

JANUSZ S. BIENIŃ, KRZYSZTOF SZAFRAN
Warszawa

Analiza morfologiczna języka polskiego w praktyce

1 Co to jest analiza morfologiczna?

Przez *analizę morfologiczną* rozumiemy pewną operację (program lub algorytm ją realizujący nazwiemy *analizatorem morfologicznym*), która dla każdego słowa stanowiącego dane wejściowe produkuje pewien jego opis. Aby dokładnie opisać dane wejściowe analizy, trzeba wskazać, co traktujemy jako słowo. Bez względu na to, czy interesuje nas tekst pisany czy mówiony, słowo jest napisem, tj. ciągiem znaków. Znaki te mogą odpowiadać literom i innym znakom piśmiennym lub mniej lub bardziej pośrednio reprezentować wymowę słowa. W przypadku słów pisanych trzeba w szczególności podjąć decyzję, jak traktujemy tzw. formy złożone, takie jak *będę czytać* czy *bardziej pociągający*. Trzeba również rozwiązać przeciwstawny problem, które słowa traktujemy jako dwa lub więcej wyrazów pisanych łącznie — jak już zwracano na to uwagę [28], traktowanie jako jednostki np. słowa *Czyś* w zdaniu «*Czyś to wiedział?*» wymagałoby w konsekwencji uznania, że wyraz *CZY* odmienia się przez osobę¹.

Kwestie te dla użytkownika programu analizy morfologicznej mają duże znaczenie praktyczne, bo określają sposób przygotowania danych wejściowych. Aby móc precyzyjnie wypowiadać się na ten temat, wskazane jest wprowadzenie m.in. takich pojęć jak *wyraz alfabetyczny* i *wyraz grafemiczny* ([4]). Nie będziemy jednak tutaj rozwijać tego tematu, bo o wiele istotniejsza jest forma wyniku analizy morfologicznej.

Etymologicznie „analiza” to rozkładanie, rozbiór, zaś słownik PAN [10] podaje m.in. taką definicję tego wyrazu «... myślowe wyodrębnienie cech, właściwości lub składników badanego zjawiska czy też przedmiotu ...». Analiza morfologiczna to więc przypisanie analizowanemu słowu pewnych własności

¹Postaci hasłowe wyrazów, reprezentujące leksemy lub wyrazy paradygmatyczne, zapisujemy kapitalikami. Cudzysłowy francuskie stosujemy do cytowania fraz zawierających interpunkcję.

morfologicznych, wśród których może być lecz nie musi rozkład tego słowa na elementy prostsze. Wynika stąd jasno, że mamy tyle różnych analiz morfologicznych, ile jest różnych definicji morfologii. Nawet w ramach jednej definicji morfologii konkretne analizatory morfologiczne mogą się różnić pod względem zestawu cech, które przypisują analizowanemu słowu.

Jedną z najważniejszych cech, które może przypisywać słowu analiza morfologiczna, jest jego *postać hasłowa*. W wielu zastosowaniach praktycznych informacja o postaci hasłowej jest całkowicie wystarczająca, a operacja jej ustalenia — wykonywana zarówno ręcznie jak i automatycznie — nosi nazwę *hasłowania*.

Jeśli chcemy jako wynik analizy otrzymać bogatszą informację niż tylko postać hasłowa, to stajemy przed problemem, jaki zestaw kategorii morfologicznych wybrać, i jaki przyjąć dla nich repertuar wartości. Szczególny kłopot sprawia tutaj kategoria rodzaju — dowodem na to, że jest to ciągle temat kontrowersyjny, jest m.in. najnowsza propozycja Marcina Wolińskiego wyróżniania 8 wartości kategorii rodzaju ([35]). Problemów tych można częściowo uniknąć, traktując oddzielnie kategorie morfologiczne — w wąskim rozumieniu tego słowa — i kategorie morfosyntaktyczne, patrz [4].

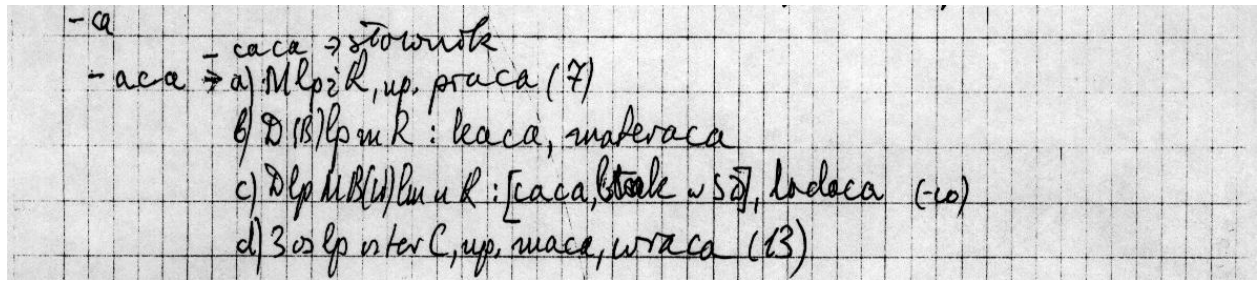
Za skrajny przykład analizy morfologicznej można uznać automatyczną korektę ortograficzną, która przyporządkowuje analizowanemu słowu tylko informację, czy jest ono uznane za poprawne.

2 Analizator SAM

2.1 Indeks Tokarskiego

Jan Tokarski był jednym z redaktorów największego słownika języka polskiego [10] składającego się z 11 tomów i zawierającego ponad 120 000 haseł. Był on odpowiedzialny pierwotnie za projekt morfologicznej części haseł. W czasie wieloletniej pracy nad słownikiem gromadził dane językowe związane z analizą morfologiczną. Dane te były zorganizowane w postaci unikalnego rękopiśmiennego indeksu, zawierającego zarówno ogólne reguły jak też wyczerpującą listę wyjątków (patrz rysunek 1); rękopis ten znajduje się obecnie w Archiwum Polskiej Akademii Nauk w kolekcji dokumentów biograficznych do dziejów kultury polskiej (symbol III-369, pozycja wykazu 19).

Niestety stan zdrowia nie pozwolił Tokarskiemu na dokończenie pracy nad indeksem; dalsze jego rozwijanie przejął Zygmunt Saloni, który opublikował ostateczną wersję, zatytułowaną *Schematyczny indeks a tergo polskich form wyrazowych Jana Tokarskiego* ([30], por. także [12] i [27]), kilka lat po śmierci To-



Rysunek 1: Fragment rękopisu prof. Tokarskiego.

karskiego. Saloni używał komputera we wszystkich stadiach pracy redakcyjnej, co było możliwe dzięki pomocy Instytutu Informatyki Uniwersytetu Warszawskiego.

(mam XII I	mieć	
-mam mIV N		imam, omam
mam žIV IG	mama	
-mam I I	-mać	mniemam, imam, dumam, trzymam (70)
mam VIa i	mamić	omam (4)
(nam Za D	my	
-nam mIV N		ignam, Uznam
-nam nIII IG	-namo	dynam
-nam žIV IG	-nama	panam, izodynam!
-nam I I	-nać	żegnam, zapinam, przekonam, zrzymam, dożynam (200)
-nam 2 formy	-na +)m	winnam, powinnam (patrz: Vm)
-pam I I	-pać	ćpam, stąpam (15)

Rysunek 2: Fragment indeksu Tokarskiego

Spójrzmy na przykładowy fragment indeksu pokazany na rysunku 2.

Każda reguła, z bardzo nielicznymi wyjątkami, zajmuje dokładnie jeden wiersz i składa się z czterech pól.

Pierwsze z nich traktowane jest jako nagłówek pozycji indeksu. Reguły ułożone są w porządku alfabetycznym odwróconym (*a tergo*) ze względu na to pole. Zawiera ono wzorzec, który określa dla jakich słów dana reguła ma zastosowanie. Wzorzec ten może odpowiadać zarówno końcowej części słowa, jak np. w regułach

-mam mIV N		imam, omam
-mam I I	-mać	mniemam, imam, dumam, trzymam (70)

jak też całemu słowu lub części końcowej słowa, jak np. w regułach

mam žIV IG	mama	
mam VIa i	mamić	omam (4)

albo wyłącznie całemu (jednemu) słowu, np.

(mam XII I	mieć	
------------	------	--

Tak więc te ostatnie reguły tworzą słownik wyjątków, podczas gdy reguły pierwszego i drugiego rodzaju są regułami ogólnymi. W szczególności reguły ogólne mogą mieć zastosowanie również do słów nowych, nie notowanych jeszcze w słownikach.

Drugie pole specyfikuje własności morfologiczne słów, które mogą być analizowane z wykorzystaniem reguły opisywanej przez dany artykuł hasłowy. Ze względu na oszczędność miejsca używana notacja jest bardzo zwarta, choć nie zawsze łatwo czytelna. Na przykład, słowo *mam* analizowane zgodnie z pierwszą regułą z rysunku 2. jest czasownikiem należącym do XII grupy koniugacyjnej² w pierwszej osobie czasu teraźniejszego.

Trzecie pole zawiera instrukcję opisującą, w jaki sposób utworzyć postać hasłową dla analizowanego słowa; w przypadku słowa *mam* może to być zarówno MAMA (rzeczownik — reguła trzecia, rysunek 2) jak też MIEĆ (czasownik — reguła pierwsza, tamże).

Czwarte pole zawiera przykłady, które mniej lub bardziej bezpośrednio pokazują, jak produktywna jest dana reguła; w szczególności liczba w nawiasie podaje przybliżoną liczbę odpowiednich artykułów hasłowych w słowniku Doroszewskiego³.

Oczywiście, nie jest tutaj możliwe podanie pełnego opisu reguł indeksu Tokarskiego, nawet w naszym bardzo niewielkim przykładzie znajdują się pewne konstrukcje, które wymagająby dłuższych wyjaśnień.

2.1.1 Komputerowa realizacja indeksu Tokarskiego

Dzięki dysponowaniu tekstem indeksu na nośniku komputerowym⁴, pierwsza implementacja opartego na nim analizatora morfologicznego — nazwanego Systemem Analizy Morfologicznej — była gotowa niemal w tym samym momencie, w którym ukazało się pierwsze wydanie drukowanej wersji indeksu. Implementacja ta została przygotowana przez Krzysztofa Szafrana dla potrzeb jego pracy doktorskiej ([24]), a następnie rozbudowana do wersji SAM-95 ([25], [26]).

Działanie analizatora zilustrujemy wynikiem przetworzenia tekstu «*Psa mam tam.*». Oto on:

```
Psa %  
{ {(G) < pies(mIV)+ } } %  
mam %
```

²Tak w *Indeksie* oznaczane są czasowniki nieregularne, w słownikach występujące bez numeru grupy.

³Brak liczby oznacza, że pole zawiera wszystkie znane redaktorowi indeksu formy opisywane danym wierszem.

⁴Nie jest przypadkiem, że skład komputerowy i łamanie książki wykonał Krzysztof Szafran.

```

{ { *m(6!) < mój(A)+ } }%
{ *m(3) < mieć(XII)+ }%
{ (1) < mieć(XII)+ }%
{ (1G) < mama(żIV)+ }%
{ (i) < mamić(VIa)+ } }%
tam.%
{ { *m() < ta()+ } }%
{ *m(6) < ten(A)+ }%
{ () < tam()+ }%
{ (1G) < tama(żIV)+ } }%

```

Jak widać, tylko pierwsze słowo ma jednoznaczną interpretację (na poziomie morfologicznym, bo na poziomie morfosyntaktycznym mamy synkretyzm dopełniacza i biernika). Dla pozostałych słów występujące niejednoznaczności są dwojakiego rodzaju. Pierwsze, zdecydowanie ważniejsze mają charakter czysto morfologiczny — *tam* może oznaczać zarówno zaimek wskazujący TAM jak też dopełniacz liczby mnogiej rzeczownika TAMA.

Drugi rodzaj niejednoznaczności związany jest z faktem, że pewne słowa w tekście polskim mogą reprezentować nie jedną, ale dwie zapisane łącznie formy. Zjawisko to potraktowane zostało w indeksie marginesowo — przedstawiony tam opis dopuszcza interpretacje praktycznie nie występujące w języku polskim. Co więcej, nie istnieje zadowolający opis tego zjawiska. W konsekwencji wszelkie interpretacje *mam* jako *ma+m* i *tam* jako *ta+m* powinny zostać zignorowane.

Analizator SAM operuje nie tylko w zakresie niemal 120 000 haseł wspomnianego wcześniej słownika PAN, ale może również — w przypadku ewentualnych haseł nie notowanych w słowniku — sugerować klasyfikację morfologiczną i odpowiadającą jej postać hasłową. Dla celów badawczych dostępny jest bezpłatnie pod adresem <ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/>.

3 Zastosowania analizatora SAM

3.1 Analiza syntaktyczna

Najwcześniejszym zastosowaniem analizatora SAM było jego wykorzystanie w eksperymentach zmierzających do stworzenia analizatora syntaktycznego oparteo na gramatyce formalnej Marka Świdzińskiego ([29]); wyniki tych eksperymentów zostały przedstawione m.in. w referatach [2] i [6], są one dostępne również w Internecie ([3]).

Kategorie morfologiczne wykorzystywane w gramatyce Świdzińskiego są w istocie kategoriami morfosyntaktycznymi, stąd konieczność konwersji wyników analizatora wyrażonych w nietradycyjnych kategoriach czysto morfologicznych. Z technicznego punktu widzenia konwersja ta nie stwarza żadnych problemów, ponieważ i tak informacje uzyskane w wyniku analizy muszą być uzupełnione np. o wymagania składniowe czasowników. Te dodatkowe informacje pobierane są z odpowiedniego słownika; w trakcie tej operacji można nie tylko posiadane informacje uzupełnić, ale — w przypadku różnego rodzaju wyjątków — można i należy zastąpić je całkowicie innym zestawem własności. Dotyczy to m.in. wyrazów, które analizator SAM słusznie traktuje jako nieodmienne, a które składniowo są np. przymiotnikami (*khaki*). Inny przykład to rzeczowniki męskie. Dla typowego takiego rzeczownika analizator rozpoznaje, że jest on rodzaju męskiego i trzeba na tym etapie wskazać tylko, o który podrodzaj chodzi — jest to niezbędne dla ustalenia, jakie synkretyzmy morfologiczne należy uwzględnić przy przejściu na poziom morfosyntaktyczny. Jednak rzeczownikom męskim odmiennym jak rzeczowniki żeńskie analizator przypisuje rodzaj żeński — wartość tę należy całkowicie wyeliminować i zastąpić odpowiednią wartością rodzaju męskiego.

Warto podkreślić, że choć rozróżnienie poziomu czysto morfologicznego i morfosyntaktycznego zostało wprowadzone już w książce [1], to właśnie prace nad zintegrowaniem analizatora morfologicznego z analizatorem syntaktycznym uzmysłowiły w pełni korzyści płynące z tego rozróżnienia.

3.2 Korpus słownika frekwencyjnego

Korpus słownika frekwencyjnego to pięć zestawów próbek po 100 000 słów wylosowanych z autentycznych tekstów z lat 1963–1967 należących do 5 stylów — tekstów popularnonaukowych, drobnych wiadomości prasowych, publicystyki, prozy artystycznej i dramatu artystycznego — na potrzeby badań frekwencji słów języka polskiego ([19]).

Pierwotnie korpus miał formę taśm papierowych wyperforowanych na dalekopisie (czego konsekwencją był brak rozróżnienia małych i dużych liter). Został on wczytany do komputera przez Bronisława Rocławskiego (wówczas na Uniwersytecie Gdańskim) i zapisany na taśmie magnetycznej; niestety, w trakcie tej operacji do korpusu wkradły się pewne przekłamania. Taśma magnetyczna została zapisana na komputerze ODRA 1204 w standardzie, który szybko wyszedł z użycia. W związku z tym taśma z korpusem trafiła w ręce Krzysztofa Szafrana, który w Instytucie Informatyki Uniwersytetu Warszawskiego na podstawie list frekwencyjnych dla poszczególnych stylów opracowywał tzw.

tom zbiorczy, opublikowany jako *Słownik frekwencyjny polszczyzny współczesnej* [20]. Krzysztof Szafran za pomocą specjalnie przygotowanego programu odczytał taśmę na komputerze SM-4 i zapisał jej zawartość na bardziej nowoczesnych nośnikach, w wyniku czego korpus stał się dostępny również na dyskietkach stosowanych w komputerach osobistych.

Choć słownictwo korpusu jest już częściowo przestarzałe, korpus ten nadal ma dużą wartość m.in. dla badań składniowych. Z tego względu Janusz S. Bień wystąpił z inicjatywą dokonania korekty korpusu i udostępnienia go w bardziej nowoczesnej formie. Pierwszy krok w tym kierunku stanowiła zrealizowana pod jego kierunkiem praca magisterska Marty Nazarczuk ([22]). Janusz S. Bień wykonał również eksperyment polegający na przetworzeniu stylu popularnonaukowego dwoma korektorami ortograficznymi: polskiej firmy TiP i węgierskiej firmy Morphologic ([21, s. 153], [37]); współautor tego drugiego narzędzia, Robert Wołosz, przetworzył nim również inne style i udostępnił nam wyniki. Operacje te pozwoliły nie tylko wykryć błędy literowe i przekłamania powstałe przy wczytywaniu taśm, ale także odtworzyć z dużym prawdopodobieństwem rozróżnienie dużych i małych liter.

Wartość korpusu bierze się przede wszystkim stąd, że — jak piszą autorzy słownika frekwencyjnego⁵ —

Homonimie morfologiczną i syntaktyczną usuwano, różnicując formy homonimiczne przez dopisywanie umownych symboli cyfrowych. Ze względu na ograniczoną pojemność pamięci maszyny liczba symboli gramatycznych została ograniczona do 63; ułożony kod jest kodem pozycyjnym.

Autorzy piszą jednak również

Zwracamy uwagę, że w opracowywanym materiale leksykalnym symbolizacja odpowiednich cech gramatycznych dotyczy tylko słowoform i haseł homonimicznych, nie obejmuje więc ona całości badanego słownictwa. Słowoformy i hasła nie kodowane oznaczają więc formy nie będące homonimami, których funkcje morfologiczno-syntaktyczne łatwo odczytać z samej postaci wyrazu (por. *domami*).

Nasunęło się zatem interesujące pytanie, czy *funkcje morfologiczno-syntaktyczne*, które człowiek może *łatwo odczytać z samej postaci wyrazu*, mogą być automatycznie dopisane za pomocą analizatora morfologicznego SAM. Częściowa odpowiedź była znana z góry: funkcje czysto morfologiczne tak, czysto syntaktyczne raczej nie. Podjęto zatem zadanie dopisania do wszystkich

⁵Wszystkie tomy pracy [19] zawierają tekst oryginalnej instrukcji redakcyjnej. Przytoczony cytat pochodzi z Instrukcji II, z punktu *Opis gramatyczny. I. kod fleksyjny*.

słów korpusu tych własności morfologicznych, które dają się rozpoznać automatycznie — otrzymany wynik nazywamy *wzbogaconym korpusem słownika frekwencyjnego*, w skrócie WKSF. Ponieważ postać hasłowa należy do własności morfologicznych, wzbogacony korpus pozwolił na stworzenie dla niego konkordancji hasłowanej — praca ta z inicjatywy prof. dr hab. Jadwigi Sambor była sfinansowana z funduszu badań statutowych Katedry Językoznawstwa Ogólnego i Bałtystyki Uniwersytetu Warszawskiego. Mamy nadzieję, że realizatorzy tego ciekawego eksperymentu — wśród których był Marcin Woliński i Maciej Ogrodniczuk — przedstawią jego szczegóły w osobnej publikacji. Tutaj tylko przedstawimy w charakterze przykładu drobny fragment tej konkordancji:

		ćma	
D1074	ich w słowa, podobne były do	ciem	SPGF-----P tłukących zapamiętałe w
D1077		Ćma	SSNF-----P przeleciała mi koło ucha, a
E0175	Dziwka. Dlaczego to zrobiłaś? .. Bo	ćma	SSNF-----P leci do światła .. Siedziałem
E1210	Dokładnie pani opowiada. Ja tę	ćmę	SSAF-----P to widzę, jak lata koło lampy.
E1225	.. Jurgacz, niech pan złapie tę	ćmę.	SSAF-----P
D1077	pochopnie powziąłem myśl zglądzenia	ćmy.	SSGF-----P
		ćmić	
E1965	jest taki. Jest! Ten, co go	ćmi,	VS---3TON---P albo cknij jak chce palić, a
E1992	hulaj, że aż się w oczach	ćmi.	VS---3TON---P A Walik to nie jeździ po
D0509	.. Ładny gips! Przed wrotami,	ćmiąc	V---W--N----P papierosa, czekał na niego
		ćwiartka	
D1598	będzie to nic kosztowało. Najwyżej	ćwiartkę ..	SSAF-----P dodał na wszelki wypadek.
D1598	kiwnął głową .. Tak. Postawicie nam	ćwiartkę,	SSAF-----P a my już damy facetowi radę.

3.3 Wspomaganie sporządzania skorowidzów

W środowisku akademickim często stosowanym narzędziem do składu tekstów jest bezpłatny system $\text{T}_{\text{E}}\text{X}$ [17] i jego odmiana $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X} 2_{\epsilon}$ (przy ich użyciu były przygotowane do druku m.in. publikacje [1, 20, 30]). Użytkownicy systemu często stosują również bezpłatny edytor tekstów Emacs, wyposażony w specjalne rozszerzenia — takie jak AUC TeX i RefTeX — ułatwiające współpracę z systemem $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X} 2_{\epsilon}$ [7]. Narzędzia te w szczególności ułatwiają sporządzanie skorowidzów, stanowiących istotny element obszerniejszych prac naukowych.

Jedną z możliwych metod sporządzania indeksu czy skorowidza składa się z m.in. następujących etapów:

1. sporządzenie listy wszystkich słów występujących w tekście;
2. wybranie spośród nich tych, które mają stanowić hasła w indeksie;
3. przejrzanie dla każdego hasła jego wszystkich wystąpień w tekście i odpowiednie oznaczenie tych, które mają być odnotowane w indeksie;

Edytor Emacs z rozszerzeniem RefTeX znacznie ułatwia wykonanie tych operacji, ale istotny problem stanowi fakt, że wyrazy w indeksie występują w swojej postaci hasłowej, zaś w tekście występują słowa stanowiące ich formy fleksyjne.

Problem ten został rozwiązany przez Kingę Izdebską [15, 16], która rozbudowała RefTeX o możliwości współpracy z analizatorem morfologicznym SAM. Po dokładnym przetestowaniu i usunięciu ewentualnych usterek, rozszerzenie to będzie dostępne bezpłatnie — podobnie jak sam edytor Emacs — na zasadach tzw. Licencji GNU Swobodnego Oprogramowania.

4 Inne zastosowania

Analizator SAM jest wykorzystywany również przez badaczy nie związanych organizacyjnie w żaden sposób z jego autorem. Znamy trzy takie zastosowania.

Pierwsze z nich to finansowany przez KBN projekt POLENG *Komputerowe tłumaczenie z języka polskiego na angielski tekstów informatycznych umieszczonych na stronach WWW* zrealizowany pod kierunkiem Krzysztofa Jassemę na Wydziale Matematyki i Informatyki Uniwersytetu Adama Mickiewicza w Poznaniu. Jednym z etapów było hasłowanie korpusu polskich tekstów informatycznych. Za pomocą analizatora SAM przetworzono ponad 1 milion słów, a wyniki omówiono w artykule [13].

Analizator SAM był również wykorzystywany w innym projekcie finansowanym przez KBN, mianowicie w zrealizowanym w Instytucie Informatyki Politechniki Śląskiej projekcie *Translacja tekstów w języku polskim na język migowy*. Jest o tym mowa w artykułach [11, 23].

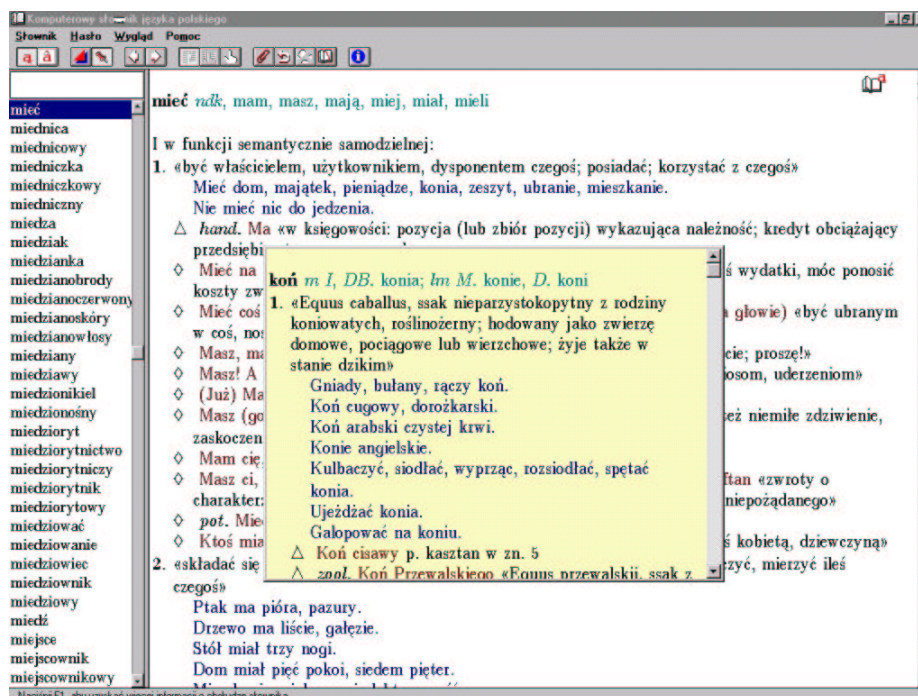
Jeszcze inne zastosowanie to wykorzystanie analizatora SAM do weryfikacji korektora ortograficznego MSPELL zintegrowanego ze wspomnianym wcześniej systemem składu $\text{T}_{\text{E}}\text{X}$. Praca ta, zrealizowana w Instytucie Matematyki Uniwersytetu Gdańskiego, jest omówiona w referacie [9].

Oczywiście, tak szerokie wykorzystanie analizatora SAM — i to niekiedy w sposób odległy od jego pierwotnego przeznaczenia — nie mogło odbywać się bezproblemowo. Od jego użytkowników napłynęło sporo uwag krytycznych, które zostaną uwzględnione w przyszłych naszych pracach.

5 Inne analizatory

Dla pełności obrazu należy wspomnieć, że analizator SAM nie jest jedynym praktycznym zastosowaniem indeksu Tokarskiego. Jest na nim oparty również moduł analizy morfologicznej komputerowego słownika języka polskiego [18], przygotowanego przez spółkę Litterae dla wydawnictwa PWN; bliższe informacje na ten temat można znaleźć m.in. w artykule [5]. Analizie morfologicznej może być poddawany również sam tekst definicji słownikowych, co pozwala odnaleźć łatwo objaśnienia występujących w nim słów; jest to szczególnie po-

żyteczne dla cudzoziemców uczących się języka polskiego — por. [31]. Ilustracja 3 pokazuje właśnie sytuację, gdy w trakcie przeglądania artykułu hasłowego dla MIEĆ użytkownik postanowił zapoznać się ze znaczeniem słowa *konia*; po ustaleniu postaci hasłowej tego słowa program wyświetlił artykuł hasłowy KOŃ.



Rysunek 3: Komputerowy Słownik Języka Polskiego PWN w działaniu

Inna zasługująca na uwagę komercyjna implementacja analizy morfologicznej była już wspomniana wcześniej. Robert Wołosz, absolwent Uniwersytetu Jagiellońskiego od lat związany z uniwersytetem w Peczu, nawiązał współpracę z węgierską firmą Morphologic, dla której przygotował dane lingwistyczne dotyczące języka polskiego. Są one wykorzystywane na dwa sposoby: w korektorze ortograficznym i we właściwym analizatorze morfologicznym. Programy te, o nazwie POMOR, są dostępne na rynku mniej więcej od 1995 roku, kiedy były prezentowane w kularach konferencji *Język i Technologia* w Poznaniu [32]. Choć od dłuższego czasu są one stosowane w praktyce (por. np. [36]), bliższe informacje o nich stały się szerzej dostępne dopiero po obronie pracy doktorskiej [37]. Oto zaczerpnięte z niej w sposób dość przypadkowy przykłady analizy (s. 141, 134, 138, 144):

```

PROCES: proces[Sm3]=PROCES+[11];
        proces[Sm3]=PROCES+[41]
INKRYMINOWANE: inkryminowany[Adj]=INKRYMINOWAN+e[05]=E;
                inkryminować[Vndk]=INKRYMIN+owane[b05]=OWANE
NIEZNANA: nie[NEG]=NIE+znać[Vndk]=ZNA+na[b06]=NA
TRZECI: trzeci[Adj]=TRZEC+i[01]=I;
        trzeci[Adj]=TRZEC+i[09]=I

```

Innymi analizatorami, opracowanymi w ramach międzynarodowego projektu badawczego, ale dostępnymi — o ile nam wiadomo — tylko na zasadach komercyjnych, są analizatory POLEX opisany w pracy [33] i LEXAN opisany w pracy [34]. W niewielkiej opublikowanej próbce udało nam się znaleźć tylko jedno słowo występujące również w próbce Wołosza. Oto przykładowe wyniki POLEXa ([33], s. 111, 114):

```
proces={proces(N310,1) proces(N310,4)}
poszarpane={poszarpany(ADJPAP,3) poszarpany(ADJPAP,11)
             poszarpany(ADJPAP,18) poszarpany(ADJPAP,20)
             poszarpany(ADJPAP,24) poszarpany(ADJPAP,28)}
nieznani={nieznani(ADJ1,19) nieznany(ADJ,27)}
```

W wyniku współpracy międzynarodowej powstał również na Uniwersytecie Warszawskim analizator POLLEX [8], nie dysponujemy jednak żadnym przykładem jego wyników.

Jeśli pominąć różnice notacyjne w sposobie prezentowania wyników, dostrzeżemy istotną różnicę w traktowaniu przymiotników. POMOR stosuje oznaczenia form wprowadzone przez Salonię, np. 05 oznacza *mianownik i biernik rodzaju nijakiego, mianownik i biernik rodzaju niemęskosobowego, deprecjatywny wariant mianownika rodzaju męskosobowego*, natomiast POLEX (a także, jak się wydaje, POLLEX) dla każdej funkcji składniowej wypisuje odrębny kod.

Można zatem powiedzieć, że analizator POMOR, podobnie jak SAM, jest analizatorem ściśle morfologicznym, podczas gdy pozostałe można nazwać analizatorami morfosyntaktycznymi.

Dla porównania, oto wyniki analizatora SAM⁶ dla słów występujących w powyższych przykładach:

```
proces %
{{(N) < proces(mIV::m3)+ } }%
inkryminowane %
{{(5) < inkryminowany(A::[p])+ } }%
{(A5) < inkryminować(IV::ndk)+ } }%
nieznana %
{{(6) < nieznany(A::[p])+ } }%
trzeci %
{{(1,9) < trzeci(A::+)+ } }%
poszarpane %
{{(A5) < poszarpać(IX::dk (się))+ } }%
nieznani %
```

⁶Użyto tu wersji analizatora SAM-99.

{(9) < nieznany(A::[p])+ } }%

Obszerne próbki wyników SAMa i innych analizatorów morfologicznych oraz ich dyskusję (dokonaną przy nieco kontrowersyjnych założeniach metodologicznych) można znaleźć w raporcie [14].

6 Podsumowanie

Przez wiele lat brak komputerowych analizatorów morfologicznych dla języka polskiego stanowił barierę nie tylko dla prac badawczych, ale i dla programów przetwarzających teksty polskie w celach czysto użytkowych. Nie ulega obecnie wątpliwości, że bariera ta została już przełamana — istnieje wiele analizatorów, opartych na różnych metodologiach i przeznaczonych do różnych celów (myśmy wymienili tutaj tylko najważniejsze z nich) i są one stosowane w praktyce. Trzeba jednak dobitnie podkreślić, że po 8 latach od powstania analizatora SAM jest on nadal jedynym analizatorem dostępnym bezpłatnie do celów niekomercyjnych (naukowych i dydaktycznych).

7 Summary

Morphological analysers of Polish in practice

We discuss first the notion of morphological analysis, stressing its dependence on various technical and conventional decisions. We consider lemmatization to be a special case of morphological analysis, spell-checking being an extreme case of it.

We focus next on the practical applications of a specific morphological analyser, named SAM, which has been developed by Krzysztof Szafran. SAM has been used in particular to provide data for a parser of Polish, to verify and supplement morphological tags in a corpus of Polish, and to extend the support of index creation for publications prepared with GNU Emacs and L^AT_EX 2_ε programs.

We mention some other morphological analysers of Polish and their availability; for the time being SAM is still the only one available free of charge for research and educational purposes.

Literatura

- [1] Bień, J.S. 1991. *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Rozprawy Uniwersytetu Warszawskiego t. 383. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

- [2] Bień, J.S. 1997. Komputerowa weryfikacja formalnej gramatyki Świ-
dzińskiego. *Biuletyn Polskiego Towarzystwa Językoznawczego* zeszyt LII,
s. 147–164.
- [3] Bień, J.S. 2000. Zestaw testów do weryfikacji i oceny analizatorów ję-
zyka polskiego. Sprawozdanie merytoryczne [nieznacznie zmodyfikowa-
ne] z projektu KBN 8 T11C 002 13. Instytut Informatyki UW. ftp:
//ftp.mimuw.edu.pl/pub/users/polszczyzna/tajp/.
- [4] Bień, J.S. 2001. O pojęciu wyrazu morfologicznego. W. Gruszczyński, W.,
Andrejewicz, U., Bańko, M., Kopcińska, D. (red.), *Nie bez znaczenia ...*
Białystok: Wydawnictwo Uniwersytetu w Białymstoku, s. 67–77 (ISBN
83-89031-01-9).
- [5] Bień, J.S., Linde-Usiekiewicz, J. 1998. Elektroniczne słowniki języka
polskiego. *Postscriptum* nr 24-25 (zima '97 — wiosna '98), s. 17–25.
- [6] Bień, J.S., Szafran, K., Woliński, M. 2000. Experimental parsers of Po-
lish. 3. *Europäische Konferenz "Formale Beschreibung slavischer Spra-
chen, Leipzig 1999"*. Linguistische Arbeitsberichte 75, Institut für Lingu-
istik, Universität Leipzig, 2000, pp 185–190. December 2000.
- [7] Bień, J.S. 2002. Gnu Emacs 21 i L^AT_EX 2_ε — piszemy artykuł naukowy. In
Proceedings of the XIII European T_EX Conference, April 29–May 3, 2002,
Bachotek, Poland, pp 105–111.
- [8] Bogacki, Ch. 1997. POLLEX — un dictionnaire électronique morphologi-
que du polonais. *BULAG* (Bulletin de Linguistique Appliquée et Généra-
le, Université de Franche Comté, ISSN 0758 6787) Numéro Spécial Actes
FRACTAL '97, pp 55–63.
- [9] Bzyl, Wł. 1999. Detection and correction of spelling errors in marked-up
documents, *Paperless T_EX. EuroT_EX 99 Proceedings*, pp 290–307.
- [10] Doroszewski, W. 1958-1969. *Słownik języka polskiego PAN* pod red.
W. Doroszewskiego. Wiedza Powszechna - PWN 1958-1969. Przedruko-
wany przez PWN w 1997 r., również dostępny na CD-ROM jako ISBN
83-01-12321-4.
- [11] Fabian, P., Migas, A., Suszczańska, N. 1999. Zastosowania analizy mor-
fologicznej i składniowej w procesie rozpoznawania mowy. *Technologia
mowy i języka* t. 3, Poznań, s.155–165.

- [12] Gladney, F.Y. 1994. Jan Tokarski Redivivus. *Journal of Slavic Linguistic* 2(2), pp. 304–317.
- [13] Graliński, F. 2000. Hasłowanie korpusu polskich tekstów informatycznych (1.2 mln słów), *Technologia mowy i języka* t. 4, Poznań, s.147–153.
- [14] Hajnicz, E., Kupść, A. 2001. Przegląd analizatorów morfologicznych dla języka polskiego. Raport Instytutu Podstaw Informatyki PAN Nr 937, Warszawa, grudzień 2001.
- [15] Izdebska, K. 2001. Tworzenie skorowidzów w systemie $\text{\LaTeX} 2_{\epsilon}$ dla dokumentów w języku polskim. Praca magisterska napisana pod kierunkiem dra Krzysztofa Szafrana. Warszawa: Instytut Informatyki Uniwersytetu Warszawskiego. 89 s., 2 dyskietki.
- [16] Izdebska, K. 2001. Wykorzystanie Gnu Emacsa i Reftex podczas tworzenia indeksów dla dokumentów $\text{\LaTeX} 2_{\epsilon}$. *Biuletyn Polskiej Grupy Użytkowników Systemu \TeX* , zeszyt 17, grudzień 2001 (ISSN 1230-5650), s. 45–50.
- [17] Knuth, D.E. 2001. *Computers & Typesetting, Volumes A–E (Revised edition)*, Addison-Wesley.
- [18] *Komputerowy słownik języka polskiego*, wydanie drugie. Warszawa: Wydawnictwo Naukowe PWN, 1998. ISBN 83-01-12504-7.
- [19] Kurcz, I., Lewicki, A., Masłowski, W., Sambor, J., Woronczak, J. 1974–1977. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Tom I-V, Uniwersytet Warszawski.
- [20] Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., Woronczak, J. 1990. *Słownik frekwencyjny polszczyzny współczesnej*, Kraków: Instytut Języka Polskiego PAN.
- [21] Prószyński, G. 1995. Humor (High-speed Unification MORphology). A Morphological System for Corpus Analysis. Heile Rettig (ed.). Proceedings of the First European Seminar LANGUAGE RESOURCES FOR LANGUAGE TECHNOLOGY. Tihany, Hungary, September 15 and 16, pp 149–158.
- [22] Nazarczuk, M. 1997. *Wstępne przygotowanie korpusu „Słownika frekwencyjnego polszczyzny współczesnej” do dystrybucji na CD-ROM*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa: Instytut Języka Polskiego Uniwersytetu Warszawskiego. 59 s., płyta CD.

- [23] Suszczańska, N., Forczek, M., Migas, A. 2000. Wieloetapowy analizator morfologiczny, *Technologia mowy i języka* t. 4, Poznań, s. 155–165.
- [24] Szafran, K. 1993. Automatyczna analiza fleksyjna tekstu polskiego (na podstawie *Schematycznego indeksu a tergo* Jana Tokarskiego). Niepublikowana praca doktorska, Wydział Polonistyki UW.
- [25] Szafran, K. 1996. Analizator morfologiczny SAM-95 — opis użytkowy. Raport Instytutu Informatyki Uniwersytetu Warszawskiego TR 96-05 (226). <ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95>
- [26] Szafran, K. 1997. Automatyczne hasłowanie tekstu polskiego. *Polonica*, tom XVIII. IJP PAN: Kraków 1997, s. 51–63.
- [27] Szafran, K. 2001. Kilka uwag o *Schematycznym indeksie a tergo polskich form wyrazowych* Jana Tokarskiego. Gruszczyński, W., Andrejewicz, U., Bańko, M., Kopcińska, D. (red.), *Nie bez znaczenia...* Białystok: Wydawnictwo Uniwersytetu w Białymstoku, s. 243–254. (ISBN 83-89031-01-9).
- [28] Świdziński, M. 1981. O spójnikach i partykułach odmiennych przez osobę. *Acta Universitatis Lodzianensis, Folia Linguistica* 2, Łódź, s. 283–284.
- [29] Świdziński, M. 1992. *Gramatyka formalna języka polskiego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.
- [30] Tokarski, J. 2001 *Schematyczny indeks a tergo polskich form wyrazowych*. Opracowanie i redakcja Zygmunt Saloni. Wydanie drugie. Warszawa: Wydawnictwo Naukowe PWN.
- [31] Tomczak, K. 1998. O możliwościach wykorzystania elektronicznych słowników języka polskiego w nauczaniu cudzoziemców. *Postscriptum* nr 24-25 (zima '97 — wiosna '98), s. 36–41.
- [32] Vetulani, Z., Abramowicz, W., Vetulani, G. [red.] 1996. *Język i technologia*. Warszawa: Akademicka Oficyna Wydawnicza PLJ.
- [33] Vetulani, Z., Martinek, J., Obrębski, T. 1998. Grażyna Vetulani. *Dictionary Based Methods and Tools for Language Engineering*. Poznań: Wydawnictwo Naukowe UAM.
- [34] Vetulani, Z. et al. 1998. *Unambiguous coding of the inflection of Polish nouns and its application in electronic dictionaries — format POLEX* [in Polish and English]. Poznań: Wydawnictwo Naukowe UAM.

- [35] Woliński, M. 2001. Rodzajów w polszczyźnie jest osiem. Gruszczyński, W., Andrejewicz, U., Bańko, M., Kopcińska, D. (red.), *Nie bez znaczenia ...* Białystok: Wydawnictwo Uniwersytetu w Białymstoku, s. 303–305. (ISBN 83-89031-01-9).
- [36] Wołosz, R. 1996. Komputerowa weryfikacja informacji o wyrazach z kwalifikatorem 'dawny' w SJPDor., *Slavica Quinqueecclesiensia* II, Pécs, pp 239–251.
- [37] Wołosz, R. 2000. Efektywna metoda analizy i syntezy morfologicznej w języku polskim. Niepublikowana rozprawa doktorska. Wydział Polonistyki, Uniwersytet Warszawski, Warszawa.

8 Uwagi do wersji elektronicznej

Niniejszy artykuł ukazał się — z licznymi błędami drukarskimi — w *Biuletynie Polskiego Towarzystwa Językoznawczego* zeszyt LVII (2001) na s. 171–184; planowane jest opublikowanie erraty w następnym zeszycie *Biuletynu*.

Za zgodą redaktora *Biuletynu* Prof. Kazimierza Polańskiego autoryzowana wersja tekstu jest udostępniona w Internecie.

Aktualnie (19.02.2003) pliki z artykułem formacie Postscript i PDF (odpowiednio JSB-KS-PTJ01.ps i JSB-KS-PTJ01.pdf) znajdują się pod adresami

http://www.orient.uw.edu.pl/~jsbien/publikacje/JSB-KS-PTJ01.*
http://www.mimuw.edu.pl/~jsbien/publikacje/JSB-KS-PTJ01.*

W przyszłości adresy te mogą ulec zmianie.

Ze względów technicznych wersja elektroniczna różni się od wersji drukowanej również pod względem formalnym, w szczególności uległ zmianie podział tekstu na strony.