietę stanowi `Prop`, zaś forma podstawowa leksemu rozpoczynana jest wielką literą. Jednak większość nazw mających odmianę przymiotnikową (*Kowalski, Góralska, Puławska, Stołeczna*) traktowana jest jak zwykłe przymiotniki (pisane małą literą). Jako jedyny z badanych analizatorów XeLDA rozpoznaje złożoną nazwę własną *Stany Zjednoczone* (choć już nie *Zielona Góra*), niepotrzebnie tylko opatruje takim opisem zarówno wyraz *Stany*, jak i *Zjednoczone*; wystarczyłby jeden z nich.

Bogaty jest także zestaw rozpoznawanych skrótów, nie są one jednak w żaden sposób rozkodowywane. Skróty nazw własnych pisane wielkimi literami (*PCK, PKP, PTTK, NATO, USA*) opisywane są przez `Prop+Abbr+InvNum+InvCase+NOUN`. Ciekawe, że w ten sposób zinterpretowany został także skrót *MHz.* Inne skróty pisane bez kropki z małej litery (mgr, dr, mg, mld) mają natomiast etykietę `Noun+Unit+NOUN`, co w przypadku pierwszych dwóch ewidentnie nie jest poprawne. Warto zwrócić uwagę na poprawne zinterpretowanie formy *ONZ-owski* jako `Adj+M3+Sg+NomAccVoc+ADJ`. Natomiast z niewiadomych przyczyn skróty zakończone kropką (ds., hab., inż., proc., tzw., ww., itp, itd.) mają aż trzy interpretacje: `Adj+Abbr+ADJ, Noun+Abbr+InvNum+InvCase+NOUN` i `Adv+Abbr+ADV`. Ciąg znaków *proc.* ma tylko jedną interpretację; może to spowodować kłopoty w wypadku wystąpienia dopełniacza l.mn. od słowa *proca* na końcu zdania: rozpoznany zostanie skrót, a koniec zdania będzie przeoczony. Podobnie rzecz się ma z liczbą kończącą zdanie. Natomiast dokładnie odwrotna sytuacja dotyczy nierozpoznawanego skrótu *ul.*: odnaleziony zostanie rzeczownik *ul* wraz z fikcyjnym końcem zdania. Inicjały (*A.*) oznaczane jako *Init* nie były testowane, podobnie rzecz się ma ze skrótami typu *USD* opatrywanymi etykietą *Curr*; nie jest w ten sposób oznaczany symbol $.

Program zdaje się interpretować ruchome klityki czasownikowe dołączane nie do czasownika, tj. *-(e)m, -(e)ś, -(e)śmy* itd., w każdym razie w przypadku spójników, jako że zostały rozpoznane formy *alem, aleś, jakem, jakeś, żebyś* (etykietowane `Conj=być+`⟨*osoba*⟩`+`⟨*liczba*⟩`+CJ/AUX`). Dotyczy to w szczególności wszystkich form przymiotnikowych zakończonych na **-im, -ym** — *całym, ochronnym, odpowiednim, ostatnim* (etykietowanych `Adj+`⟨*rodzaj*⟩`+Sg+Nom+być+1P+Sg+ADJ`), oczywiście z wyjątkiem form liczebnikowych *jednym, drugim* (etykietowanych przez `Adj+Num`), a także formy *żadnym*, błędnie zaliczonej do tejże klasy. Jednak forma *pięknie* została odrzucona, podobnie *ażebym, idźże, zróbże.* Dotyczy to także partykuły *czyżbyś* i zaimka *coś*, choć przewidziane zostały etykiety `PT/AUX` i `PR/AUX`. Nie są także akceptowane formy przyimków ze skróconą formą zaimka, jak np. *doń, zeń.* Inne formy należące do powyższych klas nie były testowane.

Analizator nie interpretuje adresów elektronicznych, nie została także rozpoznana forma *XIV-wieczny*, przy czym w ogóle nie są interpretowane liczby rzymskie. Forma *6-letni* także nie uzyskała poprawnej interpretacji, jednak opatrzona została poprawną sugestią. Natomiast etykieta `SYMB` służąca do opisu symboli pojawiła się

jedynie dla znaku '@' (pełna etykieta to `+Spec+SYMB`), ale już nie dla '$', '&' i '*'. Sugerowana interpretacja tych symboli (`open+EXCL`, `open+NOUN`, `open+ADV`) całkowicie mija się z rzeczywistością.

W taksonomii XeLD-y uwzględniona została także etykieta `CMPND`, mająca w założeniu opisywać przymiotniki złożone z dwóch części (*północno-zachodni*). Jedyny przypadek, w jakim klasa taka została rozpoznana, stanowiła forma *kujawsko-* opatrzona etykietą `Compound+First+CMPND`, jednak już następująca po niej forma *-pomorskim* (w oryginalnym tekście formy te były rozdzielone karetką) nie miała interpretacji. Podobnie zapisana forma *filtro-⟨CR⟩-wentylacyjne* została już jednak zignorowana. W typowym przypadku, kiedy forma taka pisana jest bez żadnych spacji, analizator zdaje się niezdolny do rozbicia jej na części składowe (testowane przez nas formy to *biało-czerwona, polsko-francuski, obronno-ochronne*).

Dla wszystkich nierozpoznanych wyrazów XeLDA formułuje sugestie odmiany. Są one mniej liczne niż w przypadku SAM-a, lecz generowanie ich w sposób automatyczny w przypadku braku interpretacji jest niewątpliwą zaletą. Niestety, rezultatem jest znacznie mniejsza liczba „trafień". W przypadku wybranych neologizmów udało się zasugerować poprawne analizy rzeczowników (*metnik, kluczówka, wiertopięt, pracowstręt*), lecz całkowicie chybione zostały propozycje odmiany czasowników (*komputeruje, piotruje*). Przyczyną jest wyjątkowo prymitywny algorytm tworzenia takich sugestii: rozważa on jedynie ostatnią samogłoskę (ew. końcówkę pustą), bez żadnych wymian wewnątrztematowych. W efekcie pojawiają się sugestie całkowcie obce językowi polskiemu. Przede wszystkim, po znakach *ć, ń, ś, ź* nie mogą pojawiać się samogłoski; następuje tu wymiana na *ci, ni, si, zi*, po których dopiero pojawia się samogłoska. Tak więc propozycje *buźa, buźo, mielića, mielićo, dońa, dońo* są z natury rzeczy niepoprawne. Natomiast samogłoska *y* nie może się pojawiać po niektórych spółgłoskach, co dyskwalifikuje sugestie *miely, biennaly, piotrujy, aky, szczecińsky*.

8.2.2 Analiza form niepoprawnych

Analizator XeLDA odrzuca większość form niepoprawnych. Podobnie jednak jak SAM, ma tendencje do nadmiarowych odmian rzeczownikowych. Dotyczy to końcówek *-a, -u* (*trąda, groba, domka, kutru, skuteru, biesu*), ale bez błędów w wymianie wewnątrztematowej (**trędu, *swądu, *gróbu, *kwiateka, *domeku, *kutru, *skuteru*). Podobną przyczynę ma błędna interpretacja pochodzących z plików tekstowych słów *czeka, połowa, ubiega, wynika* jako form dopełniacza i biernika leksemów *czek, połów, ubieg, wynik*. Pozostałe zaakceptowane formy niepoprawne to: *ucznie, żołądzie, giezem, biesu, księg, idej, wylągu, weźmij, uczoni, dniowy*.

Jeśli w słowniku występuje przymiotnik, to przysłówek jest prawdopodobnie tworzony od niego automatycznie. Prowadzi to do interpretowania wyrazów *całodziennie* oraz *okazało, stowarzyszenie* (z pliku `Publicystyka`) jako pochodnych przymiotników *całodzienny, okazały, stowarzyszony*, odpowiednio. Dotyczy to także błędnego wiązania poprawnych przysłówków *prawie, właśnie* z przymiotnikami *prawy, własny*; od przymiotnika *prawy* pochodzi przysłówek *prawo [postępować]*, który zresztą także posiada taką interpretację.

8.3 Statystyka analizy plików tekstowych

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach. W pliku `kraj` nie rozpoznano 8 form wyrazowych, w tym 1(6) nazw własnych i 2 liczb. W pliku `kultura` 5 nazw własnych, z czego dwie to słowo obce *Prix*. W pliku `świat` nie rozpoznano 14 form, w tym 12 nazw własnych i 2 skrótów nazw własnych. Natomiast w pliku `publicystyka` nie rozpoznano 21 form, w tym 5(9) nazw własnych i 5(9) skrótów (skrót *ul.* uznany został za rzeczownik kończący zdanie). Liczby w nawiasach oznaczają ogół nazw własnych i skrótów, w tym przymiotniki pochodzące od nazw własnych oraz skróty nazw własnych. Ponadto zdarzały się przypadki (2, 3, 0, 8 odpowiednio), w których nazwa własna została rozpoznana jako inny wyraz (zazwyczaj przymiotnik). Brak także analiz dla następujących wyrazów: *biennale, odrzwia, strzelec*.

Ponadto wykryto poważne błędy w algorytmie odmiany czy też metodzie reprezentacji. Dotyczy to przede wszystkim zaimków. Po pierwsze, zaimki osobowe występujące w 1 osobie *on, ona, ono, one, oni* mają jednoznacznie określony rodzaj (nie dotyczy to zaimków w osobie 1 i 2 *ja, ty, my, wy*): *on* — męski, *ona* — żeński, *ono* — nijaki, *one* — niemęskoosobowy, *oni* — męskoosobowy. Tymczasem XeLDA przypisuje im różne rodzaje w zależności od analizowanej formy. Ponadto wykryto następujące błędy odmiany:

- *nim* — brak celownika l.mnogiej leksemów *oni, one* (fraza *ku nim*);
- *niemu* — brak celownika l.pojedynczej leksemu *ono*, choć *on* został uwzględniony;
- *nie* — brak biernika l.pojedynczej leksemu *ono*;
- *wszystkim* — nadmiarowy narzędnik dla rodzaju męskoosobowego, brak rodzaju nijakiego dla celownika;
- *swych* — brak dopełniacza l.mnogiej leksemu *swój* (dla dowolnego rodzaju);
- *swoje* — brak mianownika, biernika i wołacza leksemu *swój* l.mnogiej dla rodzaju żeńskiego i nijakiego (są rozważane jedynie dla rodzaju Meskozwierzęcegi i męskorzeczowego).

Podobne błędy jak w odmianie zaimka dzierżawczego *swój* mogą się pojawiać w przypadku zaimków *mój, twój*, odpowiednie formy nie zostały jednak przetestowane.

Oprócz tego forma *to* wiązana jest jedynie z zaimkiem *ten*: brak więc zarówno spójnika/partykuły, jak i zaimka rzeczownego *to*. Z kolei *im* jest klasyfikowane prawidłowo jako pochodna *oni, one*, ale za to brak spójnika.

Błędy odmiany wykryto także dla wielu rzeczowników. I tak, nadmiarowy biernik przypisano formom *kierunku, końca, mieszka, półmroku, roku, wyniku* wyrazów rodzaju męskorzeczowego (`M3`), dla którego biernik jest równy mianownikowi, a także formom *jak, pułkownik, stańczyk* wyrazów rodzaju męskoosobowego bądź męskozwierzęcego, dla których biernik jest równy dopełniaczowi, a nie mianownikowi. Podobny błąd popełniono umieszczając nadmiarowy biernik l.mnogiej w opisie formy rzeczowników rodzaju męskoosobowego *obywatele, właściciele* i opuszczając go w opisie form *obywateli, właścicieli*, choć wyrazy o podobnej odmianie *gospodarze, funkcjonarusze* opisane są poprawnie, a biernik dla rodzaju męskoosobowego (także l.mnogiej) zawsze jest równy dopełniaczowi. Zbędnym tak dopełniaczem, jak i biernikiem, został opatrzony wyraz *Targówku*.

Jedyny wyjątek od reguły zrównującej biernik rzeczowników męskorzeczowych z mianownikiem stanowią niektóre nazwy własne, które w wyniku swoistej antropomorfizacji odmieniane są jak wyrazy męskoosobowe (forma *Panasonica* została poprawnie uznana zarówno za dopełnacz, jak i biernik).

Z kolei brak dopełniacza liczby pojedynczej dla form wyrazów rodzaju męskorzeczowego *cyklu, kraju, masażu*, dopełniacza liczby mnogiej dla wyrazu rodzaju żeńskiego *budowli*, mianownika, wołacza liczby mnogiej form wyrazów rodzaju męskoosobowego *obrońcy, przywódcy* (w l.mnogiej zawsze równego mianownikowi); biernika i wołacza liczby mnogiej formy *części* (gdy tymczasem formy *ludności, gotowości* opatrywane są takim opisem, na dodatek nadmiarowo, gdyż nie występują w liczbie mnogiej). W końcu forma *Maziarz* została uznana za biernik zamiast mianownika.

Następnie wszystkim formom przymiotnikowym zakończonym na **-im, -ym** — *całym, ochronnym, odpowiednim, ostatnim* przypisano nadmiarowy narzędnik liczby mnogiej rodzaju męskoosobowego, zaś w opisie tych zakończonych na **-ymi** — *fabrycznymi, mieszkalnymi, oddanymi, potężnymi, prywatnymi* — odpowiadających narzędnikowi liczby mnogiej dowolnego rodzaju — opuszczono rodzaj męskoosobowy.

Kolejny problem związany jest z akceptowaniem liczby mnogiej dla odsłowników; powoduje to uznanie za formy gerundialne takich wyrazów jak *ograniczeń, pięć, poleceń, pomieszczeń, rozstrzygnięć, ukryć*. Następujące wyrazy nie zostały uznane za rzeczowniki: *bombardowanie, notowanie, ograniczenie, polecenie, rozstrzygnięcie, uregulowanie*, lub też zostały tak oznakowane nadmiarowo: *bycie, poznanie, przeprowadzenie*.

Inne rzeczowniki, dla których uwzględniono liczbę mnogą, choć zazwyczaj uznawane są za wyrazy wyłącznie występujące w liczbie pojedynczej, to *biotechnologia, dezaktywacja, gotowość, leczo (tych lecz), ludność, obrona, psychologia, sprawiedliwość*. Wątpliwości budzi także przypisywanie liczby mnogiej nazwom miast (formy *Bolonii, Gdański, Hagi*), regionów (*Śląska*) i krajów (*Finlandii, Jugosławii, Polski, Serbii, Szwajcarii, Szwecji*). Istnieje także wyraźna nadmiarowość w dokładaniu form rzeczownikowych do przymiotników. Jeśli dopuścić nawet istnienie takich rzeczowników, to mają one odmianę przymiotnikową. Tymczasem przysłówki *cicho, dawno, nowo, wolno* zinterpretowane zostały dodatkowo jako wołacz l.pojedynczej wyrazów rodz. żeńskiego *cicha, dawna, nowa, wolna*.

Podobne wątpliwości budzą rzeczowniki rodzaju żeńskiego: *dnia (tej dni), ta dęba, ta dwa, młodsza (tej młodszy), można, trza (tej trzy)*; rodzaju męskoosobowego: *czysta, mieszkaniowiec (tego mieszakniowa, te mieszkaniowy!), znany*; rodzaju męskorzeczowego: *oburącz, oprócz* (etykietowane przez biernik, bez dopełniacza); rodzaju nijakiego: *nowo, paso (tym pasem), właśnie*. Ponadto wyraz *dużo* uzyskał interpretację (poza przysłówkową): `+Adj+Num+MFN+InvNum+InvCase+ADJ`. Podobny błąd to uznanie za przysłówek słowa *wobec*.

Oprócz tego błędnie opatrzono aspektem niedokonanym czasowniki *przekonać, przesłuchać*, zaś dokonanym czasownik *uczestniczyć*. Co gorsza, formie *zechcą* przypisano czas teraźniejszy, chociaż czasownik *zechcieć* został poprawnie opatrzony aspektem dokonanym. Co więcej, czasownikom *skutkować, ustępować, znać* nadmiarowo przyznano aspekt dokonany (poza poprawnym niedokonanym).

Ponadto wyrazowi *wojewoda* błędnie przypisano rodzaj żeński zamiast męskoosobowego, w związku z czym forma *wojewody* została nadmiarowo potraktowana jako mianownik, biernik i wołacz liczby mnogiej; wyrazowi *rok* nadmiarowo przypisano rodzaj męskoosobowy; *zwane* zostało zbędnie uznane za imiesłów od *zwiać*, *czterdziestotysięczny* opatrzony został etykietą `Adj` zamiast `NumOrd`, formę *została* zbędnie potraktowano jako imiesłów przymiotnikowy przeszły; *Panasonic, Targówek* są to nazwy własne, a nie zwykłe rzeczowniki, gdy tymczasem *Rzeczpospolita* jest raczej wyrazem pospolitym, a nie nazwą własną, choć używanym w ściśle określonym kontekście.

Opatrywanie przyimków wymaganymi przez nie przypadkami jest niewątpliwie korzystną regułą, szczególnie przydatną przy procesie dezambiguacji (procedura dostępna w pakiecie XeLDA), jednak i tutaj nie ustrzeżono się wielu błędów. I tak wymagania dopełniacza zabrakło w opisie przyimka *z* i w szczególnych wypadkach *za* (*za dawnych czasów, za komuny*), zaś biernika — w opisie przyimków *miedzy, o, po, ponad, pod, przed* i w szczególnych wypadkach *mimo* (*mimo to, mimo wszystko*). Wyjątek stanowi przyimek *jako*, który nie wymaga żadnego konkretnego przypadka (*jako kobieta uważam, jako kobiecie nie wypada mi, traktuj mnie jako przyjaciela*).

Powyższa analiza przekłada się na następujące wyniki statystyczne:

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
$\|S\|$	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	8	5	14	21	48
errors(S)	31	19	38	55	143
generated	472	121	308	963	1864
F-generated	318	94	230	678	1320
less	14	6	17	63	100
non-exist	8	2	9	14	33
more	36	11	22	99	168
correct	411	101	257	778	1547
F-correct	266	78	187	513	1044
error-rate	0.073	0.176	0.139	0.067	0.088
less-error-rate	0.030	0.049	0.055	0.065	0.054
more-error rate	0.113	0.117	0.072	0.103	0.090
precision	0.870	0.835	0.834	0.808	0.830
F-precision	0.836	0.830	0.813	0.757	0.791
average-#correct	0.986	0.981	0.990	0.974	0.980

8.4 Wnioski

Analizator XeLDA charakteryzuje się bogatym słownikiem, przy tym daje się bez trudu zauważyć, że program rozpoznał większość słów w wybranych w sposób przypadkowy tekstach prasowych, większe trudności mając ze specjalnie zestawionym zbiorem słów testowych. Ponieważ braki w słowniku dotyczą głównie słów rzadkich, archaicznych, można mieć wrażenie, że był on konstruowany na podstawie zbioru rzeczywistych tekstów współczesnych, co jest niewątpliwą zaletą. W ten sposób tłumaczyć można także stosunkowo bogaty zestaw popularnych nazw własnych i skrótów. Dobre wrażenie psuje jednak występowanie w słowniku wielu nieistniejących leksemów, powiązanych z rzeczywistymi formami innych wyrazów.

Niestety, także wiele błędów odmiany, jakie udało się wykryć podczas testów, jest poważną wadą tego analizatora.

Zaletą XeLD-y jest także idea sugerowania, podczas standardowej analizy, interpretacji dla nieznanych wyrazów. Jednak niska skuteczność tej procedury zmniejsza szansę posługiwania się nią w praktyce.

9 AMOR

AMOR jest analizatorem dostępnym spod DOS-u; litery polskie są zapisywane w kodzie latin2. Wywołanie programu bez parametrów oznacza pracę w trybie interakcyjnym — na ekranie pojawia się analiza słów wprowadzonych z klawiatury. Program można także wywołać z dwoma parametrami: pierwszy oznacza plik źródłowy, drugi — wynikowy.

9.1 Oznaczenia PoS

Podstawą oznaczeń części mowy jest klasyfikacja Saloniego (por. Saloni i Świdziński (1998)). Analizator AMOR rozróżnia następujące części mowy: rzeczownik, czasownik, przymiotnik, liczebnik (główny i zbiorowy), przysłówek, imiesłów (przymiotnikowy czynny, przysłówkowy współczesny i uprzedni) oraz dodatkowo zaimek. Liczebniki porządkowe oraz imiesłowy przymiotnikowe bierne etykietowane są jako przymiotniki, zaś odsłowniki — jako rzeczowniki rodzaju nijakiego. Poza przysłówkami, wśród wyrazów nieodmiennych wyróżniane są przyimki, spójniki, partykuły i wykrzykniki.

Rzeczowniki etykietowane są typem deklinacji (podawanym w nawiasie) oraz przypadkiem, liczbą i rodzajem (rozróżniane są rodzaje `żeński, nijaki, męskoosobowy, męskozwierzęcy, męskorzeczowy`).

Czasowniki etykietowane są typem koniugacji (podawanym w nawiasie) oraz aspektem (`dk, ndk`). Formy osobowe oznaczane są ponadto przez osobę, liczbę i czas (przeszły oraz osobno teraźniejszy i przyszły), zaś formy czasu przeszłego dodatkowo przez rodzaj (`męski, żeński, nijaki` dla liczby pojedyńczej oraz `męskoosobowy, niemęskoosobowy` dla liczby mnogiej). Tryb oznajmujący jako domyślny nie jest jawnie oznaczany; pojawiają

się etykiety trybu rozkazującego i przypuszczającego. Bezokoliczniki oznaczane są etykietą `bezokolicznik` oraz `czasownik`, co można uznać za informację redundantną, zwłaszcza że formy osobowe i bezosobowe takiej etykiety nie posiadają. Wśród czasowników wyróżnione są czasowniki niewłaściwe (np. *można, widać, trzeba*), jednak brak czasownika *wolno*.

Przymiotniki etykietowane są typem deklinacji (podawanym w nawiasie) oraz przypadkiem (przy czym wołacz, zawsze tożsamy z mianownikiem, jest ignorowany), liczbą i rodzajem (rozróżniane są rodzaje `żeński, nijaki, męski, męskorzeczowy, męskożywotny` dla liczby pojedyńczej oraz `męskoosobowy, niemęskoosobowy` dla liczby mnogiej). Brak oznaczenia rodzaju oznacza rzecz jasna jego dowolność. Stopień przymiotnika nie jest zaznaczany, zaś forma podstawowa występuje w tym samym stopniu, co forma analizowana. Natomiast stopień przysłówka jest podawany, jednak forma podstawowa czasem jest podana w stopniu równym, a czasem jest tożsama z formą analizowaną.

Liczebniki główne etykietowane są informacją o części mowy (`liczebnik`) oraz przypadkiem i rodzajem. Liczebniki zbiorowe oznaczane są częścią mowy (`liczebnik zbiorowy`) oraz przypadkiem. W obu wypadkach, formę podstawową stanowi mianownik w rodzaju niemęskoosobowym liczebnika głównego (*trzy, pięć*).

Podstawową klasę zaimków stanowią zaimki osobowe (*ja, ty, on*); są one etykietowane za pomocą przypadka, liczby i rodzaju tak jak rzeczowniki i odatkowo oznaczane przez osobę. Pomocniczo jako zaimki oznaczane są inne zaimki rzeczowne, uznane za podklasę rzeczowników (np. *kto, co, nikt*) oraz przymiotne, uznane za podklasę przymiotników (*mój, swój, ten*) — etykieta `(zaimI) przymiotnik`.

Większość etykiet podawana jest pełnymi słowami.

W poniższej tabeli przedstawiamy formy podstawowe podawane przez analizator dla poszczególnych części mowy.

PoS	rzeczownik	czasownik	przymiotnik	imiesłów przysł.	imiesłów przym.
forma podstawowa	mianownik liczba poj.	bezokolicznik	mianownik, l.poj.,r. męski	bezokolicznik	mianownik, l.poj.,r. męski

Jak widać, imiesłowy przymiotnikowe (czynne) nie są wiązane z czasownikiem, od którego pochodzą.

9.2 Wyniki analizy dla listy słów

9.2.1 Analiza form poprawnych

W słowniku brak jest najprawdopodobniej leksemów *dwunóg, ciacho, czyż(N), gazda, grot, grodź, moczymorda, muł, wieść(N), bość, mierznąć, powodzić, ugrząźć, wieścić, żółcić, długodzioby*, w tym także słów archaicznych *imienie, liście, pleć, prześl, rośnia, źrebiec, mędrszy, cudzoziemczy*, co jest przyczyną braku analizy dla różnorakich ich form. Brak także przysłówka *prawo*, przysłówka/partykuły *wręcz* oraz spójnika/partykuły *niczym*.

Ponadto wykryte zostały następujące błędy i osobliwości:

- *lata* — brak powiązania z rzeczownikiem rok (lub rzeczownika *plurale tantum lata*, tak jak w przypadku *człowiek — ludzie*);
- *mężczyźnie* — brak miejscownika liczby pojedynczej;
- *dłoniami* — zostało odrzucone, jedyną akceptowaną formą narzędnika l.mn. jest *dłońmi*;
- *wstąg* — także zostało odrzucone, jedyną akcpetowaną formą dopełniacza l.mn. jest *wstęg*;
- *pasmie, pasem* — nie są uznawane za formy rzeczownika *pasmo*;
- *woźnice* — zostało uznane za niepoprawne, za to błędnie przyjęto za mianownik i wołacz liczby mnogiej *woźnicy*;
- *przysłów* — zostało odrzucone, natomiast akceptowany jest błędny dopełniacz l.mn. *przysłowi*;
- *łazęgów* — odrzucony dopełniacz liczby mnogiej (akceptowana jest forma *łazęg*), choć wyraz jest etykietowany rodzajem męskoosobowym;
- *sypialni, alei* — brak dopełniacza liczby mnogiej; jedyne akceptowane formy to *sypialń, alej*;
- *awaryj, kniej* — odrzucone archaiczne formy dopełniacza liczby mnogiej, choć zaakceptowana została analogiczna forma *stacyj*;
- *płocie* — brak biernika liczby mnogiej dla rzeczownika *płoć*;

- *boi* — brak dopełniacza liczby mnogiej rzeczownika *boja*;
- *pierze* — brak wołacza;
- *działa* — brak wołacza liczby mnogiej rzeczownika *działo*;
- *pełźli, pełźnie* — formy odrzucone, akceptowane są jedynie *pełzli, pełznie*;
- *ugrzązł, wylągł* — formy odrzucone, chociaż zaakceptowane zostały formy rodzaju żeńskiego *ugrzęzła, wylęgła* oraz rzadsze *wylęgł* i niepoprawne *ugrzęzł*. Zignorowany został także rozkaźnik *ugrzęźnij*;
- *zaprzągł* potraktowany został jako forma od *zaprząc*, zaś błędne *zaprzęgł* — jako forma od *zaprzęgnąć*, gdy tymczasem żeńska forma *zaprzęgła* wiązana jest z oboma tymi czasownikami. Brak odpowiednich czasowników *odprząc, odprzęgać*.
- *wylęgła* — brak imiesłowu przymiotnikowego przeszłego (ew. przymiotnika);
- *gródź* — brak rozkaźnika dla *grodzić*, jest jedynie *grodź*;
- *mędrszy* odrzucone, jest natomiast *mądrzejszy*;
- *sroższy* — nie jest akceptowany, chociaż słownik zawiera wyraz *srogi*.
- *gorąco* — brak przysłówka we wszystkich stopniach, chociaż jest rzeczownik;
- *dziób* — brak rozkaźnika od *dziobać*;
- *nie* — jest formą zaimka osobowego *ono*, ale brak powiązania z zaimkiem *one*;
- brak rzadkich przymiotników odzwierzęcych *ćmy, gzowy, muli, dziczy*, ale są *karpi, lwi, psi*. Jest także *muszy*, jednak odrzucana jest forma *musi* (*musi przyjaciele*);
- brak rzadkiej odmiany *zębu* (roślina *koński ząb*);
- brak frazeologicznej formy *dęba* (*stawać dęba*);
- *żołędzi, grodzi, powodzi* — nadmiarowy biernik liczby mnogiej;
- *ojczyźnie* — nadmiarowy wołacz liczby pojedynczej;
- *pragnę, gaśnie* — błędnie przypisany aspekt dokonany, i na skutek tego czas przyszły;
- *głąb* ma przypisany rodzaj męskozwierzęcy, co nie odpowiada ani *głąbowi kapuścianemu*, ani zwrotowi *ty głąbie*. Ewidentnie brak także wyrazu rodzaju żeńskiego *głąb jeziora*;
- liczebnik *półtora* ma przypisany rodzaj niemęskoosobowy zamiast męskiego/nijakiego, oraz przypadek mianownik, gdy tymczasem występuje we wszystkich przypadkach.

Analizator rozpoznał wyraz *żebyś* jako formę spójnika z końcówką osobową *żebym*. Program posiada opcję umożliwiającą rozpoznawanie ruchomych klityk czasownikowych dołączonych nie do czasownika, ale jedynie tych nie wymagających dodania *e* (*-m, -ś, -śmy*, ale już nie(*-em, -eś, -eśmy*). Tak więc akceptowana jest forma *alem*, lecz odrzucona jest forma *jakem*. Ponadto *coś* zostało zinterpretowane jedynie jako „zwykły" zaimek rzeczowny.Nie są także analizowane formy przyimków ze skróconą formą zaimka, jak np. *doń, zeń*. Formy typu *prostu, polsku*, występujące wyłącznie po przyimku *po* potraktowane zostały jako specjalne formy przymiotnika (typ odmiany oznaczony `(ku)`) występujące wyłącznie w celowniku (jako analogia do *po swojemu*). Rozróżniane są też niektóre wykrzykniki (*aha, ciach, och*).

Natomiast wśród innych analizatorów wyróżnia AMOR fakt, że dla przymiotników *sypialny, ryżowy* nie jest rozważana męskoosobowa forma *sypialni, ryżowi*, w jakiej ewidentnie wyrazy te nie występują. Ponadto wołacz *Polsko* jest uwzględniany wyłącznie dla form pisanych z dużej litery.

Program ignoruje wszystkie znaki nie będące literami (nie są one nawet przepisywane na plik wynikowy zawierający wyniki analizy), w tym liczby, znaki przestankowe itp., przy czym wszystkie one traktowane są obligatoryjnie jako separatory słów. W ten sposób formy *biało-czerwona, polsko-francuski* zinterpretowane zostały jako para ⟨przysłówek, przymiotnik⟩. Jednak jeśli odpowiednia forma zakończona na **-o** nie jest przysłówkiem (np. we frazach *obronno-ochronne, filtro-wentylacyjne* z pliku *PUBLICYSTYKA*), jest ona ignorowana. Rzecz jasna uniemożliwia to także rozpoznanie form przymiotnikowych typu *ONZ-owski* (skrót *ONZ* jest akceptowany) czy *XIV-wieczny* (liczby rzymskie nie są rozpoznawane) oraz adresów elektronicznych.

W słowniku analizatora znajduje się wiele nazw własnych. Nie są one w żaden specjalny sposób etykietowane, jednak wyróżniane są poprzez zapisanie ich z dużej litery. Natomiast skróty wyróżniane są etykietą `skrot`. Jednak słownik zdaje się zawierać więcej skrótów popularnych nazw własnych (*PCK, PKP, NATO, USA*, choć *PTTK* już nie) niż skrótów wyrażeń złożonych z wyrazów pospolitych (*ds., tzw.*, ale już nie mgr, dr, hab., inż., płk., mld, np., ul., proc.). Ignorowanie znaków przestankowych utrudnia rozpoznawanie skrótów zakończonych kropką (ew. rozpoznawane są bez tej kropki). W szczególności dotyczy to skrótów *ul., proc.* interpretowanych jako odpowiednie formy rzeczownikowe.

Analizator nie rozpoznaje neologizmów tworzonych w sposób regularny ze znanych słów.

9.2.2 Analiza form niepoprawnych

Analizator AMOR odrzuca większość form niepoprawnych. Spośród słów umieszczonych na liście form niepoprawnych, wykryto tylko dwa wyjątki:

- *ugrzęzł* — forma osobowa czasu przeszłego czasownika *ugrzęznąć*, gdy tymczasem poprawna forma *ugrzązł* nie jest akceptowana;
- *dniowy*; analizator akceptuje także poprawną formę *dzienny*.

9.3 Statystyka analizy plików tekstowych

Dla omawianego analizatora uzyskano następujące wyniki statystyczne na badanych plikach. W pliku `kraj` nie rozpoznano 27 form wyrazowych, w tym 5 nazw własnych i 21 liczb. W pliku `kultura` nie rozpoznano 7 form, w tym 5 nazw własnych (3 z nich to słowa obcojęzyczne) i 2 liczb. W pliku `świat` nie rozpoznano 22 form, w tym 13 nazw własnych, 4 liczb i 1(5) skrótów. W pliku `publicystyka` nie rozpoznano 41 form, w tym 9 nazw własnych, 5(12) skrótów i 12 liczb (dwa skróty, *ul.* i *proc.* potraktowane zostały jako rzeczowniki). Jak poprzednio, liczby w nawiasach oznaczają ogół skrótów, w tym skróty nazw własnych. Ponadto zdarzały się przypadki (0, 4, 0, 5 odpowiednio), w których nazwa własna została rozpoznana jako wyraz pospolity.

Brak także analiz dla następujących wyrazów: *biennale, iberystyka, odrzwia, sanitariaty, podusić, wentylatorowa* oraz przysłówka/partykuły *równie*.

Zanalizowane zostały rzeczowniki o odmianie przymiotnikowej (etykieta `rzp`), w szczególności nazwiska (*Krasiński, Skłodowska*). Jednak nazwiska tożsame z przymiotnikiem pospolitym nie są w żaden sposób wyróżniane.

Dla rzeczownika *rok* w ogóle nie są uwzględniane formy liczby mnogiej — ani wiązane z *rok*, ani z *lata*. Wyraz *to* traktowany jest wyłącznie jako forma zaimka wskazującego *ten*. Brak zaimka rzeczownego *to* (odpowiednika *co*), a przede wszytkim spójnika *to*. Forma *wszystkim* powiązana została z zaimkiem przymiotnym *wszystek*; brak odniesienia do ziamków rzeczownych *wszystko, wszyscy* (które to formy także są wiązane z przymiotnikiem *wszystek*). Na dodatek wyrazy te nie występują w liczbie pojedynczej, więc zamiast miejscownika i narzędnika l.poj., w opisie powinien się znaleźć celownik l.mnogiej. Dotyczy to także przymiotnika *niektóre*, dla którego za formę podstawową uznano nieistniejącą formę *niektóry*, oraz nadmiarowo uwzględniono liczbę pojedynczą (por. Doroszewski i Kurkowska (1973)).

Natomiast zaimek/przymiotnik *któryś* występuje, zgodnie z Markowski (1999); Doroszewski i Kurkowska (1973), jedynie w liczbie,pojedynczej, więc nie powinien być wiązany z formą *których*.

Kolejny problem stanowią zaimki dzierżawcze. Zaimki o odmianie przymiotnikowej *mój, swój* są rozpoznawane, jednak zaimki *jej, jego, ich* traktowane są wyłącznie jako formy zaimków osobowych. Jest to zapewne poprawne z czysto morfologicznego punktu widzenia, jednak zaimki te pełnią również rolę zaimków dzierżawczych, a w szczególności łączą się z rzeczownikami we wszystkich przypadkach (np. *ich książką, ich książki*).

Ponieważ odsłowniki nie zostały wydzielone jako odrębna część mowy, nie są rozróżniane wypadki, w których dany wyraz jest zarówno rzeczownikiem, jak i odsłownikiem (np. *przyjęcie, uregulowanie, zainteresowanie*). Wyrazy takie występują zarówno w liczbie pojedynczej, jak i mnogiej, a także posiadają wołacz. Brak wołacza (nie tylko w przypadku wyrazów „gerundio-podobnych") liczby pojedynczej wystąpił dla słów *stowarzyszenie; dziennikarstwo, gotowości, ludności, prawo, roku, sprawiedliwości, ultimatum*, liczby mnogiej — dla słów *bombardowania, notowania, pomieszczenia, ukrycia, uregulowania, wyjścia, zalecenia; archiwa, dobra, dni, miejsca, państwa, podziemia, społeczeństwa, studia*.

Natomiast wyrazy będące wyłącznie odsłownikami nie mają liczby mnogiej (ani wołacza). Błędne rozpatrywanie liczby mnogiej wystąpiło dla słów *budowanie, przeprowadzenie, schronienie, utrzymanie, wznoszenie, zakwaterowanie, zarządzanie*. Inne wyrazy, dla których nadmiarowo została uwzględniona liczba mnoga, to *obrona, ochrona, psychologia*.

Dla wyrazów *odkażania, porozumienia, zachowania* można mieć wątpliwości, czy występują one w liczbie mnogiej. Jeśli jednak tak, to należy dla nich uwzględniać wołacz.

Ponadto nadmiarowy celownik został uwzględniony dla formy *go*, zaś nadmiarowy biernik — dla formy *kolei*.

Z drugiej strony, brak biernika w opisach form *twórców, mieszkańców*. Czasownikom *odwołać, uczestniczyć* błędnie przypisano aspekt. Nie jest dla nas także jasne, dlaczego formie liczebnikowej *jeden* przypisano wszystkie przypadki (bez wołacza).

Na koniec, została wykryta literówka w formie podstawowej słowa *szanse* — *szanse***a**.

Wyniki te przekładają się na następujące wyniki statystyczne:

	KRAJ	KULTURA	ŚWIAT	PUBLICYSTYKA	RAZEM
$\|S\|$	425	108	273	820	1626
$\|S_F\|$	304	83	202	615	1204
unknown	27	6	22	41	96
errors(S)	30	12	24	60	126
generated	436	107	272	848	1663
F-generated	319	81	205	637	1242
less	4	1	6	19	30
non-exist	0	0	1	1	2
more	5	0	2	16	23
correct	425	106	262	810	1603
F-correct	308	80	196	599	1183
error-rate	0.071	0.111	0.081	0.073	0.077
less-error-rate	0.013	0.0093	0.022	0.022	0.018
more-error rate	0.011	0.000	0.0074	0.019	0.014
precision	0.975	0.991	0.963	0.955	0.964
F-precision	0.966	0.988	0.956	0.940	0.952
average-#correct	1.068	1.039	1.044	1.040	1.048

9.4 Wnioski

AMOR pomyślany został jako „czysty" analizator morfologiczny — dlatego ignoruje liczby, znaki przestankowe i specjalne itp. Kłopot w tym, że znaki te w ogóle nie pojawiają się na pliku wynikowym, w związku z czym znajdująca się na nim informacja zostaje w sposób istotny zubużona. Tak więc korzystając z tego narzędzia należy dodatkowo zadbać o to, by znaki te nie zaginęły.

Program dostarcza dość bogatą informację morfosyntaktyczną, jednak występuje w niej pewien brak konsekwencji (np. stopniowanie przysłówków, ale już nie przymiotników).

Istotną zaletą omawianego analizatora jest fakt odrzucenia większości błędnych form wyrazowych. Mimo to jednak wykryto sporo błędów w odmianie wyrazów, co w sposób istotny obniża wartość bieżącej wersji programu.

10 Podsumowanie

W niniejszym opracowaniu omówiono i przebadano na jednolitym zestawie testów sześć analizatorów: Gram, PoMor, SAM, LEM, XeLDA i AMOR. Ponieważ jednak podejście do zadania analizy morfologicznej zastosowane w prezentowanych programach jest wyraźnie niejednolite, ich porównanie jest ewidentnie zadaniem trudnym, a wyniki przeprowadzonych przez nas testów porównawczych nie są do końca jednoznaczne.

Wszystkie wspomniane analizatory wyróżniają podstawowe części mowy: rzeczowniki, czasowniki, przymiotniki, przysłówki, liczebniki (główne). PoMor i SAM, stosując się do zasad zawartych w indeksie Tokarskiego (1993), identyfikują także imiesłowy (przymiotnikowe czynne, bierne i przeszłe, przysłówkowe współczesne i uprzednie), odsłowniki, zaimki rzeczowne. PoMor etykietuje także przyimki, gdy tymczasem SAM oznacza jedynie przyimki posiadające formy zakończone na **-e**. Pozostałe części mowy grupowane są w klasie *nieodmienne*. Z kolei AMOR rozróżnia imiesłowy przysłowne współczesne i uprzednie, a przymiotnikowe jedynie czynne. XeLDA rozróżnia kilka klas czasowników, podobnie jak LEM, zaś AMOR wyróżnia czasowniki niewłaściwe. Tak XeLDA, jak Gram i LEM, identyfikują wszystkie rodzaje imiesłowów. Najbogatszy zestaw części mowy, w tym części nieodmiennych, posiada analizator LEM; XeLDA, AMOR i Gram także rozróżniają wiele nieodmiennych części mowy.

Wszystkie analizatory podają rodzaj, przypadek i liczbę dla rzeczowników i przymiotników (PoMor i SAM w postaci zakodowanej), stopień dla przymiotników i przysłówków (przy czym SAM nie odróżnia stopnia wyższego od najwyższego, zaś AMOR stopniuje jedynie przysłówki); formę, aspekt, tryb, czas (teraźniejszy i przyszły traktowane łącznie przez SAM-a, LEM-a i prawdopodobnie Gram'a), osobę, rodzaj (dla czasu przeszłego) i liczbę dla czasowników. Istotną cechą LEM-a jest niezależne oznakowanie każdego z tych parametrów, jednak nie

ma to charakteru pozycyjnego, gdyż kolejność ich umieszczenia jest zmienna. LEM opatruje formy bezosobowe (*honorowano, rozpoczęto*) oznaczeniem czasu przeszłego; inne analizatory tego nie czynią, uznając tę własność za domyślną. Ponadto imiesłowy i odsłowniki są wiązane z odpowiadającymi im czasownikami przez Gram'a, SAM-a, PoMor'a i XeLD-ę, zaś przez LEM-a — nie. AMOR wiąże z czasownikiem imiesłowy przysłowne, zaś odsłowników w ogóle nie rozważa. Cenną własnością LEM-a i XeLD-y jest podawnie dla przyimków informacji o przypadku, w jakim muszą wystąpić wymagane przez nie rzeczowniki.

Zaimki traktowane są przez omawiane analizatory w sposób niejednolity. LEM wyróżnia zaimki rzeczowne, przymiotne i przysłowne, AMOR — rzeczowne i częściowo przymiotne, XeLDA — osobowe, względne jako główne kategorie oraz dzierżawcze, wskazujące i pytajne jako podkategorie trzeciej, najogólniejszej grupy. Pozostałe analizatory rozpatrują jedynie zaimki rzeczowne. Zauważmy, że SAM, Gram, PoMor i XeLDA przypisują zaimkom *sobie, się, siebie* formę podstawową *się*, gdy tymczasem LEM i AMOR — *siebie*. AMOR wyróżnia też partykułę *się*.

W testowanej przez nas wersji Gram'a nie są ujawniane bardziej szczegółowe informacje dotyczącej analizowanej formy, poza jej formą podstawową (i kanoniczną) oraz częścią mowy, do której rozważany wyraz należy, więc nie mogłyśmy sprawdzić poprawności rzeczywistego, pełnego etykietowania.

SAM, PoMor i Gram traktują zaimki *ona, ono, oni, one* jako formy zaimka *on*, LEM, XeLDA i AMOR rozważają je oddzielnie. Podobnie Gram i SAM uznają wyraz *ludzie* za formę rzeczownika *człowiek*, Gram wiąże także formy *lat, latach* z wyrazem *rok*. Natomiast AMOR uważa wyrazy *człowiek, ludzie* za oddzielne leksemy, LEM i XeLDA traktują w ten sposób także wyrazy *rok, lata*. Zauważmy, że SAM i AMOR wiążą formy *lata, latach* jedynie z formą podstawową *lato*, w ogóle ignorując liczbę mnogą słowa *rok*.

Ponadto SAM nadaje wyrazom *plurale tantum* rodzaj `blp`, LEM — `G*`, XeLDA — `M23FN`, zaś AMOR oznacza typ odmiany `plur`, zaś rodzaj zależy od konkretnego wyrazu (np. *drzwi* są rodzaju nijakiego, a *ludzie* — męskoosobowego).

Ustalenie, czy niektóre wyrazy to spójniki czy partykuły, albo też partykuły czy przysłówki (zaimki przysłowne), może być w wielu przypadkach problematyczne. Wśród przebadanych analizatorów, jedynie LEM, XeLDA i AMOR rozróżniają nieodmienne części mowy (czyni to także Gram, ale nie w testowanej przez nas wersji). W poniższej tabeli przedstawiony został sposób etykietowania wyrazów nieodmiennych (bez zaimków i wykrzykników) przez LEM-a, AMOR-a i XeLD-ę.

wyraz	spójnik	przysłówek	zaim.przysł.	partykuła
a	A,L,X	–	–	L
albo	A,L,X	–	–	L
ale	A,L,X	–	–	L
czy	A, L	–	–	A,L,X
czyli	A, X	–	–	L
czyż	–	–	–	A, X
chociaż	A,L,X	–	–	L
dopiero	–	X	–	L, A
dotąd	L	L, X	L	A
dziś	–	L, X	–	A
gdzie	–	–	L, X	A
i	A,L,X	–	–	A, X
iż	A,L,X	–	–	–
jak	A,L,X	–	L, X	–
jako	X	–	L	A, X
jednak	A,L,X	–	–	L
jeszcze	A,L,X	–	–	–
jeśli	A,L,X	–	–	–
już	–	–	–	A,L,X
kiedy	A,L,X	–	L, X	A
lecz	A,L,X	–	–	–
może	–	–	–	A,L,X
nie	–	–	–	A,L,X
niemal	–	L	–	A, X
niezbyt	–	L	–	A, X
nim	A,L,X	–	–	–
niż	A,L,X	–	–	–
oburącz	–	L, X	–	A

wyraz	spójnik	przysłówek	zaim.przysł.	partykuła
odtąd	–	L, X	–	A
omal	–	L, X	–	A
ponieważ	A,L,X	–	–	–
prawie	–	L	–	A, X
przeciwko	–	L, X	–	–
również	L, X	–	–	A
skoro	A,L,X	–	–	–
tak	–	–	L	A,L,X
także	L	–	–	A, X
też	L	–	–	A,L,X
tu	–	L	X	L, X
tuż	–	L	–	A,L,X
tylko	–	–	–	A,L,X
wciąż	–	L, X	–	A
więc	L, X	–	–	L, A
właśnie	–	A, X	–	L
wprawdzie	L, A	–	–	L, X
wprost	–	L, X	–	A
wraz	L, X	–	–	A
wręcz	–	L	–	L
wszędzie	–	L, X	–	A
za	–	L	–	X
zanim	A,L,X	–	–	–
zazwyczaj	–	L	–	A, X
zresztą	–	L	–	A,L,X
zwłaszcza	–	–	–	A,L,X
że	A,L,X	–	–	L

Ponadto, spośród powyższych wyrazów, PoMor uznaje za przysłówki *dziś* oraz *wprost*. Natomiast jako przyimki wszyscy traktują *przeciwko* oraz *za*, zaś AMOR i PoMor — także *niż*. Wątpliwości stwarza także wyraz *mimo*: Gram, LEM, PoMor i XeLDA traktują go jako przyimek, PoMor także jako przysłówek, zaś AMOR — jako spójnik. XeLDA i Gram traktują także jako przyimek wyraz *niczym* (*piękna niczym róża*); zaś SAM — jako wyraz nieodmienny.

Jak już wspominaliśmy we wstępie, w testach pojawiła się (celowo bądź przypadkowo) spora liczba form kontrowersyjnych, archaicznych bądź dopiero wchodzących do języka. I tak, forma *książęciu* zaaprobowana została jedynie przez SAM-a (wszystkie analizatory akceptują nowszą formę *księciu*), wyraz *mełty* — jedynie przez PoMor'a (wszyscy uznają *mielony*). Forma *stacyj* została uwzględniona w zasadzie przez wszystkich, ale już analogiczna forma *awaryj* została odrzucona przez AMOR-a i XeLD-ę. Rozkaźnik *ból* czasownika *boleć*, nie uznawany przez specjalistów za poprawny, został zaakceptowany przez SAM-a i Gram'a. Czasownik *mielić*, akceptowany w nowszych słownikach i traktowany jako niepoprawny w starszych, nie jest rozpoznawany przez żaden analizator, przy czym XeLDA uznaje słowo *mieli* jako formę czasownika *mleć* (zamiast powszechnie wiązanej z tym czasownikiem formy *miele*). Kolejną budzącą wątpliwości formą jest *domie*. PoMor uwzględnia ją jako rzadki wołacz wyrazu *dom*; nie wykluczają tego także słowniki. Jednak LEM, SAM, XeLDA przypisują jej także miejscownik, co już poprawne nie jest (Gram także akceptuje tę formę, lecz z braku opisu nie wiadomo, czy tylko wołacz, czy także miejscownik). Wszystkie analizatory nie tylko przypisują wyrazowi *niektóre* formę podstawową *niektóry*, ale także (ze względu na standardową regularność odmiany przymiotnikowej) uwzględniają w jej opisie rodzaj nijaki l.pojedynczej, chociaż zarówno Markowski (1999), jak i Doroszewski i Kurkowska (1973) wyraźnie stwierdzają, że wyrazy *niektórzy, niektóre* występują wyłącznie w liczbie mnogiej. Na koniec, jedynie AMOR nie traktuje słów *sypialni, ryżowi* jako form męskoosobowych przymiotników *sypialny, ryżowy*. Tym bardziej dziwi, że nie uwzględniono podobnego wyjątku dla wyrazu *niektóre*.

Rodzaj męskozwierzęcy, uwzględniany przez PoMor'a i XeLD-ę dla wyrazów *dureń, tysiąc* oraz samego PoMor'a dla *żołądź* (XeLDA w ogóle tego wyrazu nie rozpatruje) dotyczy gry w karty i w tym kontekście jest poprawny. W szczególności, dla tych wyrazów biernik jest równy dopełniaczowi (*grać w durnia, w tysiąca*). Dotyczy to także w pewnym stopniu nazw firm (*mam Fiata* czy *Panasonica*). Jedyny taki przypadek w testach to rozpoznanie formy *Panasonica* przez XeLDę, przy czym biernik przypisany jest tej formie pomimo potraktowania jej jako wyrazu rodzaju męskorzeczowego. Na tej samej zasadzie PoMor przypisuje rodzaj męskozwierzęcy wyrazowi *dolar* (*mam dolara*, ale *mam dolary*), XeLDA zaś, co dziwniejsze, zarówno rodzaj męskozwierzęcy, jak i męskorzeczowy.

PoMor i XeLDA traktują formę *premier* także jako jako nieodmienny wyraz rodzaju żeńskiego. Odpowiada to zwrotowi *panią premier*, jednak w podobnym kontekście może się pojawić wiele rzeczowników rodzaju męskoosobowego (*pani magister, pani mecenas, pani minister*), i nie jest dla nas oczywiste, czy istnieje wystarczające uzasadnienie dla traktowania tych przypadków jako odrębnych leksemów.

SAM i AMOR nie rozpoznają liczb, pozostałe testowane analizatory — tak (przy czym PoMor rozpoznaje także liczby rzymskie, zaś Gram wręcz podaje ich denotację za pomocą liczb arabskich). LEM rozpoznaje wyłącznie liczby naturalne, na które rozbija również daty itp. Gram i XeLDA rozpoznają daty i godziny, PoMor przetwarza daty, godziny i liczby ułamkowe całościowo, jednak interpretuje je jako ciąg cyfr porozdzielanych znakami specjalnymi, więc trudno uznać, że je rozpoznaje.

PoMor i XeLDA mają bogate słowniki skrótów, w tym skrótów nazw własnych, jednak XeLDA ma problemy z opisem skrótów wielu wyrażeń pospolitych. AMOR rozpoznaje sporo skrótów, lecz sporo także ignoruje, Gram rozpoznaje jedynie popularne skróty wyrażeń złożonych z wyrazów pospolitych, zaś SAM i LEM ignorują je wszystkie. Gram i XeLDA posiadają bogate słowniki nazw własnych, także PoMor i AMOR rozpoznają wiele z nich. SAM posiada specjalne oznakowanie dla nazw własnych, jednak wśród testowanych wyrazów rozpoznał jedynie imię *Maria*.

LEM, Gram i XeLDA identyfikują znaki przestankowe, PoMor wyróżnia kropkę i myślnik (łącznik), zaś SAM i AMOR ignorują je wszystkie.

Porównanie zawartości słowników rozważanych analizatorów na podstawie przeprowadzonych testów jest zadaniem trudnym. Słownik analizatora PoMor jest niezwykle bogaty, zwłaszcza języka oznaczanego kodem 2069, zawierającego wiele słów archaicznych i regionalnych. Dotyczy to również w dużym stopniu SAM-a. Takie bogactwo ma jednak swoje wady, gdyż zwiększa liczbę analiz nieadekwatnych w danym kontekście. Z drugiej strony, w SAM-ie daje się odczuć brak nowszych wyrazów, takich jak *biotechnologia*. XeLDA zawiera wiele słów nieistniejących, utworzonych sztucznie od prawidłowych form.

We wszystkich testowanych analizatorach dało się wykryć pewne błędy. Na przykład wszystkie analizatory poza PoMor'em i AMOR-em akceptują nieporawne formy *koniami*, natomiast poza wymienionym jedynie SAM i XeLDA przyjmują poprawną formę *końmi*. Wiele błędów wyraźnie wynika z omyłek przy tworzeniu słownika, na przykład pojawiające się częściej bądź rzadziej we wszystkich badanych analizatorach przypisanie czasownikom błędnego aspektu, czy też potraktowanie przez LEM-a przyimka *pod* jako formy czasownika *podać*. Także brak analiz dla form *dęba, książęciu* czy też *wstąg* wynika najprawdopodobniej z konstrukcji słownika. Niestety, pewne

błędy ewidentnie mają swe źródło w programie. Nadmiarowość akceptowanych form wykazywana przez SAM-a (dotyczy to także XeLD-y, i w pewnym stopniu Gram'a) jest zgodna z przyjętym przez autorów założeniem, nie może więc być traktowana jako błąd. Ale już brak formy przymiotnikowej *susi* trudno wytłumaczyć w inny sposób niż błędem w programie. Podobnie brak biernika w opisie form *twórców, mieszkańców* w analizatorze AMOR.

Wyniki analizy statystycznej dla plików tekstowych są dość zbliżone. Podstawowym źródłem niewielkich różnic jest większe bądź mniejsze bogactwo słownika nazw własnych i skrótów, oraz ignorowanie liczb (SAM, AMOR). Z drugiej strony LEM przypisuje niejednokrotnie kilka różnych części mowi temu samemu wyrazowi nieodmiennemu, co ma wpływ na wartość współczynników. Owe różnice w podejściu niwelowane są poprzez współczynniki liczone wyłącznie dla słów odmiennych.

Decydując się na ocenę zaprezentowanych analizatorów na podstawie przeprowadzonego w niniejszym raporcie porównania, należy cały czas pamiętać, że większość z nich jest wciąż dopracowywana i modyfikowana, w szczególności ich słowniki ulegają staej rozbudowie; jednak przy usuwaniu starych błędów niestety mogą pojawiać się nowe.

Bibliografia

Bień, Janusz S. (2001) ,,O pojęciu wyrazu morfologicznego". Włodzimierz Gruszczyński, red., *Nie bez znaczenia... prace ofirowane profesrowi zygmuntowi saloniemu z okazji 15 000 dni pracy naukowej.* Białystok: Wydawnictwo Uniwersytetu w Białymstoku, 67–77.

Doroszewski, Witold, red. (1997) *Słownik języka polskiego.* Warszawa: Państwowe Wydawnictwo Naukowe.

Doroszewski, Witold i Halina Kurkowska. (1973) *Słownik poprawnej polszczyzny.* Warszawa: Państwowe Wydawnictwo Naukowe.

Kurcz, I., A. Lewicki, J. Sambor, K. Szafran i J. Woronczak. (1990) *Słownik frekwencyjny polszczyzny współczesnej.* Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Markowski, Andrzej, red. (1999) *Słownik poprawnej polszczyzny.* Warszawa: Państwowe Wydawnictwo Naukowe.

McShane, Marjorie J. (2001, January) ,,Polish inflection fit for man and machine". Raport MCCS-01-325, New Mexico State University.

Podracki, Jerzy, red. (2001) *Wielki słownik ortograficzno-fleksyjny.* Warszawa: Wydawnictwo Horyzont.

Saloni, Zygmunt i Marek Świdziński. (1998) *Składnia współczesnego języka polskiego.* Warszawa: Państwowe Wydawnictwo Naukowe, czwarte (zmienione) wyd.

Szafran, Krzysztof. (1996) ,,Analizator morfologiczny sam-95. opis użytkowy". Raport TR 96-05 (226), Instytut Informatyki UW.

Szymczak, Mieczysław. (1975) *Słownik ortograficzny języka polskiego.* Warszawa: Państwowe Wydawnictwo Naukowe, pierwsze wyd.

Tokarski, Jan. (1993) *Schematyczny indeks* a tergo *polskich form wyrazowych (opracowanie i redakcja Zygmunt Saloni).* Warszawa: Państwowe Wydawnictwo Naukowe.

Vetulani, Zygmunt, Jacek Martinek, Tomasz Obrębski i Grażyna Vetulani. (1998) *Dictionary based methods and tools for language engineering.* Poznań: Wydawnictwo Naukowe UAM.

A Porównanie wyników działania analizatorów dla testowych zbiorów słów

A.1 Formy poprawne

W poniższych tabelach przyjęto następujące oznaczenia: – to brak analizy, + to poprawna analiza, <x to brak formy fleksyjnej x, >x to nadmiarowe rozpatrzenie formy x. Oznaczenia x są zgodne z przyjętymi w danym analizatorze. ! wskazuje na błąd, ? na wątpliwości, ** na brak testu dla danego słowa w danym analizatorze. Dla analizatora `Gram` zaznaczone zostało jedynie, czy dana forma podstawowa i/lub kanoniczna została rozważona.

RZECZOWNIKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
trądu	trąd	+	+	+	+	+	>M2	+
swędu	swąd	+	+	+	+	+	–	+
grobu	grób	+	+	+	+	+	+	+
wrogu	wróg	>GD	+	+	+	+	>A,G	+
wrogowi	wróg	+	+	+	+	+	+	+
wrogowie	wróg	$+^{c}$	+	+	+	+	+	+
mózgu	mózg	+	+	+	+	+	>A	+
szerszenia	szerszeń	$+^{a}$	+	+	+	+	–	+
ucznia	uczeń	$+^{a}$	+	+	+	+	+	+
uczniowie	uczeń	+	+	+	+	+	+	+
uczniów	uczeń	$+^{a}$	+	+	+	+	+	+
pereł	perła	+	+	+	+	+	–	+
pociągu	pociąg	+	+	+	+	+	>A	+
żołędzi	żołądź	$>N^{b}$	+	+	+	+	–	>b
	żołędź	!!	–	–	–	–	–	–
żołądź	żołądź	+	+	+	+	+	–	+
rąk	ręka	+	+	+	+	+	+	+
rękoma	ręka	+	+	+	+	+	+	+
rękami	ręka	+	+	+	+	+	+	+
oczyma	oko	+	+	+	+	+	+	+
oczami	oko	+	+	+	+	+	+	+
psem	pies	+	+	+	+	+	+	+
kwiatku	kwiatek	>G	+	+	+	+	>A,G	+
kwiatka	kwiatek	+	+	+	+	+	>A	+
domku	domek	+	+	+	+	+	>A	+
garbów	garb	+	+	+	+	+	+	+
końmi	koń	+	–	–	+	+	+	+
okoniami	okoń	+	–	+	+	+	–	+
dłoniami	dłoń	+	+	+	+	+	–	–
dłońmi	dłoń	+	+	+	+	+	+	+
kutra	kuter	+	+	+	+	+	+	+
kutrami	kuter	+	+	+	+	+	+	+
skutera	skuter	+	+	+	+	+	+	+
skuterami	skuter	+	+	+	+	+	+	+
głąba	głąb	$+^{a}$	–	–	>41m3	>41m3	>M2	+!
głąby	głąb	$+^{a,c}$	<Gp	+	+	+	>M2	+!
biesowi	bies	+	+	+	+	+	+	+
woźnice	woźnica	$+^{c,e}$	+	+	+	+	–	–
woźnicy	woźnica	<L	+	+	+	+	>N	>mn,w
	woźnik^{d}	+	–	–	–	+	–	–
woźnych	woźny	$+^{f}$	+	+	+	+	+	$+^{f}$
	woźna	–	–	+	+	+	–	–

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
ciem	ćma	+	+	+	+	+	−	+
drzazdze	drzazga	+	+	+	+	+	−	+
gaździe	gazda	$+^{e}$	+	+	+	+	+	−
gwieździe	gwiazda	$+^{b}$	+	+	+	+	+	+
ksiąg	księga	+	+	+	+	+	+	+
wstęg	wstęga	+	+	+	+	+	+	+
wstąg	wstęga	−	−	−	+	+	−	−
stacji	stacja	$+^{b}$	+	+	+	+	+	+
stacyj^{d}	stacja	+	+	+	−	+	+	+
sypialni	sypialnia	$+^{b}$	+	+	+	+	<G,Pl	<d,lm
	sypialny	+?	+?	+?	+?	+?	+?	−
sypialń	sypialnia	+	+	+	+	+	+	+
agonii	agonia	$+^{b}$	<pl	+	+	+	+	+
autonomii	autonomia	$>\text{pl}^{b}$	>pl	+	+	+	+	+
awarii	awaria	$+^{b}$	+	+	+	+	+	+
awaryj^{d}	awaria	+	+	+	−	+	−	−
kniei	knieja	$+^{b}$	+	+	+	+	+	+
kniej	knieja	+	+	+	+	+	+	−
alej	aleja	+	+	+	+	+	+	+
alei	aleja	$+^{b}$	+	+	+	+	+	<d,lm
idei	idea	$+^{b}$	+	+	+	+	+	+
idee	idea	$+^{a,c}$	+	+	+	+	+	+
buź	buzia	+	+	+	+	+	−	+
buzi	buzia	$+^{b}$	+	+	+	+	<G,Pl	+
mężczyźnie	mężczyzna	$<\text{L}^{e}$	+	+	+	+	+	<msc
samolocie	samolot	+	+	+	+	+	+	+
artykule	artykuł	+	+	+	+	+	+	+
przemyśle	przemysł	+	+	+	+	+	+	+
mrozie	mróz	+	+	+	+	+	+	+
chlebie	chleb	+	+	+	+	+	+	+
pokoi	pokój	+	+	+	+	+	+	+
liścia	liść	+	+	+	+	+	+	+
	liście^{d} (n)	+	−	−	+	+	−	−
nożu	nóż	+	+	+	+	+	+	+
nóż	nóż	$+^{a}$	+	+	+	+	+	+
pawiu	paw	+	+	−	+	+	>A	+
insekcie	insekt	+	+	+	+	+	+	+
agencie	agent	+	+	+	+	+	+	+
Turcy	turek	+	−	+	+	+	+	+
księciu	książę	+	+	+	+	+	+	+
książęciu	książę	+	−	−	−	−	−	−
źrebięciu	źrebię	+	+	+	+	+	+	+
źrebcu	źrebiec^{d}	<D	<D	+	+	+	>A	−
źrebca	źrebiec^{d}	$+^{a}$	+	+	+	+	+	−
źrebcowi	źrebiec^{d}?	+	+	+	+	+	+	−
księdzu	ksiądz	+	+	+	+	+	+	+
kamieniu	kamień	>G	+	+	+	+	+	+
łazęgów	łazęga	−	+	−	+	+	−	−
łazęg	łazęga	+	+	+	+	+	+	+
imienia	imię	+	+	+	+	+	+	+
	imienie^{d}	+	−	−	−	+	−	−
wylęgu	wyląg	+	+	+	+	+	−	−
	wylęg	+	+	+	+	+	>A	+
zębie	ząb	+	+	+	+	+	+	+
zęba	ząb	+	+	+	+	+	+	+
zębu	ząb	+	−	−	+	+	+	−

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
dębu	dąb	+	−	+	+	+	+	+
dęba	dąb	+	−	−	+	+	+!	−
durniów	dureń	+[a]	+	+	+	+	+	+
durniu	dureń	+	+	−	+	+	+	+
więźniu	więzień	+	+	+	+	+	+	+
minucie	minuta	<L	+	+	+	+	+	+
kapuście	kapusta	<L	+	+	+	+	+	+
ojczyźnie	ojczyzna	<L	+	+	+	+	+	>w
formule	formuła	<L	+	+	+	+	+	+
cesze	cecha	<L	+	+	+	+	+	+
stóp	stopa	+	+	+	+	+	+	+
pluskiew	pluskwa	+	+	+	+	+	+	+
cegieł	cegła	+	+	+	+	+	+	+
próśb	prośba	+	+	+	+	+	+	+
cnocie	cnota	<L	+	+	+	+	+	+
cnót	cnota	+	+	+	+	+	+	+
szkół	szkoła	+	+	+	+	+	+	+
dziele	dzieło	+	+	+	+	+	+	+
przęśle	przęsł	+	+	+	+	+	+	+
	przęśl[d]	+	−	−	−	+	−	−
piśmie	pismo	+	+	+	+	+	+	+
pasm	pasmo	+	+	+	+	+	+	+
pasmie	pasmo	−	+	+	+	+	−	−
paśmie	pasmo	+	+	+	+	+	+	+
przysłów	przysłowie	+	−	+	+	+	−	−
zwierząt	zwierzę	+	+	+	+	+	+	+
muzeum	muzeum	+[c]	+	+	+	+	+	+
zimorodek	zimorodek	+	+	+	+	+	>A	+
moczymorda	moczymorda	−	+	+	+	+	+!	−
wielkanoc	wielkanoc	+	+	−	−	−	−	−
den	dno	+	+	+	+	+	+	+
	dna	+	−	−	−	−	+	+
pasem	pasmo	+	+	+	+	+	−	−
	pas	+	+	+	+	+	+	+
	paso	−	−	−	−	−	!!	−
mąk	mąka	+	+	+	+	+	+	+
	męka	+	−	+	+	+	+	+
gmin	gmin	+	+	+	+	+	−	+
	gmina	+	+	+	+	+	+	+
grot	grot	+[a]	+	+	+	+	+	−
	grota	+	+	+	+	+	+	+
moście	most	+	+	+	+	+	+	+
	mość	+[a,c]	+	+	−	−	−	+
płocie	płot	+	+	+	+	+	+	+
	płoć	+[c]	+	+	+	+	+	<b,lm
głąb	głąb(N,Gi)	+[a]	+	+?	+	+	+	+?
	głąb(N,Gp)	+?	−	+?	+	+	−	+?
	głąb(N,Ga)	+?	−	+?	+?	+?	+!	+!
	głąb(N,Gf)	+	+	+?	+	+	−	−
	głąb(Adv)	−	−	−	−	−	+	−
	głąb(fraz)	−	−	−	+	+	−	−

a – przewidziany brak biernika
b – przewidziany brak miejscownika
c – przewidziany brak wołacza
d – arachaizm
e – wyraz rodzaju męskiego o odmianie żeńskiej, potraktowany jako rodz. żeński
f – rzeczownik o odmianie przymiotnikowej, potraktowany jako przymiotnik.

CZASOWNIKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
idę	iść	+	+	+	+	+	+	+
szłam	iść	+	+	+	+	+	+	+
szedłem	iść	+	+	+	+	+	+	+
poszedł	pójść	+	−	+	+	+	+	+
pójdźmy	pójść	+	−	+	+	+	−	+
ciął	ciąć	+	+	+	+	+	+	+
tnie	ciąć	+	+	+	>dk	>dk	+	+
tnę	ciąć	−	+	+	>dk	>dk	+	+
wezmę	wziąć	+	+	+	+	+	+	+
ciągnę	ciągnąć	+	+	+	+	+	+	+
wyląg	wylęgnąć	+	−	−	+	+	−	−
	wyląc	+	−	−	+	+	−	−
	wylec	−	−	−	−	!!	−	−
wylęgła	wylęgnąć	+	−	+	+	+	−	+
	(im.przym.przesz)	+	−	+	+	+	−	−
	wyląc	−	−	−	+	+	−	−
	(im.przym.przesz)	−	−	−	+	+	−	−
	wylec	−	−	−	−	!!	−	−
pragnę	pragnąć	+	+	+	+	+	+	+!
gaśnie	gasnąć	+	+	+	+	+	+	+!
rósł	rosnąć	+	+!(Ap)	+	+	+	+	+
	róść	−	−	+	+	+	−	−
pełźli	pełznąć	>ł	+	−	+	+	−	−
pełzli	pełznąć	>ł	+	−	+	+	>PRT	+
pełznie	pełznąć	+	+	−	+	+	+!	+
pełźnie	pełznąć	+	+	+	+	+	−	−
pełza	pełzać	+	+	+	+	+	+!	+
sroży	srożyć	+	+	+	+	+	−	+
brakuje	brakować	+	+	+	+	+	+	+
mierznie	mierznąć	+	+	+	+	+	+	−
woła	wołać	+	+	+	+	+	+	+
	wół	!!	−	−	−	−	!!	−
jadł	jeść	+	+	+	+	+	+	+
	jadło(N)	−	−	!!	−	−	−	−
plotła	pleść	+	+	+	+	+	−	+
plótł	pleść	+	+	+	+	+	−	+
wiódł	wieść	+	+	−	+	+	+	+
wiodła	wieść	+	+	−	+	+	+	+
wiodę	wieść	+	+	−	+	+	+	+
bódł	bóść	+	−	+	+	+	+	−
bodła	bóść	+	−	+	+	+	+	−
	bodnąć	>ł	−	−	−	−	−	−
wiezie	wieźć	+	+	+	+	+	+	+
mogę	móc	+	+	+	+	+	+	+
znajdź	znaleźć	+	−	+	+	+	+	+
	znajść	!!	−	−	−	−	−	−
mieli	mieć	+	+	+	+	+	+	+
	mielić	−	−	−	−	−	−	−
	mleć	−	−	−	−	−	+!	−
mielił	mielić	−	−	−	−	−	−	−
mielić	mielić	−	−	−	−	−	−	−
mleć	mleć	+	+	+	+	+	+	+
mełł	mleć	+	+	+	+	+	+	+
miele	mleć	+	+	+	+	+	−	+

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
ugrząźć	ugrząźć	+	+!!	+	+	+	−	−
ugrzęznąć	ugrzęznąć	+	+	+	+	+	−	+
ugrzązł	ugrzęznąć	+	−	+	+	+	−	−
	ugrząźć	−	+!!	−	+	+	−	−
ugrzęzła	ugrzęznąć	+	−	+	+	+	−	+
	(im.przym.przesz.)	+	−	+	+	+	−	−
	ugrząźć	−	+!!	−	+	+	−	−
	(im.przym.przesz.)	−	−	−	+	+	−	−
ugrzęźnij	ugrząźć	+	+!!	+	+	+	−	−
	ugrzęznąć	−	−	−	+	+	−	−
zaprzągł	zaprząc	+	+	+	+	+	−	+
	zaprzęgnąć	+	+	−	+	+	−	−
zaprzęgła	zaprząc	+	+	+	+	+	−	+
	(im.przym.przesz.)	!!	−	−	−	−	−	−
	zaprzęgnąć	+	+	−	+	+	−	+
	(im.przym.przesz.)	!!	−	−	!!	!!	−	−
odprzągł	odprząc	−	+	+	+	+	−	−
	odprzęgnąć	+	−	−	+	+	−	−
odprzężmy	odprząc	−	+	+	+	+	−	−
wlecz	wlec	+	+	+	+	+	+!	+
wlecze	wlec	+	+	+	+	+	+!	+
wlokę	wlec	+	+	−	+	+	+!	+
wlekę	wlec	+	+	**	+	+	−	+
wlokła	wlec	+	+	−	+	+	+!	+
wlókł	wlec	+	+	−	+	+	+!	+
wlekł	wlec	+	+	+	+	+	−	+
wlekła	wlec	+	+	+	+	+	−	+
wzięli	wziąć	+	+	+	+	+	+	+
weźmy	wziąć	+	+	+	+	+	+	+
weźmie	wziąć	+	+	+	+	+	+	+
są	być	+	<Ai	+	+	+	+	+
jest	być	+	<Ai	+	+	+	+	+
będą	być	+	<Ap	+	+	+	+	+
czytali	czytać	+	+	+	+	+	+	+
bogaciła	bogacić	+	+	+	+	+	+	+
bogacę	bogacić	+	+	+	+	+	+	+
myli	myć	+	+	+	+	+	+	+
	mylić	+	+	+	+	+	+	+
zaprzągłszy	zaprząc	+	?	−	+	+	−	+
	zaprzęgnąć	+	?	−	+	+	−	−
	zaprzągłszy	−/+	+	−	−/+	−/+	−	−/+
wypełzły	wypełznąć	+	+	−	+	+	−	+
	(im.przym.przesz.)	+	−	+	+	+	−	−
zdziczały	zdziczeć	+	+	+	+	+	+	+
	(im.przym.przesz.)	+	+	+	+	+	+	+
uczenia	uczenie	+	+	+	−	−	−	+
	uczyć	+	−/+	+	+	+	>lm	−
	uczeń	!!	−	−	−	−	−	−

PRZYMIOTNIKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
bielszy	biały	+	+	$+^{j}$	+	+	+	$+^{j}$
weselszy	wesoły	+	+	$+^{j}$	+	+	−	$+^{j}$
mądry	mądry(Adj)	+	+	+	+	+	+	+
	mądry(N)	−	+	−	−	−	−	−
mędrszy	mądry	+	−	−	+	+	−	−
mądrzejszy	mądry	+	+	$+^{j}$	+	+	+	$+^{j}$
wyższy	wysoki	+	+	$+^{j}$	+	+	+	$+^{j}$
lepszy	dobry	+	+	$+^{j}$	+	+	+	$+^{j}$
starszy	stary	+	+	$+^{j}$	+	+	+	$+^{j}$
sroższy	srogi	+	+	$+^{j}$	+	+	−	−
bogatszy	bogaty	+	+	$+^{j}$	+	+	+	$+^{j}$
milszy	miły	+	+	$+^{j}$	+	+	+	$+^{j}$
dalszy	daleki	+	+	$+^{j}$	+	+	+	$+^{j}$
niższi	niski	+	+	$+^{j}$	+	+	+	$+^{j}$
więksi	większy	−	+!	+	−	−	−	+
	duży	−	+	−	+	+	+	−
	wielki	+	−	−	+	+	−	−
łatwiejszy	łatwy	+	+	$+^{j}$	+	+	+	$+^{j}$
źli	zły (Adj)	+	+	+	+	+	+!	+
	zły (N)	−	+	−	−	−	−	−
susi	suchy	−	+	+	+	+	+	+
	sus	!!	−	−	−	−	−	−
suchszy	suchy	+	+	$+^{j}$	+	+	+	$+^{j}$
lekcy	lekki	+	+	+	+	+	+	+
wrodzy	wrogi	−	+	+	+	+	+	+
wrogiemu	wrogi	−	+	+	+	+	+	+
akademiccy	akademicki	+	+	+	+	+	+	+
mączny	mączny	+	+	+	+	+	+	+
dzikszy	dziki	+	+	−	+	+	−	$+^{j}$
dzikiej	dziki	+	+	+	+	+	+	+
psi	psi	+	+	+	+	+	−	+
mielony	mleć	+	+	+	+	+	+	−
	mielony(Adj)	−	−	+	−	−	−	+
mełty^{d}	mleć	−	−	−	+	+	−	−
długodzioby	długodzioby	+	+	+	+	+	+	−
dwunodzy	dwunogi	−	+	+	+	+	+	+
dwunogiemu	dwunogi	−	+	+	+	+	+	+
dwunogich	dwunogi	−	+	+	+	+	+	+
dwojakiego	dwojaki	+	+	+	+	+	−	+
dwojacy	dwojaki	+	−!	+	+	+	−	+
	dwojak	!!	−	−	!!	!!	!!	−
dwojakie	dwojaki	+	+	+	+	+	−	+
całodobowy	całodobowy	+	+	+	+	+	+	+
dzienny	dzienny	+	+	+	+	+	+	+
całodzienny	całodzienny	+	+	+	+	+	+	+
codzienny	codzienny	+	+	+	+	+	+	+
dwudniowy	dwudniowy	+	+	+	+	+	+	+
dwutygodniowy	dwutygodniowy	+	+	+	+	+	+	+

PRZYSŁÓWKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
najjaśniej	jaśnieć	!!	−	−	−	−	−	−
	jaśnie^{d}	+^{g}	−	−	−	−	−	−
	jasno	+^{g}	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
częściej	często	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+
lepiej	dobrze	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
	lepiej^{h,i}	?	−	−	−	−	−	−
lżej	lekko	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
	lżej^{h,i}	?	−	+^{j}	−	−	−	−
dalej	daleko	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+
	dalej	−	+(APP)	−	−	−	−	−
wyżej	wysoko	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
szerzej	szeroko	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
niżej	nisko	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
węziej	wąsko	+	+	+^{j}	+^{k}	+^{k}	−	+^{j}
ciszej	cicho	+	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
	ciszeć	−	!!	−	−	−	−	−
goręcej	gorąco	+	+	+^{j}	+^{k}	+^{k}	−	−
więcej	dużo	−	+	+^{j}	+^{k}	+^{k}	+^{k}	+^{j}
	więcej^{h,i}	?	−	−	−	−	−	−
	wiele	−	−	−	+	+	−	−
srodze	srodze	+	−	+	+	+	−	+
dużo	dużo	+	+	+	+^{k}	+^{k}	+^{k}	+
	dużo(N)	−	−	−	!!	!!	−	−
	dużo(Adj)	−	−	−	−	−	!!	!!
oburącz	oburącz(Adv)	−	+	+	−	−	−	−
	oburącz(Part)	+	−	−	+	+	+	+
	oburącz(N)	−	−	−	−	−	!!	−
wprost	wprost(Adv)	−	+	+	+	+	+	−
	wprost(Part)	+	−	−	+	+	+	+
dziennie	dziennie	+	+	+	+	+	+	+
codziennie	codziennie	+	−	−	+	+	+	+

g – brak oznaczenia stopnia najwyższego
h – brak oznaczenia stopnia wyższego
i – brak oznaczenia przysłówka
j – formę podstawową stanowi przymiotnik (przysłówek) w stopniu wyższym bądź najwyższym, a nie równym
k – formę podst. przysłówka stanowi przymiotnik w st. równym

NIEJEDNOZNACZNE

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
dziób	dziób	+	+	+	+	+	+	+
	dziobać	+	+	−	−	−	−	−
	dzióbać	+	−	+	+	+	+	−
pal	pal	+^{a}	+	−	+	+	+	+
	pala(N)	−	−	−	+	+	+	−
	palić	+	+	+	+	+	+	+
śledź	śledź	+	+	+	+	+	−	+
	śledzić	+	+	+	+	+	+	+
śledzi	śledź	+	+	+	+	+	−	+
	śledzić	+	+	+	+	+	+	+
broń	broń	+	+	+	+	+	+	+
	bronić	+	+	+	+	+	+	+

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
broni	broń	$+^{b}$	+	+	+	+	+	+
	bronić	+	+	+	+	+	+	+
trap	trap	+	+	+	+	+	+	+
	trapić	+	+	+	+	+	+	+
	trap(ONO)	+	−	−	+	+	−	−
trop	trop	+	+	+	+	+	+	+
	tropić	+	+	+	+	+	+	+
stop	stop(N)	$+^{a}$	+	+	+	+	+	+
	stop(Excl)	+	+	−	+	+	−	−
	stopić	+	−	+	+	+	+	+
	stopa	!!	−	−	−	−	−	−
płaszcz	płaszcz	$+^{a}$	+	−	+	+	+	+
	płaszczyć	+	+	+	+	+	+	+
piec	piec(N)	$+^{a}$	+	−	+	+	+	+
	piec(V)	+	+	+	+	+	+	+
piekło	piekło	$+^{a,c}$	+	+	+	+	+	+
	piec(V)	+	+	+	+	+	+	+
strzygą	strzyc	+	+	+	+	+	+	+
	strzyga	+	+	+	+	+	+	+
strzyże	strzyc	+	−	+	+	+	+	+
	strzyża	+	+	+	+	+	+	+
krój	krój	$+^{a}$	+	+	+	+	+	+
	kroić	+	+	+	+	+	−	+
klucz	klucz	$+^{a}$	+	−	+	+	<N	+
	kluczyć	+	+	+	+	+	+	+
kluczy	klucz	+	+	+	+	+	+	+
	kluczyć	+	+	+	+	+	+	+
dwój	dwója	+	+	+	+	+	+	+
	dwoić	+	+	+	+	+	−	+
trój	trója	+	+	+	+	+	+	+
	troić	+	+	+	+	+	−	+
szalej	szalej	$+^{a}$	+	−	+	+	+	+
	szaleć	+	+	+	+	+	+	+
pieprz	pieprz	$+^{a}$	+	−	+	+	+	+
	pieprzyć	+	+	+	+	+	+	+
grodź	grodź	+	−	−	+	+	+	−
	grodzić	+	+	+	+	+	−	+
gródź	gródź	+	+	+	+	+	−	+
	grodzić	+	−	−	+	+	+	−
grodzi	grodź	−	−	−	+	+	−	−
	gródź	+	+	+	+	+	−	>b,lm
	grodzić	+	+	+	+	+	+	+
	gród	!!	−	!!	−	−	−	−
rośnie	rosnąć	+	+!(Ap)	+	+	+	+	+
	róść	−	−	+	+	+	−	−
	rośnie^{d}(Adj)	+	−	−	+	+	$+^{k}$	−
	rośnia^{d}	+	−	−	−	+	+	−
pleć	pleć(V)	+	+	+	+	+	+	−
	pleść	+	+	+	+	+	−	+
	pleć^{d}(N,Gf)	+	−	−	+	+	−	−
nadejść	nadejście	+	+	+	+!	+!	+!	+
	nadejść	+	−	+	+	+	+	+
napaść	napaść(N)	+	+	+	+	+	+	+
	napaść(V)	+	+	+	+	+	+	+
	napaść(się)	+	?+	?+	?+	?+	?+	?+
wiedźmy	wiedźma	$+^{a,c}$	+	+	+	+	+	+
	wieść(V)	+	+	−	+	+	+	+

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
wieść	wieść(N)	+	+	+	+	+	+	−
	wieść(V)	+	+	−	+	+	+	+
	wieścić	+	+	+	+	+	+	−
brać	brać(N)	+	−	+	+	+	+	+
	brać(V)	+	+	+	+	+	+	+
	bracić^{d}	+	−	−	−	+	−	−
więź	więź	+	+	+	+	+	+	+
	więzić	+	+	+	+	+	+	+
pasie	pas	+	+	+	+	+	+	+
	paść	+	+	+	+	+	+	+
	paso	−	−	−	−	−	!!	−
poręczy	poręcz	+^{b}	+	+	+	+	+	+
	poręczyć	+	+	+	+	+	+	+
	poręcze^{d}(Gn)	+	−	−	−	+	−	−
miotła	miotła	+	+	+	+	+	+	+
	miotło^{d}(N)	+	−	−	+	+	−	−
	mieść	+	+	+	+	+	−	+
powodzi	powódź	+^{b}	+	+	+	+	−	>b,lm
	powód	!!	−	−	−	−	−	−
	powodzić	+	+	+	>dk	>dk	+!	−
działa	działo	+^{a}	+	+	+	+	+	<w,lm
	dział	!!	−	−	−	−	!!	−
	działać	+	+	+	+	+	+	+
	dziać	+	+	+	+	+	−	+
zboczy	zbocze	+	+	+	+	+	+	+
	zbocz^{d}(N)	+	−	−	−	+	−	−
	zboczyć	+	+	+	+	+	+	+
pośle	poseł	+	+	+	+	+	+	+
	posłać	+	+	+	+	+	+	+
pił	piła	+	+	+	+	+	+	+
	pić	+	+	+	+	+	+	+
tył	tył	+^{a}	+	+	+	+	+	+
	tyć	+	+	+	+	+	+	+
para	para	+	+	+	+	+	+	+
	par	+^{a}	+	+	+	+	+	−
	parać	+	+	+	+	+	+	+
pierze	pierze	+^{a}	+	+	+	+	+	<w
	pierz^{d}(N)	+	−	−	−	+	−	−
	prać	+	+	+	+	+	+	+
ból	ból	+^{a}	+	+	+	+	+	+
	boleć	+	−	+	−	−	−	−
boi	boja	+	+	+	+	+	+	<d,lm
	bój	!!	−	−	−	−	−	−
	boj	!!	−	−	−	−	−	−
	bać	+	+	+	+	+	+	+
bój	bój	+^{a}	+	+	+	+	+	+
	bać	+	+	+	+	+	−	−
tyka	tyka	+	+	+	+	+	+	+
	tyk(N)	!!	−	−	−	−	!!	−
	tykać	+	+	+	+	+	+	+
lata	rok	+^{a}	−?	+	+	+	−?	−
	lata(pl)	−?	+	−?	−?	−?	+	−
	lato	+^{a}	+	+	+	+	+	+
	latać	+	+	+	+	+	+	+
ćmi	ćmić	+	+	+	+	+	+	+
	ćmy	+	−	−	+	+	+	−

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
musi	musieć	+	+	+	+	+	+	+
	musić[d]	+	−	−	+	+	−	−
	muszy	+	+	+	+	+	−	−
	mus	!!	−	−	−	−	−	−
karpi	karp	+	+	+	+	+	−	+
	karpi(Adj)	−	−	+	+	+	+	+
karpie	karp	+[a,c]	+	+	+	+	−	+
	karpa	+	+	+	+	+	+	+
	karpi(Adj)	−	−	+	+	+	+	+
gzowi	giez	+	−	−	+	+	−	+
	gzowy	+	+	+	+	+	+	−
mule	muł	+	+	+	+	+	+	−
	muli(Adj)	−	+	+	+	+	+	−
ryżowi	ryż	+	+	+	+	+	+	+
	ryżowy	+	+	+	+	+	+	−
święci	święty	+	+	+	+	+	+	+
	święcić	+	−	+	+	+	+	+
czyści	czysty	+	+	+	+	+	+	+
	czyścić	+	+	+	+	+	+	+
	czysta(N)	−	−	−	−	−	!!	−
gorszy	gorszyć	+	+	+	+	+	+	+
	zły	+	+	+[j]	+	+	+	+[j]
żółci	żółć	+[b]	+	+	+	+	+	+
	żółty	+	+	+	+	+	+	+
	żółcić	+	−	+	+	+	+	−
biegli	biegły (Adj)	+	+	+	+	+	+	+
	biec	+	+	+	+	+	+	−
	biegnąć	+	+	+	+	+	>PRT	+
wroga	wróg	+[a]	+	+	+	+	+	+
	wrogi	−	+	+	+	+	+	+
lwie	lew	+	+	+	+	+	+	+
	lwi	+	+	+	+	+	+	+
dwunogi	dwunogi	−	+	+	+	+	+	+
	dwunóg	+	−	−	+	+	+	−
wolno	wolno (Adv)	+	+	+	+[k]	+[k]	+[k]	+
	wolno (V)	−	−	−	+	+	−	−
	wolna (N)	−	−	−	−	−	!!	−
dziczej	dziko	+	+	−	+[k]	+[k]	−	−
	dziczeć	+	+	+	+	+	+	+
	dziczy(Adj)	+	+	+	+	+	+	+
srożej	srogo	+	+	+[j]	+[k]	+[k]	−	+
	srożeć[d]	+	−	−	−	+	−	−
bogaciej	bogato	+	+	+[j]	+[k]	+[k]	+[k]	+
	bogacieć[d]	+	+	−	+	+	+	−
gorzej	źle	+	+	+[j]	+[k]	+[k]	+[k]	+[j]
	gorzej[h,i]	?	−	−	−	−	−	−
	gorzeć	+	+	+	+	+	+	+
durni	dureń	+	−	+	>m2	>m2	>M2	−
	durny	+	+	+	+	+	+	+
wrogi	wróg	+	−	−	+	+	+	−
	wrogi(Adj)	−	+	+	−	+	+	+
cudzoziemcze	cudzoziemiec	+	+	+	+	+	+	+
	cudzoziemczy[d]	+	−	−	−	+	−	−
księży	ksiądz	+	+	+	+	+	+	+
	księża[d](N)	+	−	−	−	+	−	−
	księży	+	+	+	+	+	+	+

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
prawie	prawo	+	+	+	+	+	+	+
	prawie(Adv)	−	+	+	−	−	$+^{k}$!	−
	prawie(Part)	+	−	−	+	+	+	+
lecz	leczyć	+	+	+	+	+	+	+
	lecz	+	+	+	+	+	+	+
	leczo	−	−	−	−	−	!!	−
bez	beza	+	+	+	+	+	+	+
	bez(N)	+	+	−	+	+	+	+
	bez(P)	+	+	+	+	+	+	+
albo	alba(N)	+	+	+	+	+	+	−
	albo(P)	+	+	+	+	+	+	+
może	móc	+	+	+	+	+	+	+
	może	+	+	+	+	+	+	+
wręcz	wręczyć	−	+!	+	+	+	+	+
	wręcz(Adv)	−	+	+	−	−	−	−
	wręcz(Part)	+	+	−	+	+	−	−
ma	mój	$+^{l}$	+	+	$+^{l}$	$+^{l}$	+	+
	mieć	+	+	+	+	+	+	+
czyż	czyż(N)	+	+	−	+	+	+	−
	czyż	+	−	−	+	+	+	+
ciach	ciacho	+	+	+	+	+	+	−
	ciach(Ono)	+	+	+	+	+	−	+
uczeni	uczeń	!!	−	−	−	−	!!	−
	uczony(Adj)	+	+	+	+	+	+	+
	uczony(N)	−	−	−	+	+	−	−
	uczyć	+	−	+	+	+	+	−
nie	nie	+	+	+	+	+	+	+
	on	+!	(*)	+	+	+	(*)	(*)
	one	(*)	+	(*)	(*)	(*)	<FN	−
	ono	(*)	+	(*)	(*)	(*)	−	+

l – zaimek przymiotny oznaczany jako przymiotnik

INNE

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
pięcioro	pięcioro	−	+	**	+	+	+	+
pięciorgu	pięcioro	−	+	**	+	+	+	+
coś	coś	+	+	+?(N)	$+^{m}$	$+^{m}$	+	+
	co+ś	−	−	−	+	+	−	−
niczym	nic	−	+	+	+	+	+	+
	niczym	+	−	+	−	−	+	−
polsku	polsku	+	+	−	+	+	+	+
prostu	prostu	+	−	−	+	+	+	+
właśnie	właśnie	+	+	+	+	+	$+^{k}$	+
	właśnie(N)	−	−	−	−	−	!!	−
między	między	+	+	+	+	+	+	+
omal	omal	+	+	+	+	+	+	+
ale	ale	+	+	+	+	+	+	+
zeń	z niego	−	+	+!	+?	+?	−	−
doń	do niego	−	+	+	+?	+?	−	−
czmych	czmych	+	+	+	+	+	−	−
żebyś	żeby(m)	−	−	+	+	+	+	+
biało-czerwona		−	+	+	+	+	−	+/−
polsko-francuski		−	−	+	+	+	−	+/−

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	XeLDA	AMOR
alem	ale	−	−	−	+	+	+	−
półtora	półtora	+	+	+	+	+	+!	+!
kilkadziesiąt	kilkadziesiąt	+	+	+	+	+	+	+
wiele	wiele(Num)	+	+	+	+	+	+?!	−
	wiele/u(Adj)	−	−	−	+	+	+!	+
	wiele(Part)	−	−	−	−	−	−	+
wielu	wiele/u	+	+	+	+	+	−	+
27.4.1987	–	−	−	??	+?	+?	+	−
17:02	–	−	−	+	+?	+?	+	−
0.75	–	−	−	−	+?	+?	+	−
$45.000	–	−	−	−	−	−	−	−
Jan	Jan	−	−	+	+	+	+	+
Kowalski	kowalski(Adj)	+	+	+	+	+	+	+
Grzegorza	Grzegorz	−	−	+	+	+	+	+
Grzymały	Grzymała	−	−	−	−	−	−	−
Białegostoku	Białystok	−	−	+	+	+	+!	−
Wielkopolska	wielkopolski	+	+	−	+	+	+	+
	Wielkopolska	−	−	+	+	+	−	+
Zielona	zielona	+	+	+	+	+	+	+
Góra	góra	+	+	+	+	+	+	+
Zielona Góra		−	−	−	−	−	−	−
Puławska	Puławska/ski	−	−	+?	+	+	+	+
Madalińskiego	Madaliński	−	−	+?	−	−	−	−
Stołeczna	stołeczny	+	+	+	+	+	+	+
XIV-wieczny	–	−	−	−	+	+	−	−
http://www.ipipan.waw		−		+	−	−	−	−
aniak@ipipan.waw.pl		−		+	−	−	−	−
XIV	14	−	−	+	+/−	+/−	−	−
mgr	magister	−	−	+	+	+	$+!^n$	−
dr	doktor	−	−	+	+	+	$+!^n$	−
hab.	habilitacja	−	−	+	$+^n$	$+^n$	$+!^n$	−
	habilitowany	−	−	−	$+^n$	$+^n$	$+!^n$	−
inż.	inżynier	−	−	+	$+^n$	$+^n$	$+!^n$	−
PCK	–	−	−	−	+	+	+	+
PTTK	–	−	−	−	+	+	+?	−
NATO	–	−	−	−	+	+	+	+
PKP	–	−	−	−	+	+	+	+
MHz	–	−	−	−	+	+	+!	−
USA	–	−	−	−	+	+	+	+
aha	aha	+	+	**	+	+	+	+
miau	miau	+	+	−	+	+	−	−
metnik	metnik	−/+	−	−	−	−	−/+	−
komputeruje	komputerować	−/+	−	−	−	−	−/−	−
kluczówka	kluczówka	−/+	−	−	−	−	−/+	−
wiertopięt	wiertopięt	−/+	−	−	−	−	−/+	−
pracowstręt	pracowstręt	−/+	−	−	−	−	−/+	−
piotruje	piotrować	−/+	−	−	−	−	−/−	−

m – brak określenia przypadka (dla zaimków rzeczownych?)
n – skrót nierozkodowany

Poniżej prezentujemy uproszczoną analizę statystyczną dla listy form poprawnych. W analizie tej nie brano pod uwagę wyrazów z zestawu INNE, pominięto też słowa archaiczne oraz alternatywne bezokoliczniki, by nie preferować analizatorów posiadających je w swoich słownikach. W rezultacie rozważane są 482 leksemy (w tym 453 poprawne), uzyskane dla 335 analizowanych form wyrazowych. Jeśli dane słowo nie zostało przetestowane w danym analizatorze (oznaczenie '**' w tabeli), uznawane jest za rozpoznane poprawnie. Jedynym liczonym współczynnikiem jest precision = correct/453.

	SAM	LEM	Gram	PoMor	XeLDA	AMOR
unknown	26	42	47	9	79	58
less	58	3	?	0	5	7
non-exist	17	1	2	5	12	1
more	25	2	?	7	28	6
correct	355	377	489	430	319	354
precision	0.784	0.832	0.859	0.949	0.704	0.782

A.2 Formy niepoprawne

W poniższych tabelach ograniczyliśmy się do następujących oznaczeń: − to oczekiwany brak analizy, + to niepoprawna analiza, ? w przypadku wątpliwości.

RZECZOWNIKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	Xelda	AMOR
trędu	trąd	+	−	−	−	−	−	−
trąda	trąd	+	−	−	−	−	+	−
swądu	swąd	+	−	−	−	−	−	−
swęda	swąd	−	−	−	−	−	−	−
oknia	okno	−	−	−	−	−	−	−
okni	okno	−	−	−	−	−	−	−
lasi	las	+	−	−	−	−	−	−
	laszy^d	−	−	−	−	+	−	−
gróbu	grób	+	−	−	−	−	−	−
groba	grób	+	−	−	−	−	+	−
mozgu	móżg	−	−	−	−	−	−	−
ucznie	uczeń	+	+	−	−	−	+	−
uczni	uczeń	+	+	+	−	−	−	−
szersznia	szerszeń	−	−	−	−	−	−	−
żołądzi	żołądź	−	−	+	−	−	−	−
żołędź	żołędź	+	−	−	−	−	−	−
żołądzie	żołądź	−	−	+	−	−	+	−
	żołąd	−	−	−	+	+	−	−
oczmi	oko	−	−	−	−	−	−	−
oczymom	oko	−	−	−	−	−	−	−
kwiateka	kwiatek	+	−	−	−	−	−	−
domeku	domek	+	−	−	−	−	−	−
domka	domek	+	−	−	−	−	+	−
koniami	koń	+	+	+	−	−	−	−
okońmi	okoń	+	−	−	−	−	−	−
słońmi	słoń	+	−	−	−	−	−	−
kutera	kuter	+	−	−	−	−	−	−
kutru	kuter	+	−	−	−	−	+	−
skutru	skuter	+	−	−	−	−	−	−
skuteru	skuter	+	−	−	−	−	+	−
skutra	skuter	+	−	−	−	−	−	−
gzu	giez	−	+	−	−	−	−	−
giezem	giez	+	−	−	−	−	+	−
biesu	bies	+	−	−	−	−	+	−
bsem	bies	+	−	−	−	−	−	−
piesem	pies	+	−	−	−	−	−	−
gwiaździe	gwiazda	+	−	−	−	−	−	−
drzezdzie	drzazga	−	−	−	−	−	−	−
drzeździe	drzazga	−	−	−	−	−	−	−

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	Xelda	AMOR
księg	księga	+	−	−	−	−	+	−
męk	męka	+	+	−	−	−	−	−
gródzi	gro(ó)dź	−	−	−	−	−	−	−
źrebieca	źrebiec	+	−	−	−	−	−	−
źrebiu	źreb	+?	−	−	+?	+?	−	−
bronii	broń	−	−	−	−	−	−	−
sypialnii	sypialnia	+	−	−	−	−	−	−
agoni	agon	+	−	−	−	−	−	−
	agonia	+	−	−	−	−	−	−
autonomi	autonomia	+	−	−	−	−	−	−
awari	awaria	+	−	−	−	−	−	−
idej	idea	+	−	−	−	−	+	−
ideje	idea	−	−	−	−	−	−	−
ćmów	ćma	−	−	−	−	−	−	−
cechie	cecha	−	−	−	−	−	−	−
senie	sen	−	−	−	−	−	−	−
ząbie	ząb	+	−	+	−	−	−	−
ząby	ząb	+	−	+	−	−	−	−
dureniu	dureń	+	−	−	−	−	−	−
kamniu	kamień	+	−	−	−	−	−	−
cegł	cegła	−	−	−	−	−	−	−
cegłów	cegłą	−	−	−	−	−	−	−
prośb	prośba	+	−	−	−	−	−	−
cnot	cnota	+	−	−	−	−	−	−
grót	grota	+	−	−	−	−	−	−
szkoł	szkoła	+	−	−	−	−	−	−
pasmów	pasmo	−	−	−	−	−	−	−
wylągu	wylęg	−	−	+	−	−	−	−
	wyląg	+	−	−	−	−	+	−

CZASOWNIKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	Xelda	AMOR
szłem	iść	−	−	−	−	−	−	−
szedłam	iść	−	−	−	−	−	−	−
poszłem	pójść	−	−	−	−	−	−	−
poszedłam	pójść	−	−	−	−	−	−	−
pójdzijmy	pójść	−	−	−	−	−	−	−
jedł	jeść	−	−	−	−	−	−	−
jadli	jeść	−	−	−	−	−	−	−
jedły	jeść	−	−	−	−	−	−	−
bra	brać	−	−	−	−	−	−	−
skakam	skakać	−	−	−	−	−	−	−
skakaj	skakać	−	−	−	−	−	−	−
skaka	skakać	−	−	−	−	−	−	−
ugrzęść	ugrzęść	−	−	−	−	−	−	−
ugrząznąć	ugrząznąć	−	−	−	−	−	−	−
ugrzęznąłem	ugrzęznąć	+	+	−	−	−	−	−
ugrzęznęła	ugrzęznąć	+	+	**	−	−	−	−
ugrzęzł	ugrzęznąć	+	−	−	−	−	−	+
ugrzązła	ugrzęznąć	−	−	+	−	−	−	−
zaprzągła	zaprząc	+	−	−	−	−	−	−
odprzęgł	odprzęgnąć	+?	+?	−	−	−	−	−
odprząż	odprzęgnąć	−	−	−	−	−	−	−
	odprząc	+	+	−	+	+	−	−

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	Xelda	AMOR
plotł	pleść	–	–	–	–	–	–	–
plótła	pleść	–	–	–	–	–	–	–
bodł	bóść	+	–	–	–	–	–	–
bódła	bóść	–	–	–	–	–	–	–
bódzie	bóść	–	–	–	–	–	–	–
wiedę	wieść	–	–	–	–	–	–	–
wiódła	wieść	–	–	–	–	–	–	–
wiodł	wieść	–	–	–	–	–	–	–
wiozie	wieźć	–	–	–	–	–	–	–
wiezła	wieźć	+	–	**	–	–	–	–
wlókła	wlec	–	–	–	–	–	–	–
wlokł	wlec	–	–	–	–	–	–	–
weznę	wziąć	–	–	–	–	–	–	–
wzięli	wziąć	–	–	–	–	–	–	–
weźmij	wziąć	–	–	–	+	+	+	–

PRZYMIOTNIKI I PRZYSŁÓWKI

formy analizowane	formy podstawowe	SAM	LEM	Gram	PoMor 1045	PoMor 2069	Xelda	AMOR
bialszy	biały	+	–	–	–	–	–	–
bielejszy	biały	–	–	–	–	–	–	–
wysoższy	wysoki	–	–	–	–	–	–	–
dobrzejszy	dobry	+	–	–	–	–	–	–
dobrszy	dobry	+	–	–	–	–	–	–
źlejszy	zły	–	–	–	–	–	–	–
duższy	duży	+	–	–	–	–	–	–
malejszy	mały	–	–	–	–	–	–	–
malszy	mały	+	–	–	–	–	–	–
miejszy	mały	–	–	–	–	–	–	–
bogaciejszy	bogaty	+	–	–	–	–	–	–
dalejszy	daleki	–	–	–	–	–	–	–
wesolszy	wesoły	+	–	–	–	–	–	–
szczodrszy	szczodry	+	–	–	–	–	–	–
mądrszy	mądry	+	–	–	–	–	–	–
starejszy	stary	–	–	–	–	–	–	–
lekszy	lekki	–	–	–	–	–	–	–
uczoni	uczony	+	–	–	–	–	+	–
mąkowy	mąkowy	–	–	–	–	–	–	–
dniowy	dniowy	+	+	–	+	+	+	+
codniowy	codniowy	–	–	–	–	–	–	–
całodziennie	całodziennie	–	–	–	–	–	+	–
codniowo	codniowo	–	–	–	–	–	–	–

Poniżej prezentujemy uproszczoną analizę statystyczną dla listy form niepoprawnych. Rozważane jest 127 leksemów dla 124 analizowanych form wyrazowych. Pokazujemy, ile takich form dany analizator potraktował jako poprawne (non-exist), ajedynym obliczanym współczynnikiem jest specyficzny error-rate = non-exist/127.

	SAM	LEM	Gram	POMor	XeLDA	AMOR
non-exist	61	10	9	3	17	2
error-rate	0.480	0.079	0.071	0.023	0.134	0.016

B Pliki tekstowe

B.1 KRAJ

Trwają egzaminy wstępne na wyższe uczelnie. A ponieważ można próbować swoich sił w kilku uczelniach, liczby kandydatów przypadających na jedno miejsce dodatkowo wzrastają. Reporterzy „Rzeczpospolitej" zebrali informacje o najbardziej obleganych kierunkach w uczelniach całej Polski. Najwięcej chętnych na jedno miejsce przypada na kierunkach uniwersyteckich, zwłaszcza psychologii. Ale równie obleganym kierunkiem jest informatyka — w Uniwersytecie Warszawskim na jedno miejsce przypada 18 kandydatów. Mniej o dwie osoby na miejsce na tym kierunku zanotował Uniwersytet Gdański. Z kolei w Poznaniu na Uniwersytecie Adama Mickiewicza na każdy indeks czeka 13 chętnych. Informatyka najbardziej oblężona jest także w Uniwersytecie Warmińsko-Mazurskim — o jedno miejsce ubiega się 11 kandydatów.Łatwiej będzie w Uniwersytecie w Białymstoku — chętnych do zgłębiania nowych technologii jest pięć razy więcej niż miejsc.

Wciąż dobre notowania wśród młodych ludzi ma filologia angielska. W Uniwersytecie Warszawskim o jeden indeks walczy 17 kandydatów. Większe szanse mają kandydaci w Uniwersytecie Śląskim (10 kandydatów na miejsce), w Katolickim Uniwersytecie Lubelskim i Uniwersytecie w Białymstoku (gdzie kandydatów jest „tylko" 6 na miejsce) oraz w Uniwersytecie Szczecińskim (5 kandydatów na miejsce).

Co uniwersytet to specjalność

Uczelnie mają też swoje własne „specjalności". I tak w Uniwersytecie Warszawskim popularna jest socjologia (15 kandydatów). Z kolei w Uniwersytecie Jagiellońskim dużym zainteresowaniem cieszy się — oprócz psychologii — biotechnologia, zarządzanie turystyką i zarządzanie administracją publiczną (po 10 osób na jedno miejsce).

Biotechnologia i japonistyka przyciągają maturzystów na poznański Uniwersytet Adama Mickiewicza (po 12 kandydatów na miejsce). Bardziej „konserwatywne" kierunki przewodzą w Uniwersytecie Mikołaja Kopernika w Toruniu — najwięcej kandydatów ubiega się o przyjęcie na kierunki administracja (25 na miejsce) i stosunki międzynarodowe (17 osób na miejsce).

Stosunki międzynarodowe i iberystyka to najbardziej oblegane studia w Uniwersytecie Wrocławskim (po 17 osób na miejsce). Kandydaci z Dolnego Śląska chętnie składali dokumenty również na dziennikarstwo i komunikację społeczną (14 kandydatów na miejsce).

W pobliskim Opolu, w tamtejszym Uniwersytecie Opolskim, najwięcej chętnych zgłosiło się na pedagogikę: resocjalizacyjną (17).W lubelskim Uniwersytecie Marii Curie-Skłodowskiej najwięcej kandydatów na jedno miejsce przypada na stosunkach międzynarodowych -15, biotechnologii -10, oraz pedagogice — 9. Na stosunki międzynarodowe łatwiej się dostać na olsztyński Uniwersytet Warmińsko-Mazurski — chętnych jest 8 osób na jedno miejsce.

W najmłodszym polskim uniwersytecie, Uniwersytecie Kardynała Stefana Wyszyńskiego w Warszawie, największym zainteresowaniem — oprócz psychologii — cieszy się politologia — 12 osób na miejsce. Na politologię łatwiej zdać na Uniwersytet Szczeciński — każdy kandydat musi wyeliminować 6 osób. Łatwiej na uczelnie techniczne Uczelnie techniczne mają zazwyczaj mniej kandydatów niż uniwersytety, chociaż na niektóre kierunki dostać się jest trudniej niż na niejeden uniwersytet. Niektóre politechniki uruchomiły nietechniczne kierunki studiów, na przykład socjologię czy administrację, i one też cieszą się dużym zainteresowaniem.

B.2 ŚWIAT

USA czekają na rozstrzygnięcie

Slobodan Miloszević, przebywający w areszcie od 1 kwietnia, może być wydany międzynarodowemu trybunałowi ds. zbrodni wojennych w byłej Jugosławii (TPI) najwcześniej w piątek — oznajmił Veselin Cerović, jeden z adwokatów byłego prezydenta. Również premier Serbii, Zoran Djindjić, podał piątek jako najwcześniejszą datę ekstradycji byłego prezydenta. Miloszević ma być przesłuchany w środę przez trybunał okręgowy, który ma trzy dni na wydanie decyzji. Od werdyktu tego obrońcy Miloszevicia będą mogli odwołać się do Sądu Najwyższego również w ciągu trzech dni. Jeśli Sąd Najwyższy odrzuci apelację, decyzja o przekazaniu Miloszevicia do Hagi będzie mogła być wykonana natychmiast.

Były prezydent wystąpił już o wyznaczenie zespołu prawników, którzy będą go bronić w Hadze. Oznacza to, iż zdaje sobie sprawę, że nie uniknie procesu przed ONZ-owskim trybunałem. Dotąd miał nadzieję, że jugosłowiański wymiar sprawiedliwości odrzuci pozew TPI. Powtarza, że wysunięte przeciwko niemu zarzuty są wynikiem machinacji politycznych NATO. W maju 1999 roku trybunał międzynarodowy oskarżył Miloszevicia o zbrodnie wojenne w Kosowie. W areszcie przebywa on od 1 kwietnia pod zarzutem malwersacji finansowych i nadużycia władzy.

Obrońcy Miloszevicia zaskarżyli już do Sądu Najwyższego dekret rządu, legalizujący współpracę Jugosławii z ONZ-owskim trybunałem w Hadze, jako sprzeczny z konstytucją. Premier Djindjić odrzucił ich zarzuty. — Chodzi tu o postępowanie karne stosowane dotąd wobec cudzoziemców, które w wyniku wydanego w sobotę

dekretu może odtąd objąć również obywateli Jugosławii - wyjaśnił. Zanim dekret wszedł w życie, jugosłowiańskie prawo nie przewidywało ekstradycji Jugosłowian do innego państwa.

Przywódcy Jugosławii mają nadzieję, że 29 czerwca w Luksemburgu państwa zamierzające dofinansować odbudowę jugosłowiańskiej gospodarki przyznają Belgradowi pomoc w wysokości 1,3 mld dolarów. Stany Zjednoczone nie potwierdziły jeszcze, czy wezmą udział w spotkaniu. Czekają na rozstrzygnięcie sprawy Miloszevicia.

B.3 KULTURA

Grand Prix dla Rosochy

Na Japońskim Biennale Ilustracji Europejskiej w Aki, Polacy zdobyli pięć nagród, w tym Grand Prix, które otrzymał Wiesław Rosocha. Na wystawie szczególnie uhonorowano też twórczość Józefa Wilkonia, znanego w wielu krajach z wyśmienitych ilustracji.

Polska, regularnie wyróżniana na Targach Książki Dziecięcej w Bolonii, w 2003 będzie ich honorowym gościem. Przygotowania już rozpoczęto: sekcja Ilustratorów zorganizuje w listopadzie kolejną wystawę z cyklu „Pro Bolonia", na której pokazane będą najlepsze projekty rysunkowe, malarskie i typograficzne. Publikowany z tej okazji katalog ma pomóc wydawcom w kontakcie z artystami. Powinno to przynieść konkretne efekty, czyli książki ilustrowane przez rodzimych twórców, a takimi właśnie powinniśmy się chwalić w Bolonii.

B.4 PUBLICYSTYKA

W sieni szarej kamienicy przy ulicy Krasińskiego na warszawskim Żoliborzu nie świeci się żarówka, lecz jaskrawe uchwyty stalowych drzwi w półmroku wskazują drogę. Słuchać zgrzyt ciężkich zasuw i „obronno-ochronne" hermetyczne odrzwia cicho ustępują. Jeszcze kilka schodków w dół i jesteśmy w bunkrze, czyli jednym z 2100 stołecznych schronów, które mogą pomieścić ćwierć miliona warszawiaków.

W całym kraju 15192 budowle ochronne wciąż drzemią pod fabrycznymi halami, urzędami i zwyczajnymi blokami. Tak zwane ukrycia, w których w razie wojennej pożogi mogłoby szukać schronienia 1,6 miliona ludzi, nie mają dziś prawdziwego gospodarza.

Przekonała się o tym posłanka Helena Góralska (UW). Odkryła schron na ponad 200 osób pod kamienicą, w której mieszka wraz z 80 sąsiadami. Wspólnota mieszkaniowa, którą utworzyli, nie chce utrzymywać schronu, który służy dziś jako piwnica. Wyróżnia się potężnymi stropami, są w nim awaryjne wyjścia, pomieszczenia stacji filtrów i zdewastowane instalacje sanitarne. Lokatorzy nie akceptują poleceń i ograniczeń sugerowanych przez władze obrony cywilnej. Podziemia to ich własność, więc dziwią się, że nie pozwala im się piwnicy użytkować, jak chcą, i na przykład... tynkować. Według instrukcji w czasie bombardowania czy ciężkiego ostrzału odpadający tynk mógłby podusić ukrytych w schronie mieszkańców.

Apele przestały skutkować

— Kiedy zaczęłam sprawdzać, kto za „cywilne" bunkry w Polsce odpowiada, okazało się, że nikt, a przepisy, które nakładały na właścicieli domów obowiązki związane z instalacjami Obrony Cywilnej, są powielaczowe, niepublikowane, a przede wszystkim już dawno się zdezaktualizowały — mówi Góralska.

— Od siedmiu lat czekamy na jasne regulacje — potwierdza płk Włodzimierz Stańczyk z Wydziału Zarządzania Kryzysowego, Ochrony Ludności i Spraw Obronnych Urzędu Wojewódzkiego w Lublinie. Czterystutysięczny Lublin ma 157 schronów (największy pod halami dawnej FSC), czterdziestotysięczny Świdnik — cztery.

— W sprawie zachowania bunkrów w kamienicach, które są własnością wspólnot mieszkaniowych czy spółdzielni, mogę tylko pisać i apelować w imieniu dobra wspólnego. A gospodarze i tak zrobią, co zechcą — mówi płk Stańczyk.

Urzędnicy Departamentu Bezpieczeństwa Powszechnego MSWiA sądzą, iż do czasu formalnego uregulowania tej kwestii warunki utrzymania schronów należałoby ustalać w drodze porozumienia między właścicielami a odpowiednim szefem komórki Obrony Cywilnej. — Nie da się już dziś zaakceptować „zalecenia o stosowaniu dotychczasowych przepisów", pochodzących z innej epoki, skoro podstawy prawne do tego są nikłe — mówi poseł Góralska.

Najbezpieczniej w salonie masażu

W schronie przy ul. Krasińskiego w Warszawie widać staranie o utrzymanie bojowej gotowości: łącznica telefoniczna ma wprawdzie czterdzieści lat, jest jednak wciąż czynna. Tuż obok zainstalowano dodatkowo elektroniczną centralkę Panasonica najnowszej generacji. Sprawna jest stacja wentylatorowa, świeżo wyremontowano sanitariaty przystosowane do przeprowadzenia dezaktywacji i odkażania.

— Nie wszędzie udaje się utrzymać odpowiedni poziom — żałuje płk Wiesław Baranowicz, wiceszef Miejskiego Inspektoratu Obrony Cywilnej w Warszawie. Baranowicz nowych użytkowników swych bunkrów szuka więc, gdzie się da. Do części schronów za symboliczną opłatą wprowadziły się już warsztaty rzemieślnicze, archiwa, magazyny, pralnie, siłownie. Swoją bazę przeniosło do warszawskiego bunkra stowarzyszenie Strzelec. W schronie

jest też salon masażu. — Warunkiem zainstalowania się w obiekcie ochronnym jest utrzymanie przez nowych użytkowników podstawowych funkcji schronu — mówi kategorycznie wiceszef MIOC.

W Bydgoszczy, jak mówi Joanna Jankowska z Urzędu Wojewódzkiego, wspólnoty mieszkaniowe postawiły Obronie Cywilnej ultimatum. Do końca roku ma się wyjaśnić, co dalej. Właścicieli irytują komory filtro- -wentylacyjne. Pytają, kto ma ponosić koszty ich utrzymania, skoro nie da się pomieszczeń użytkować jako piwnicy. W województwie kujawsko- -pomorskim przeprowadzono dopiero weryfikację schronów — jest ich dziś 201 i, podobnie jak w całej Polsce, są tylko w miastach.

W Poznaniu służby wojewody są już zmęczone brakiem formalnych rozstrzygnięć. W sprawie tzw. ukryć trwa jałowa korespondencja z prywatnymi właścicielami. Jeśli schron, to tylko prywatnie

Z najnowszego raportu sporządzonego na potrzeby szefa OC, którym jest obecnie komendant główny Państwowej Straży Pożarnej, wynika, że spośród 15 tysięcy budowli ochronnych 6927 to typowe schrony. Niemal połowa zlokalizowana została na terenie zakładów pracy. Większość „ukryć" wybudowano w piwnicach budynków w latach 1951 — 1960. Trzy czwarte schronów jest tak zaniedbanych, że wymaga natychmiastowych remontów i modernizacji.

W latach dziewięćdziesiątych, według raportu PSP, zaprzestano w Polsce wznoszenia za publiczne pieniądze budowli ochronnych, podobnie zresztą postąpiły np. Anglia, Niemcy, Czechy czy Węgry.

Dziś obowiązuje zasada dobrowolnego budowania „ukryć", i to za własne pieniądze. — Państwo gotowe jest jednak uczestniczyć w kosztach utrzymania schronów — mówi, niezbyt zresztą pewny swych słów, młodszy brygadier Witold Maziarz z Komendy Głównej PSP, reprezentujący szefa OC.

Warszawscy funkcjonariusze Obrony Cywilnej przyznają, że za własne fundusze wybudowano w ostatnich latach w stolicy i okolicach pojedyncze „ukrycia" pod mieszkalnymi apartamentami. Właściciele zastrzegają kategorycznie poufność takich informacji.

Pułkownik Baranowicz za przykład wzorcowych wręcz inwestycji podaje schrony pod świeżo oddanymi ratuszami na warszawskim Bemowie czy Targówku. Mogą już dziś służyć do natychmiastowego awaryjnego zakwaterowania kilkudziesięciu osób. To, że gospodarze dzielnic zatroszczyli się o ulokowanie w ratuszach normalnych schronów, jest wśród stołecznych inwestorów wyjątkowe. W żadnym z nowo wznoszonych w Warszawie biurowców nie ma schronów, mimo że na całym świecie „ukrycia" tworzy się wszędzie, na przykład wzmacniając stropy w podziemnych garażach. Podwójne przeznaczenie wielu obiektów w Szwecji, Szwajcarii czy Finlandii to standard. W tym ostatnim kraju ponad 70 proc. społeczeństwa ma swoje miejsca w „ukryciach". W Szwajcarii ten wskaźnik jest jeszcze wyższy. -

C Przykładowe fragmenty analizy

Poniżej przedstawiamy przykładowe fragmenty analizy dokonanej przez poszczególne analizatory. W każdym przypadku dotyczą one początkowego fragmentu pliku `Publicystyka`.

C.1 SAM

```
W %
{{() < w(E::0)+ } }%
sieni %
{{(GDVlG) < sień(żV::)+ } }%
szarej %
{{(7) < szary(A::[p])+ } }%
kamienicy %
{{(GD) < kamienica(żII::)+ } }%
przy %
{{() < przy(::0)+ } }%
ulicy %
{{(lH) < ulik(mIII::m2 I)+ }%
{(GD) < ulica(żII::)+ } }%
Krasińskiego %
{ }%
na %
{{() < na(::0 I)+ } }%
warszawskim %
{{(4) < warszawski(A::[p])+ } }%
Żoliborzu %
{ }%
nie %
{{() < nie(::0)+ }%
{(por.) < on(Zc::+ I)+ } }%
świeci %
{{(3) < świecić(VIa::ndk (się))+ } }%
się %
{{(N-akcGT) < się(Zb::+)+ } }%
żarówka, %
{{(N) < żarówka(żIII::)+ } }%
lecz %
{{() < lecz(::0)+ }%
{(i) < leczyć(VIb::ndk (się))+ } }%
jaskrawe %
{{(5) < jaskrawy(A::[ps])+ } }%
uchwyty %
{{(lN) < uchwyt(mIV::m3)+ } }%
stalowych %
{{(10) < stalowy(A::[p])+ } }%
drzwi %
{{(lNG) < drzwi(blp::)+ } }%
w %
{{() < w(E::0)+ } }%
półmroku %
{{(GLV) < półmrok(mIII::m3)+ } }%
wskazują %
{{(l3) < wskazywać(VIIIa::ndk)+ } }%
drogę. %
{{(T) < droga(żIII::)+ } }%
Słuchać %
{{(B) < słuchać(I::)+ } }%
zgrzyt %
{{(N) < zgrzyt(mIV::m3)+ } }%
ciężkich %
{{(10) < ciężki(A::[ps])+ } }%
zasuw %
{{(N) < zasuw(mIV::m3)+ }%
{(lG) < zasuwa(żIV::)+ } }%
i %
{{() < i(::0 I)+ } }%
„obronno-ochronne” %
{ }%
hermetyczne %
{{(5) < hermetyczny(A::[p])+ } }%
odrzwia %
{{(lN) < odrzwia(blp::)+ } }%
cicho %
{{() < cicho(J::s)+ } }%
ustępują. %
{{(l3) < ustępować(IV::ndk)+ } }%
Jeszcze %
{{() < jeszcze(::0)+ } }%
kilka %
{{(1) < kilka(Ka::+ I)+ }%
{(N) < kilka(żIII::II)+ } }%
schodków %
{{(lG) < schodek(mIII::m3)+ } }%
w %
{{() < w(E::0)+ } }%
dół %
{{(N) < dół(mIV::m3)+ } }%
i %
{{() < i(::0 I)+ } }%
jesteśmy %
{{(l1') < być(XII::ndk)+ } }%
w %
{{() < w(E::0)+ } }%
bunkrze, %
{{(LV) < bunkier(mIV::m3)+ } }%
czyli %
{{() < czyli(::0)+ } }%
jednym %
{{(4) < jeden(A::)+ } }%
z %
{{() < z(E::0)+ } }%
2100 stołecznych %
{{(10) < stołeczny(A::[p])+ } }%
schronów. %
{{(lG) < schron(mIV::m3)+ } }%
```

C.2 PoMor

```
Analysis of ”W”:
      w[E]=W
Analysis of ”sieni”:
      sień[Sf]=sie+ni[21]
      sień[Sf]=sie+ni[31]
      sień[Sf]=sie+ni[61]
      sień[Sf]=sie+ni[71]
      sień[Sf]=sie+ni[22]
Analysis of ”szarej”:
      szary[Adj]=szar+ej[07]
Analysis of ”kamienicy”:
      kamienica[Sf]=kamienic+y[21]
      kamienica[Sf]=kamienic+y[31]
      kamienica[Sf]=kamienic+y[61]
Analysis of ”przy”:
      przy[E]
Analysis of ”ulicy”:
      ulica[Sf]=ulic+y[21]
      ulica[Sf]=ulic+y[31]
      ulica[Sf]=ulic+y[61]
Analysis of ”Krasińskiego”:
      Krasiński[AdjS]=Krasiń+skiego[02]
Analysis of ”na”:
      na[ndm]
      na[E]
Analysis of ”warszawskim”:
      warszawski[Adj]=warszaw+skim[04]
      warszawski[Adj]=warszaw+ski[01]+m[Vsg1]
Analysis of ”Żoliborzu”:
      Żoliborz[Sm3]+u[61]
      Żoliborz[Sm3]+u[71]
      Żolibórz[Sm3]=Żoliborz+u[61]
      Żolibórz[Sm3]=Żoliborz+u[71]
Analysis of ”nie”:
      nie[ndm]
      on[Zc]=nie+[A0k2]
      on[Zc]=nie+[A0n2]
Analysis of ”świeci”:
      świecić[Vndk]=świe+ci[3]
Analysis of ”się”:
      się[ndm]
      się[Zb]+[G1]
      się[Zb]+[A1]
      się[Zb]+[N0]
Analysis of ”żarówka,”:
      żarówka[Sf]=żarów+ka[11]
Analysis of ”lecz”:
      lecz[ndm]
      leczyć[Vndk]=lecz+[i]
Analysis of ”jaskrawe”:
      jaskrawy[Adj]=jaskraw+e[05]
Analysis of ”uchwyty”:
      uchwyt[Sm3]=uchwy+ty[12]
      uchwyt[Sm3]=uchwy+ty[42]
      uchwyt[Sm3]=uchwy+ty[72]
Analysis of ”stalowych”:
      stalowy[Adj]=stalow+ych[10]
Analysis of ”drzwi”:
      drzwi[Splt]+[12]
      drzwi[Splt]+[22]
      drzwi[Splt]+[42]
      drzwi[Splt]+[72]
Analysis of ”w”:
      w[E]
Analysis of ”półmroku”:
      półmrok[Sm3]=półmro+ku[21]
      półmrok[Sm3]=półmro+ku[61]
      półmrok[Sm3]=półmro+ku[71]
Analysis of ”wskazują”:
      wskazywać[Vndk]=wskaz+ują[13]
Analysis of ”drogę.”:
      droga[Sf]=dr+ogę[41]+.[.]
Analysis of ”Słuchać”:
      słuchać[Vndk]=Słucha+ć[B]
Analysis of ”zgrzyt”:
      zgrzyt[Sm3]=zgrzy+t[11]
      zgrzyt[Sm3]=zgrzy+t[41]
Analysis of ”ciężkich”:
      ciężki[Adj]=cięż+kich[10]
Analysis of ”zasuw”:
      zasuw[Sm3]+[11]
      zasuw[Sm3]+[41]
      zasuwa[Sf]=zasu+w[22]
Analysis of ”i”:
      i[ndm]
Analysis of ”obronno-ochronne”:
      obronno[Adv]+-[lacz1]+ochronny[Adj]
                  =ochronn+e[05]
      obronny[Adj]=obronn+o[psk]+-[lacz2]
                  +ochronny[Adj]=ochronn+e[05]
Analysis of ”hermetyczne”:
      hermetyczny[Adj]=hermetyczn+e[05]
Analysis of ”odrzwia”:
      odrzwia[Splt]=odrzwi+a[12]
      odrzwia[Splt]=odrzwi+a[42]
      odrzwia[Splt]=odrzwi+a[72]
Analysis of ”cicho”:
      cichy[Adj]=ci+cho[J]
Analysis of ”ustępują.”:
      ustępować[Vndk]=ustęp+ują[13]+.[.]
Analysis of ”Jeszcze”:
      jeszcze[ndm]=Jeszcze
Analysis of ”kilka”:
      kilka[Ka]+[1]
      kilka[Sf]=kil+ka[11]
Analysis of ”schodków”:
      schodek[Sm3]=schod+ków[22]
Analysis of ”w”:
      w[E]
Analysis of ”dół”:
      dół[Sm3]=d+ół[11]
      dół[Sm3]=d+ół[41]
Analysis of ”i”:
      i[ndm]
```

C.3 LEM

```
<W w,P/Cal>W
<B>
<W sień,N/GfNpCg><W sień,N/GfNsCd>
<W sień,N/GfNsCg><W sień,N/GfNsCl>
<W sień,N/GfNsCv>sieni
<B>
<W szary,ADJ/DpNsCdglGf>szarej
<B>
<W kamienica,N/GfNsCd><W kamienica,N/GfNsCg>
<W kamienica,N/GfNsCl>kamienicy
<B>
<W przy,P/Cl>przy
<B>
<W ulica,N/GfNsCd><W ulica,N/GfNsCg>
<W ulica,N/GfNsCl>ulicy
<B>
<W>Krasińskiego
<B>
<W na,P/Cal>na
<B>
<W warszawski,ADJ/DpNpCd>
<W warszawski,ADJ/DpNsCilGpain>warszawskim
<B>
<W>Żoliborzu
<B>
<W nie,PART><W one,NPRO/ZpNpGaifnCa>
<W ono,NPRO/ZpNsGnCa>nie
<B>
<W świecić,V/AiVpMdTrfP3Ns>świeci
<B>
<W się,NPRO/ZxN*G*Ca><W się,NPRO/ZxN*G*Cg>się
<B>
<W żarówka,N/GfNsCn>żarówka
<P>,
<B>
<W lecz,CONJ><W leczyć,V/AiVpMiP2Ns>lecz
<B>
<W jaskrawy,ADJ/DpNpCnavGaifn>
<W jaskrawy,ADJ/DpNsCnavGn>jaskrawe
<B>
<W uchwyt,N/GiNpCa><W uchwyt,N/GiNpCn>
<W uchwyt,N/GiNpCv>uchwyty
<B>
<W stalowy,ADJ/DpNpCaGp>
<W stalowy,ADJ/DpNpCgl>stalowych
<B>
<W drzwi,N/G*NpCa><W drzwi,N/G*NpCg>
<W drzwi,N/G*NpCn><W drzwi,N/G*NpCv>drzwi
<B>
<W w,P/Cal>w
<B>
<W półmrok,N/GiNsCg><W półmrok,N/GiNsCl>
<W półmrok,N/GiNsCv>półmroku
<B>
<W wskazywać,V/AiVpMdTrfP3Np>wskazują
<B>
<W droga,N/GfNsCa>drogę
<P>.
<B>
<W słuchać,V/AiVb>Słuchać
<B>
<W zgrzyt,N/GiNsCa><W zgrzyt,N/GiNsCn>zgrzyt
<B>
<W ciężki,ADJ/DpNpCaGp>
<W ciężki,ADJ/DpNpCgl>ciężkich
<B>
<W zasuwa,N/GfNpCg>zasuw
<B>
<W i,CONJ><W i,EXCL>i
<B>
<P>“
<W>obronno
<P>-
<W ochronny,ADJ/DpNpCanvGaifn>
<W ochronny,ADJ/DpNsCanvGn>ochronne
<P>”
<B>
<W hermetyczny,ADJ/DpNpCnavGaifn>
<W hermetyczny,ADJ/DpNsCnavGn>hermetyczne
<B>
<W odrzwia,N/G*NpCa><W odrzwia,N/G*NpCn>
<W odrzwia,N/G*NpCv>odrzwia
<B>
<W cicho,ADV/Dp>cicho
<B>
<W ustępować,V/AiVpMdTrfP3Np>ustępują
<P>.
<B>
<W jeszcze,PART>Jeszcze
<B>
<W kilka,NUMCRD/SpZiCaGaifn>
<W kilka,NUMCRD/SpZiCnGaifn>kilka
<B>
<W schodek,N/GiNpCg>schodków
<B>
<W w,P/Cal>w
<B>
<W dół,N/GiNsCa><W dół,N/GiNsCn>dół
<B>
<W i,CONJ><W i,EXCL>i
<B>
<W być,BYC/VpMdTrfP1Np>jesteśmy
<B>
<W w,P/Cal>w
<B>
<W bunkier,N/GiNsCl>
<W bunkier,N/GiNsCv>bunkrze
<P>,
<B>
```

C.4 XeLDA

```
W
    »w+Prep+Loc+PREP
    »w+Prep+Acc+PREP
sieni
    »sień+Noun+Fem+Sg+Dat+NOUN
    »sień+Noun+Fem+Sg+Gen+NOUN
    »sień+Noun+Fem+Sg+Loc+NOUN
    »sień+Noun+Fem+Sg+Voc+NOUN
    »sień+Noun+Fem+Pl+Gen+NOUN
szarej
    »szary+Adj+Fem+Sg+GenDatLoc+ADJ
kamienicy
    »kamienica+Noun+Fem+Sg+Dat+NOUN
    »kamienica+Noun+Fem+Sg+Gen+NOUN
    »kamienica+Noun+Fem+Sg+Loc+NOUN
przy
    »przy+Prep+Loc+PREP
ulicy
    »ulica+Noun+Fem+Sg+Dat+NOUN
    »ulica+Noun+Fem+Sg+Gen+NOUN
    »ulica+Noun+Fem+Sg+Loc+NOUN
Krasińskiego
    »Krasiński+Prop+M1+Sg+Acc+NOUN
    »Krasiński+Prop+M1+Sg+Gen+NOUN
na
    »na+Prep+Loc+PREP
    »na+Prep+Acc+PREP
         warszawskim
    »warszawski+Adj+MN+Sg+InsLoc+ADJ
    »warszawski+Adj+M12+Sg+Nom=być+1P+Sg+ADJ
    »warszawski+Adj+M23FN+Pl+Dat+ADJ
    »warszawski+Adj+M3+Sg+Nom=być+1P+Sg+ADJ
    »warszawski+Adj+M1+Pl+Ins+ADJ
    »warszawski+Adj+M1+Pl+Dat+ADJ
Żoliborzu
    »Żolibórz+Prop+M3+Sg+Loc+NOUN
    »Żolibórz+Prop+M3+Sg+Voc+NOUN
    »Żoliborz+Prop+M3+Sg+Loc+NOUN
    »Żoliborz+Prop+M3+Sg+Voc+NOUN
nie
    »one+Pron+Pers+3P+Pl+M23+Acc+PPERS
    »nie+Partcl+PRTCL
świeci
    »świecić+Verb+Imperf+Ind+
         Pres+3P+Sg+VFIN
się
    »się+Pron+Refl+Acc+PREFL
    »się+Pron+Refl+Gen+PREFL
    »się+Pron+General+PRON
żarówka
    »żarówka+Noun+Fem+Sg+Nom+NOUN
,
    »,+Punct+Comma+CM
lecz
    »leczyć+Verb+Imperf+Impv+2P+Sg+VFIN
    »leczo+Noun+Neut+Pl+Gen+NOUN
    »lecz+Conj+CONJ
jaskrawe
    »jaskrawy+Adj+Neut+Sg+NomAccVoc+ADJ
    »jaskrawy+Adj+M23FN+Pl+NomAccVoc+ADJ
uchwyty
    »uchwyt+Noun+M3+Pl+Acc+NOUN
    »uchwyt+Noun+M3+Pl+Nom+NOUN
    »uchwyt+Noun+M3+Pl+Voc+NOUN
stalowych
    »stalowy+Adj+M23FN+Pl+GenLoc+ADJ
    »stalowy+Adj+M1+Pl+GenAccLoc+ADJ
drzwi
    »drzwi+Noun+M23FN+Pl+Voc+NOUN
    »drzwi+Noun+M23FN+Pl+Nom+NOUN
    »drzwi+Noun+M23FN+Pl+Gen+NOUN
    »drzwi+Noun+M23FN+Pl+Acc+NOUN
w
    »w+Prep+Loc+PREP
    »w+Prep+Acc+PREP
półmroku
    »półmrok+Noun+M3+Sg+Acc+NOUN
    »półmrok+Noun+M3+Sg+Gen+NOUN
    »półmrok+Noun+M3+Sg+Loc+NOUN
    »półmrok+Noun+M3+Sg+Voc+NOUN
wskazują
    »wskazywać+Verb+Imperf+Ind+
         Pres+3P+Pl+VFIN
drogę
    »droga+Noun+Fem+Sg+Acc+NOUN
.
    ».+Punct+Final+SENT
Słuchać
    »słuchać+Verb+Imperf+Infinit+VINF
zgrzyt
    »zgrzyt+Noun+M3+Sg+Nom+NOUN
    »zgrzyt+Noun+M3+Sg+Acc+NOUN
ciężkich
    »ciężki+Adj+M1+Pl+GenAccLoc+ADJ
    »ciężki+Adj+M23FN+Pl+GenLoc+ADJ
zasuw
    »zasuwa+Noun+Fem+Pl+Gen+NOUN
    »zasuw+Noun+M3+Sg+Nom+NOUN
    »zasuw+Noun+M3+Sg+Acc+NOUN
i
    »i+Partcl+PRTCL
    »i+Conj+CONJ
”
    »”+Punct+Misc+PUNCT
obronno-ochronne
    »obronno-ochronny+Adj+Neut+Sg+
         NomAccVoc+guessed+ADJ
    »obronno-ochronny+Adj+M23FN+Pl+
         NomAccVoc+guessed+ADJ
”
    »”+Punct+Misc+PUNCT
hermetyczne
    »hermetyczny+Adj+Neut+Sg+NomAccVoc+ADJ
    »hermetyczny+Adj+M23FN+Pl+NomAccVoc+ADJ
```

C.5 AMOR

```
W
   w (przyimki) przyimek
sieni
   sień (zV) wołacz lp, żeński
   sień (zV) dopełniacz lp, żeński
   sień (zV) celownik lp, żeński
   sień (zV) miejscownik lp, żeński
   sień (zV) dopełniacz lm, żeński
szarej
   szary (przymII) przymiotnik,
        dopełniacz lp, żeński
   szary (przymII) przymiotnik,
        celownik lp, żeński
   szary (przymII) przymiotnik,
        miejscownik lp, żeński
kamienicy
   kamienica (zII) celownik lp, żeński
   kamienica (zII) dopełniacz lp, żeński
   kamienica (zII) miejscownik lp, żeński
przy
   przy (przyimki) przyimek
ulicy
   ulica (zII) celownik lp, żeński
   ulica (zII) dopełniacz lp, żeński
   ulica (zII) miejscownik lp, żeński
Krasińskiego
   Krasiński (rzp) dopełniacz lp,
        męskoosobowy
   Krasiński (rzp) biernik lp, męskoosobowy
na
   na (przyimki) przyimek
warszawskim
   warszawski (przI) przymiotnik,
        miejscownik lp, męski/nijaki
   warszawski (przI) przymiotnik,
        celownik lm
   warszawski (przI) przymiotnik,
        narzędnik lp, męski/nijaki
Żoliborzu
nie
   ono (zai) zaimek osobowy, os.3,
        biernik lp, nijaki
   nie (part) partykuła
świeci
   świecić (c6a) 3 lp,
        czas teraźniejszy, ndk
się
   siebie (zaimI) zaimek zwrotny, biernik
   się (part) partykuła
żarówka
   żarówka (zIII) mianownik lp, żeński
lecz
   lecz (spójniki) spójnik
   leczyć (c6b) 2 lp, tryb rozkazujący, ndk
jaskrawe
   jaskrawy (przymII) przymiotnik,
        mianownik lp, nijaki
   jaskrawy (przymII) przymiotnik,
        biernik lp, nijaki
   jaskrawy (przymII) przymiotnik,
        biernik lm, niemęskoosobowy
   jaskrawy (przymII) przymiotnik,
        mianownik lm, niemęskoosobowy
uchwyty
   uchwyt (mIV) biernik lm, męskorzeczowy
   uchwyt (mIV) mianownik lm, męskorzeczowy
   uchwyt (mIV) wołacz lm, męskorzeczowy
stalowych
   stalowy (przymII) przymiotnik,
        miejscownik lm
   stalowy (przymII) przymiotnik,
        biernik lm, męskoosobowy
   stalowy (przymII) przymiotnik,
        dopełniacz lm
drzwi
   drzwi (plur) mianownik lm, nijaki
   drzwi (plur) dopełniacz lm, nijaki
   drzwi (plur) biernik lm, nijaki
w
   w (przyimki) przyimek
półmroku
   półmrok (mIII) wołacz lp, męskorzeczowy
   półmrok (mIII) dopełniacz lp,
        męskorzeczowy
   półmrok (mIII) miejscownik lp,
        męskorzeczowy
wskazują
   wskazywać (c8a) 3 lm,
        czas teraźniejszy, ndk
drogę
   droga (zIII) biernik lp, żeński
Słuchać
   słuchać (cI) bezokolicznik, czasownik ndk
zgrzyt
   zgrzyt (mIV) mianownik lp, męskorzeczowy
   zgrzyt (mIV) biernik lp, męskorzeczowy
ciężkich
   ciężki (przI) przymiotnik, miejscownik lm
   ciężki (przI) przymiotnik, dopełniacz lm
   ciężki (przI) przymiotnik, biernik lm,
        męskoosobowy
zasuw
   zasuwa (zIV) dopełniacz lm, żeński
i
   i (part) partykuła
   i (spójniki) spójnik
obronno
ochronne
   ochronny (przymII) przymiotnik,
        mianownik lp, nijaki
   ochronny (przymII) przymiotnik,
        biernik lp, nijaki
   ochronny (przymII) przymiotnik,
        biernik lm, niemęskoosobowy
   ochronny (przymII) przymiotnik,
        mianownik lm, niemęskoosobowy
```

C.6 Gram

Spis treści

Pracę zgłosił: Leonard Bolc

Adres autorek: Elżbieta Hajnicz
Anna Kupść
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Ordona 21
01-237 Warszawa

e-mail: Elzbieta.Hajnicz@ipipan.waw.pl
Anna.Kupsc@ipipan.waw.pl

Symbole klasyfikacji rzeczowej: CR: I.2.7

Nakład 140 egzemplarzy. Papier kserograficzny klasy III. Oddano do druku w grudniu 2001. Wydawnictwo IPI PAN.
ISSN 0138-06-48.