

MARCIN WOLIŃSKI

## System znaczników morfosyntaktycznych w korpusie IPI PAN

Niniejszy artykuł opisuje zasady znakowania<sup>1</sup> morfosyntaktycznego tekstów języka polskiego przyjęte dla korpusu tekstów tworzonych w ramach projektu 7 T11C 043 20 finansowanego przez Komitet Badań Naukowych w latach 2001–2004 i realizowanego w Instytucie Podstaw Informatyki PAN pod kierunkiem Adama Przepiórkowskiego.

Omawiany system znaczników opracowali Adam Przepiórkowski i Marcin Woliński. Pracy tej sekundowali Łukasz Dębowski i Elżbieta Hajnicz. W końcowej fazie do dyskusji włączył się Zygmunt Saloni.

### 1. Wprowadzenie

Korpus tekstów języka naturalnego może być cennym źródłem danych w pracach leksykograficznych, w badaniach systemu fleksyjnego, badaniach struktury składniowej języka i innych związanych ze statystycznym modelowaniem różnych aspektów języka. Jedną z najprostszych form wykorzystania korpusu jest wyszukiwanie w nim wystąpień konkretnych słów<sup>2</sup> (ukazanych w kontekście). W praktyce jednak taki sposób pracy jest mało zadowalający. Do efektywnego wykorzystania korpusu konieczna jest co najmniej możliwość wyszukiwania w nim form wyrazowych należących do zadanych leksemów lub typów leksemów.

Dla zadania wyszukania w korpusie wszystkich form należących do określonego leksemu, można sobie wyobrazić dwa rozwiązania. Pierwsze polegałoby na (automatycznym) znalezieniu dla zadanego leksemu wszystkich słów, które mogą być wykładnikami form doń należących i szukaniu w korpusie wystąpień tych słów. Drugi sposób postępowania to oznakować cały tekst korpusu, przypisując każdemu słowu jego interpretację jako wykładnika pewnej formy pewnego leksemu (dokonać analizy fleksyjnej tekstu). Wówczas wyszukiwanie może się odbywać według tego znakowania.

Pierwszy sposób postępowania ma oczywistą wadę ze względu na niejednoznaczności interpretacji (homonimie): pewne słowa mogą być wykładnikami więcej niż jednej formy

<sup>1</sup> Posługuję się terminem *znacznik* jako odpowiednikiem angielskiego *tag* oraz *znakować* na oznaczenie czynności opatrywania znacznikami. Artykuł jest tym samym eksperymentem, mającym na celu sprawdzenie, czy da się w tej materii unikać używania kalek z angielskiego. Pozostawiam Czytelnikowi ocenę wyników.

<sup>2</sup> Pojęciami słowo, forma, leksem, kategoria gramatyczna staram się posługiwać zgodnie z pracą Saloniego i Świdzińskiego (2001).

lub więcej niż jednego leksemu. W takich wypadkach otrzymuje się odpowiedź na inne pytanie, niż zostało zadane. Uzyskuje się mianowicie wszystkie słowa, które mogłyby być wykładnikami form danego leksemu (gdyby zaniedbać kontekst ich wystąpienia). Drugi sposób pozwala przy każdym wystąpieniu słowa zaznaczyć właściwą dla tego wystąpienia interpretację i w konsekwencji umożliwia udzielenie właściwej odpowiedzi.

Wydaje się zatem, że aby korpus dawał się efektywnie wykorzystać, musi on zostać oznakowany, a niejednoznaczności dotyczące słów homonimicznych muszą zostać rozstrzygnięte (ręcznie lub automatycznie).

Z punktu widzenia przetwarzania danych są tu do wykonania dwa zadania. Pierwsze polega na określeniu dla każdego słowa wszystkich form wszystkich leksemów, których może ono być wykładnikiem. Na tym etapie nie uwzględnia się kontekstu, w którym wystąpiło dane słowo. Ten proces będę nazywać *analizą fleksyjną* (lub *morfologiczną*). Drugie zadanie, czy też etap, polega na określeniu na podstawie kontekstu, jaką formę faktycznie reprezentuje dane wystąpienie słowa. Ten proces można nazwać *ujednoznacznieniem fleksyjnym*. Do automatycznego ujednoznaczniania wyników analizy fleksyjnej zwykle używa się metod statystycznych.

W niektórych pracach bardziej użyteczne są wyniki pierwszego z tych etapów, niż obu. Analizę fleksyjną można stosunkowo łatwo wykonać automatycznie z dużą dokładnością. Jest to jedynie kwestia zgromadzenia słownika fleksyjnego (rozumianego jako program komputerowy), który będzie odpowiednio obszerny i będzie opisywał formy odpowiednio precyzyjnie. Natomiast ujednoznacznianie wystąpień wymaga odwoływania się do zjawisk składniowych i semantycznych. Czasem zaś nie da się go wykonać. W następującym przykładzie nawet odwołując się do wypowiedzeń otaczających podane nie da się stwierdzić, czy drugie słowo jest rzeczownikiem, czy czasownikiem (wtedy pierwsze jest wykładnikiem formy rzeczownika połączonej z aglutynacyjną formą *-em*), treść bowiem jest w obu interpretacjach ta sama.<sup>3</sup>

(1) *Miałem miał.*

Można się spodziewać, że ujednoznacznienie dokonane za pomocą metod statystycznych będzie niedokładne, przy czym największa jest szansa na błąd w wypadku form rzadkich. Ale przecież właśnie te formy często interesują leksykografa. Dlatego w niektórych pracach może być konieczne przejrzanie większej liczby słów, które potencjalnie mogą być wystąpieniami szukanej formy.

Z kolei w automatycznej analizie składniowej wielość interpretacji morfologicznych nie musi przeszkadzać — to właśnie wykonanie analizy składniowej wypowiedzenia może dać ujednoznacznienie na poziomie fleksyjnym. Jeśli jednak na podstawie statystycznej przedwcześnie usunie się pewne interpretacje morfologiczne, uzyskanie rozbioru składniowego może być niemożliwe.

Sądzę więc, że w znakowanym korpusie tekstowym konieczna jest możliwość dostępu zarówno do wszystkich interpretacji słów, jak i do wyniku ich ujednoznacznienia.

Opracowanie obszernego korpusu tekstów wiąże się ze znacznym nakładem pracy. W związku z tym byłoby dobrze, gdyby raz stworzony korpus mógł służyć użytkownikom o różnych zainteresowaniach i potrzebach. Pierwszym krokiem ku stworzeniu takiego korpusu jest wypracowanie systemu znakowania tekstu, który byłby akceptowalny z punktu

<sup>3</sup> Jest jeszcze trzecia interpretacja, eliptyczna, odpowiadająca na pytanie *Czym miał zasypany trawnik?*

widzenia lingwistycznego, a jednocześnie pozwalał na efektywne przetwarzanie komputerowe (w szczególności efektywne automatyczne znakowanie).

Zwróćmy przy tym uwagę, że ze względu na kwestie prawne (prawa autorskie do tekstów) twórcy korpusów często nie mogą udostępnić całości tekstów, lecz jedynie wyniki ich przetwarzania (listy słów, leksemów, konkordancje). Ma to istotne konsekwencje dla użytkownika korpusu. Mianowicie jeśli idzie o analizę morfologiczną tekstu, a zwłaszcza o ujednoznaczenie homonimów, użytkownik jest zdany na pracę wykonaną przez twórców korpusu i nie jest w stanie samodzielnie poprawić błędów, zmienić zasad hasłowania ani oznaczania form.

Dlatego też w projekcie przyjęto sposób znakowania, który jest w pewnym sensie minimalny. Celem jest mianowicie jedynie dyzambiguacja fleksyjna tekstu — chodzi o to, aby jednoznacznie wskazać pewną formę w obrębie pewnego leksemu, której badane słowo jest tekstowym wykładnikiem. Notowane są więc wyłącznie cechy pozwalające odróżnić od siebie leksemy lub formy w obrębie leksemów.

Takie podejście ma dwojake uzasadnienie. Pierwszym argumentem jest nadzieja, że w ten sposób zminimalizowany zostaje wpływ przyjętych rozwiązań na system pojęciowy stosowany przez użytkowników korpusu. Skoro informacja w korpusie służy tylko do identyfikacji form, jeżeli zestawia się ją z odpowiednim słownikiem notującym dodatkowe cechy leksemów (np. wzory odmiany według jakiejś klasyfikacji), można zapewnić wyszukiwanie w korpusie według tych dodatkowych informacji. (Wymaga to bardziej zaawansowanego wykorzystania mechanizmu wyszukiwawczego korpusu, niestety może więc powodować problemy przy opisanym wyżej ograniczonym dostępie do korpusu).

Drugim argumentem jest efektywność komputerowego przetwarzania tekstów. Podstawą osiągnięcia efektywności jest niemnożenie bytów. Jeżeli na pewnym etapie przetwarzania pewne różne obiekty zachowują się identycznie, należy zadbać o to, aby były opisane wspólnie jako jeden byt. Pamiętajmy, że dla automatycznego analizatora fleksyjnego informacje o znaczeniu leksemów są niedostępne. Dlatego też na tym etapie leksemy należy traktować jako zbiory form rozumianych jako pary złożone z identyfikatora leksemu i określenia pozycji formy w paradygmacie tego leksemu. Konsekwencją takiej decyzji jest brak ujednoznaczniania semantycznego różnych znaczeń leksemów.

## 2. Segmentacja tekstu

Dla tekstu pisanego w języku polskim nie jest zupełnie oczywiste, jakie segmenty powinny być interpretowane jako wykładniki form leksemów.

W pierwszym przybliżeniu interpretacji podlegają słowa rozumiane jako sekwencje znaków nie zawierające odstępów ani znaków niealfanumerycznych (z wyjątkiem, co najmniej, kropki stanowiącej część skrótu, apostrofu pojawiającego się w słowach takich jak *Chomsky'ego*, dywizu używanego przy odmianie skrótowców). W pewnych sytuacjach interpretacji powinny jednak podlegać zarówno segmenty dłuższe, jak i krótsze.

Tradycyjnie jako fleksyjne traktuje się na przykład tzw. formy analityczne czasu przyszłego i trybu warunkowego. Za segmenty przekraczające granice słów uważane są też wyrażenia typu *po polsku*. Jednak dopuszczenie takich segmentów raczej zwiększa liczbę problemów, niż zmniejsza. Jeżeli bowiem wziąć pod uwagę wypowiedzenia

- (2) *Będzie pisał lub czytał.*

(3) *Mówił po polsku i angielsku.*

widać, że i tak konieczna jest możliwość interpretowania fragmentów takich „segmentów” (albo uznania wypowiedzenia (2) za złożone z jednej formy fleksyjnej i kropki).

Dlatego na potrzeby projektu przyjęto kategorię zasadę nieznakowania żadnych segmentów zawierających odstęp. Oznacza to konieczność „zagospodarowania” form typu *polsku* (ta została włączona do paradygmatu leksemu *POLSKI*) lub stworzenia jednostek słownika (np. leksem *INDZIEJ* dla *gdzie indziej*).

W tym miejscu warto nadmienić, że takie rozstrzygnięcie nie oznacza uznania wszelkich zależności między jednostkami oddzielonymi odstępem za składniowe. Sądzę raczej, że w komputerowym przetwarzaniu tekstu można wyróżnić wiele etapów pośrednich pomiędzy tym, co fleksyjne, i tym, co tradycyjnie uważa się za składniowe. Zjawisko interpretowania segmentów złożonych z wielu słów sytuuje się gdzieś w tym „pomiędzy”. Zapewne konieczne jest opisanie wielu zjawisk uwarunkowanych leksykalnie, zanim przystąpi się do opracowania poziomu składniowego. Co więcej, do opisu takiego posłużą zapewne jeszcze inne formalizmy niż do opisu fleksji i (wąsko rozumianej) składni.

Niekiedy jednak interpretacji podlegają segmenty krótsze niż słowa. Typowym przykładem jest wypowiedzenie następujące:

(4) *Świniam.*

W wypowiedzeniu tym występują wykładniki dwóch form wyrazowych: formy leksemu rzeczownikowego *ŚWINIA* i skróconej (aglutynacyjnej) formy czasu teraźniejszego leksemu czasownikowego *BYĆ* (Saloni i Świdziński, 2001; Tokarski, 2001). Trzeba więc uznać, że interpretowane jako wykładniki tych form są odpowiednio segmenty *Świnia* i *m* stanowiące części właściwe słowa *Świniam*. Saloni i Świdziński w swoim podręczniku przeslizgują się nad tym problemem, pisząc nieprecyzyjnie „słowa typu *świniam* składają się z dwóch form wyrazowych” (s. 96; zdanie to stoi w niejkiej sprzeczności z poglądem o unilateralności słów<sup>4</sup>).

Aglutynacyjne formy czasownika *BYĆ* pojawiają się również w nieciągłym wariacie czasu przeszłego czasowników:

(5) *Dlaczegoś to zrobił?*

Słowo *zrobił* w tym wypowiedzeniu jest wykładnikiem formy, którą Saloni (2001) nazywa *pseudoimiesłowem*. Dla konsekwencji w niniejszym opisie uznaje się, że czas przeszły czasowników zawsze jest analityczny i również w drugim wariacie szyku mamy do czynienia z pseudoimiesłowem i aglutynantem. Podobnie tryb warunkowy wyraża się konstrukcją złożoną z pseudoimiesłowu, partykuły warunkowej *BY* i aglutynantu. Pewnym argumentem za takim spojrzeniem na sprawę jest występowanie słów takich jak w poniższym przykładzie:

(6) *Potrzebowałżebyś, pytam na koniec, tego strachu wstrętnego i bezsilnej wściekłości?*  
S. Lem *Bajki robotów*

W przedstawionej tu interpretacji słowo *potrzebowałżebyś* składa się z czterech segmentów, jednym z nich jest wykładnik formy leksemu wzmacniającego *ŻE*.

Biorąc pod uwagę, że w trybie warunkowym aglutynant musi następować bezpośrednio

<sup>4</sup> Problem ten podniósł w dyskusji Mirosław Bańko, za co mu niniejszym dziękuję.

po częście *by*, być może byłoby bardziej naturalnie wprowadzić „aglutynant warunkowy” o wykładnikach *bym, byś, by, byśmy, byście*.

Rozbijane są słowa pisane z łącznikiem, np. *polsko-niemiecki* (łącznik stanowi osobny segment). Jako osobny segment jest traktowana aglutynacyjna postać *-ń* zaimka osobowego *on*, segmenty *-że, -ż* i *-ć* stanowiące wykładniki leksemów wzmacniających *że* i *ć* oraz segment *-li* stanowiący wykładnik leksemu pytającego *li*.

Znaki interpunkcyjne wypada uznać za osobne segmenty, skoro są one istotne składniowo (Świdziński, 1992). Z technicznego punktu widzenia segmenty te są nawet interpretowane jako wykładniki „leksemów” o nazwach równokształtnych z analizowanym segmentem.

### 3. Jak hasłować tekst w języku polskim

Celem hasłowania jest wskazanie dla każdego segmentu opisującej go jednostki słownika fleksyjnego (leksemu w rozumieniu jak wyżej). Oczywiście konieczne jest podanie jakiegoś umownego identyfikatora leksemu (postaci hasłowej). Utrwaloną konwencją jest tutaj podawanie słów realizujących pewną wybraną formę danego leksemu (np. mianownik liczby pojedynczej dla leksemów rzeczownikowych). Zdarza się jednak, że takie postępowanie daje ten sam identyfikator dla różnych leksemów:

- (7) (a) *Jak śmiesz<sub>ŚMIEĆ</sub> tak się zwracać do matki?*  
 (b) *Weź ze sobą od razu ten kubek ze śmieciami<sub>ŚMIEĆ</sub>.*

Można temu zaradzić, numerując leksemy o homonimicznych postaciach hasłowych. Numery homonimów z konieczności byłyby jednak nadawane arbitralnie. Takie rozwiązanie wydaje się niepraktyczne, lepiej byłoby identyfikować leksemy na podstawie ich cech.

Poręcznym elementem identyfikacji leksemów jest podawanie typu leksemu, czyli „części mowy”. W powyższych przykładach wystarczy to do jednoznacznego wskazania leksemu. Zdarza się jednak, że to za mało:

- (8) (a) *Aresztując<sub>ARESztOWAĆ</sub> przestępcę, wykazał się odwagą.*  
 (b) *Aresztowawszy<sub>ARESztOWAĆ</sub> przestępcę, zajął się algebrą abstrakcyjną.*  
 (9) (a) *Gdy wszyscy pływacy<sub>PLYWAK</sub> są nieruchomi, starter podaje sygnał startu.*  
 (b) *W kałuży uwijały się pływaki<sub>PLYWAK</sub> żółto-brzeżki.*  
 (c) *Pole magnetyczne generowane przez pływak<sub>PLYWAK</sub> powoduje lokalną zmianę impedancji akustycznej struny.*

Przyjmujemy, że w przykładzie (8) mamy do czynienia z dwoma czasownikami o homonimicznej formie hasłowej i przeciwnym aspekcie. Czasowniki te różnią się zasobem form, więc zmuszeni jesteśmy uznać je za osobne leksemy w sensie naszej definicji. Dlatego przy formach czasownikowych zawsze będziemy notować aspekt.

Podobnie tak zwane rzeczowniki dwurodzajowe (czy wielorodzajowe), jak *PLYWAK* w przykładzie (9), stanowią osobne leksemy rzeczownikowe o homonimicznej postaci mianownika liczby pojedynczej, ale różnych wartościach rodzaju. Leksemy występujące w przykładzie mają wartości rodzaju odpowiednio: męską osobową, męską zwierzęcą i męską rzeczową.

Analogicznie dla przyimków notowany jest wymagany przypadek, ponieważ przyimki o homonimicznej postaci hasłowej, ale wymagające różnych przypadków uznajemy za osobne leksemy.

To jednak nie wyczerpuje problemu:

- (10) (a) *Ślemy<sub>SLAĆ</sub> doń list za listem już od dwóch miesięcy.*  
 (b) *Codziennie ścielemy<sub>SLAĆ</sub> łóżka.*

W tym wypadku oba czasowniki są niedokonane i mają homonimiczną formę hasłową, różnią się jednak zbiorami form. Można by je rozróżnić podając oznaczenia grup odmiany. To jednak wymagałoby przyjęcia jakiegoś systemu wzorów odmiany, co wydaje się niekorzystne.

Janusz Bień proponował na potrzeby systemu MARYSIA (Bień *et al.*, 1973), aby w takich wypadkach uzupełniać identyfikator o fragmenty formy lub form różniących leksemy w liczbie wystarczającej do jednoznacznego wskazania, o który leksem chodzi. Praktyczne zastosowanie tego schematu wymagałoby wypracowania reguł określających wybór form wchodzących do identyfikatora.

Taka procedura prawdopodobnie dałaby bardziej czytelne identyfikatory, ale skomplikowałaby sposób ich generowania. Na potrzeby projektu postanowiono więc dołączać do identyfikatora leksemu liczby numerujące kolejne homonimy. Oczywiście przypisanie takich liczb będzie arbitralne i będzie miało wady wspomniane wyżej.

W tytule niniejszego artykułu mowa o znakowaniu morfosyntaktycznym. Nazwa ta bierze się z tego, że oprócz cech morfologicznych (fleksyjnych) potrzebnych do identyfikacji form, w znacznikach notowane są wymienione wyżej cechy leksemów, które nie są dla nich fleksyjne, a których znaczenie uwidacznia się na poziomie syntaktycznym (składniowym). (Porównaj także pojęcie wyrazu morfosyntaktycznego w pracy Bienia (1991)).

#### 4. Identyfikacja form w obrębie paradygmatu

Tradycyjnie wyróżniane leksemy są jednostkami dość niejednorodnymi. Na przykład czasowniki to leksemy o fleksyjnej kategorii osoby. Jednak w obrębie leksemu czasownikowego znajduje się bezokolicznik, który jest obojętny ze względu na kategorię osoby. Znaczy to, że jednorodność odmiany form nie jest podstawą grupowania ich w leksemy, że raczej istotne są funkcje, jakie formy mogą pełnić w wypowiedzeniu.

W przedstawionej tu koncepcji proponuje się konstrukcję leksemów na podstawie faktycznej regularności odmiany, a dopiero na drugim miejscu na podstawie własności składniowych. Do tego celu użyteczna jest pośrednia jednostka klasyfikacji, grupująca formy, które są nieobojętne ze względu na te same kategorie fleksyjne. Jednostkę taką zaproponował Bień (1991) nazywając ją *fleksemem*. Bień przyjmuje, że fleksem jest zbiorem wyrazów morfologicznych różniących się tylko wartościami kategorii liczby, przypadku, rodzaju i osoby. W niniejszym opisie, aby zmniejszyć liczbę różnych typów fleksemów, dopuszcza się odmiennosc w obrębie fleksemów również ze względu na inne kategorie. Cechą wyróżniającą fleksemów pozostaje jednolite (lub niemal jednolite) zróżnicowanie ze względu na właściwe im kategorie gramatyczne.

Tak więc leksemy traktowane są jako zbiory fleksemów, które z kolei są zbiorami form wyrazowych. W obrębie leksemu czasownikowego wyróżnia się kilka fleksemów

odmiennych (np. flexsem zawierający syntetyczne formy trybu rozkazującego odmienny m.in. przez osobę) i kilka flexsemów nieodmiennych (np. flexsem bezokolicznikowy).

Aby poklasyfikować leksemy i flexsemy nieodmienne trzeba odwołać się do ich własności składniowych (podobnie jak w koncepcji Salonięgo).

Klasy flexsemów otrzymują indywidualne nazwy, aby więc określić, do leksemu jakiej klasy należy dana forma, wystarczy podać nazwę flexsemu, do którego ona należy. Identyfikacja formy w obrębie flexsemu odbywa się poprzez podanie zestawu wartości przysługujących jej kategorii fleksyjnych.

Z technicznego punktu widzenia w wyniku znakowania danemu segmentowi tekstu (kształtowi ortograficznemu) przypisywany jest pewien inny kształt ortograficzny, być może rozszerzony numerem homonimu i stanowiący identyfikator leksemu oraz pewien symbol z ustalonego repertuaru, który nazwiemy znacznikiem morfosyntaktycznym. Znacznik określa klasę flexsemów i pewien zestaw wartości kategorii gramatycznych.

## 5. Kategorie gramatyczne

Jeżeli przyjąć, że uzgodnienie kategorii gramatycznych polega na równości ich wartości dla określonych form w wypowiedzeniu, to trzeba uznać, że repertuar wartości dopuszczalnych dla danej kategorii jest taki sam niezależnie od tego, do jakiej klasy gramatycznej należy rozpatrywana w danym momencie forma. Przyjęcie, że np. rzeczowniki i przymiotniki mają różne zbiory wartości kategorii rodzaju, mogłoby pozwolić na zmniejszenie liczby różnych wartości. Wtedy trzeba by jednak w nietradycyjny sposób opisać uzgodnienie, wprowadzając nietrywialny operator uzgodnienia, zdający sprawę z tego, jakie wartości do siebie „pasują”. W niniejszej pracy stosowane jest rozwiązanie tradycyjne.

Oto lista kategorii gramatycznych, które wydają się istotne dla opisu form leksemów dla języka polskiego. Dla każdej kategorii podany został zestaw wartości, które może ona przyjmować, wraz z oznaczeniami skrótowymi stosowanymi w znacznikach morfosyntaktycznych.

- **Liczba:** pojedyncza sg, mnoga pl
- **Przypadek:** mianownik nom, dopełniacz gen, celownik dat, biernik acc, narzędnik inst, miejscownik loc, wołacz voc
- **Rodzaj:** męski osobowy m1, męski zwierzęcy m2, męski rzeczowy m3, żeński f, nijaki zbiorowy n1, nijaki zwykły n2, przymnogi osobowy p1, przymnogi zwykły p2, przymnogi opisowy p3
- **Osoba:** pierwsza pri, druga sec, trzecia ter
- **Stopień:** równy pos, wyższy comp, najwyższy sup
- **Aspekt:** niedokonany imperf, dokonany perf
- **Zanegowanie:** niezanegowana aff, zanegowana neg
- **Akcentowość:** akcentowana akc, nieakcentowana nakc
- **Poprzyimkowość:** poprzyimkowa praep, niepoprzyimkowa npraep
- **Akomodacyjność:** uzgadniająca congr, rządząca rec
- **Aglutynacyjność:** aglutynacyjna agl, nieaglutynacyjna nagl
- **Wokaliczność:** wokaliczna wok, niewokaliczna nwok

Wartości kategorii **liczby** dla języka polskiego wydają się nie budzić wątpliwości. Zaznaczmy tu tylko, że liczbę traktujemy oczywiście składniowo, więc np. wszystkie formy

rzeczowników *plurale tantum* traktujemy jako mnogie, mimo że niektóre morfologicznie przypominają formy liczby pojedynczej.

Zestaw kategorii **rodzaju** przejmujemy z pracy Saloniego (1976). Zaproponowany tam opis wychodzi od zróżnicowania form biernika liczby pojedynczej i mnogiej przymiotników, co pozwala wyróżnić trzy odrębne rodzaje męskie, rodzaj żeński i nijaki. Następnie rodzaj nijaki zostaje rozbity na dwa ze względu na łączliwość z formami liczebnikowymi typu *dwa* (n2) lub *dwoje* (n1, tzw. liczebniki zbiorowe, przez Saloniego traktowane jako formy liczebników tzw. głównych różniące się właśnie rodzajem). Wreszcie zostają wyróżnione w osobne klasy rodzajowe (przymnogie) rzeczowniki *plurale tantum*. Klasa p1 obejmuje rzeczowniki wykazujące łączliwość z formami typu *ci*, p2 te rzeczowniki, które łączą się z formami *te* i *dwoje*, p3 wreszcie są łączliwe z *te* ale nie *dwoje*.

Klasyfikacja Saloniego jest chyba najbardziej szczegółową spośród powszechnie używanych. Zastosowanie jej do znakowania korpusu pozwala mieć nadzieję, że inne sposoby klasyfikacji dadzą się uzyskać poprzez proste łączenie klas tu wyróżnionych. Na przykład, jeżeli ktoś nie jest zainteresowany rozróżnieniem opartym na łączliwości z formami liczebnikowymi, może operować rodzajem nijakim stanowiącym sumę klas n1 i n2.

**Stopień** stanowi kategorię fleksyjną przysługującą przymiotnikom i przysłówkom. Również w tym wypadku mamy nadzieję, że nie utrudniamy przesadnie posługiwania się odmiennym aparatem, w którym grupy form przymiotnikowych poszczególnych stopni tworzą osobne leksemy.

Kategoria **aspektu** jest czysto słownikowa, wszystkim formom danego czasownika przysługuje jedna jej wartość. Jak wspomniano wyżej, kategoria ta została wprowadzona dla ujednoznaczenia hasłowania czasowników dwuaspektowych. Ponadto, jak się okaże dalej, jawne znakowanie aspektu pozwala uprościć notowanie własności czasowych form czasownikowych.

Kategoria fleksyjna **zanegowania** przysługuje w przedstawionym tu opisie odśłownikom i imiesłowom przymiotnikowym czynnym i biernym. Uznajemy mianowicie, że np. formy *celebrowanie* i *niecelebrowanie* należą do tego samego leksemu. Kategoria zanegowania została wprowadzona dla odróżnienia od siebie tych form. W wypadku tych trzech klas gramatycznych obecność cząstki *nie-* ma konsekwencje składniowe. Kategoria zanegowania nie dotyczy przymiotników, np. *nieładny*, w ich wypadku bowiem nie ma podobnych konsekwencji (a ŁADNY i NIEŁADNY stanowią osobne leksemy).

Kategoria **akcentowości** została wprowadzona dla odróżnienia form zaimka osobowego występujących w zdaniu na pozycji akcentowanej (np. *jego*) od form występujących na pozycji nieakcentowanej (np. *go*). **Poprzyimkowość** różnicuje formy zaimka ON które mogą wystąpić wyłącznie po przyimku (*przez niego, do-ń*) od występujących w innych kontekstach (*jego, go*). (Kategorie te po raz pierwszy zostały wprowadzone przez Saloniego (1981)).

Kategoria **akomodacyjności**, zaproponowana w artykule Bienia i Saloniego (1982), przysługuje wyłącznie formom liczebnikowym. Jej znaczenie opisano w punkcie 6.4.

**Aglutynacyjność** to kategoria odróżniająca formy niesamodzielne obligatoryjnie dołączające jakąś formę aglutynacyjną od form niedopuszczających aglutynantu. Zróżnicowanie to ujawnia się dla pseudoimiesłowów tylko niektórych czasowników (np. GNIEŚĆ: *gniótt*, ale *gniott-em*) i tylko dla nich jest notowana aglutynacyjność.

Kategorię **wokaliczności** wprowadzono w celu odróżnienia wariantu aglutynacyjnych form czasownika BYĆ, który występuje po formie kończącej się spółgłoską (np. *-em*), od wariantu posamogłoskowego (*-m*).

Powyższa lista kategorii nie uwzględnia czasu, trybu ani strony czasowników. Jak się okaże dalej, czas jest wyrażany pośrednio poprzez pogrupowanie form czasownikowych we fleksy. W podobny sposób traktowany jest tryb rozkazujący. Tryb warunkowy jest traktowany jako konstrukcja, podobnie jak strona bierna (por. p. 6.6).

## 6. Klasy leksemów (części mowy)

Klasy leksemów wyróżnione w przedstawianym tu opisie wymieniono w tabeli 1. Dla każdej klasy podano, z jakich fleksmów składają się leksemy tej klasy. W tabeli 2 dla każdej klasy fleksmów podano listę kategorii gramatycznych przysługujących formom należącym do fleksmów tej klasy.

W tabeli 2 wyraźnie wyróżnia się grupa fleksmów, które można by nazwać deklinacyjnymi — mają one fleksyjną kategorię przypadku, oraz grupa fleksmów koniugacyjnych — którym przysługuje określona wartość aspektu. Trzy klasy, odsłownik, imiesłów przymiotnikowy czynny i bierny, wydają się przynależeć do obu tych grup. Wykazują one deklinacyjne cechy odmiany, a jednocześnie wyraźnie widać ich czasownikowe pochodzenie. Przysługuje im bowiem kategoria aspektu, co widać w poniższych przykładach:

- (11) *Rozpoczęcie spawania konstrukcji nastąpiło w marcu.*
- (12) *\*Rozpoczęcie zespawania konstrukcji nastąpiło w marcu.*

W niniejszym opisie klasy te zostały włączone do leksemu czasownikowego.

### 6.1. Rzeczowniki

Rzeczowniki są odmienne przez liczbę i przypadek i mają ustaloną wartość kategorii rodzaju. Identyfikatorem leksemu rzeczownikowego jest wykładnik formy mianownika liczby pojedynczej.

Ten prosty obraz zakłócają rzeczowniki rodzaju m1, dla których istnieją dwie formy mianownika i wołacza liczby mnogiej:

- (13) (a) *Przyszli uroczy profesorowie.*
- (b) *Przyszły głupie profesory.*

Według Saloniego (1988) takie pary form istnieją dla każdego rzeczownika rodzaju m1, choć czasami dochodzi do neutralizacji, a czasami któraś z form jest tylko potencjalna.

Dla opisanego zjawiska Bień i Saloni (1982) zaproponowali wprowadzenie kategorii deprecjatywności. Różnicuje ona formy niedeprecjatywne (*profesorowie*) i deprecjatywne (*profesory*). Taki opis, konstruowany w duchu dystrybucyjnym, musi rozciągać kategorię deprecjatywności również na inne klasy gramatyczne (co najmniej przymiotniki, liczebniki i czasowniki), aby w szczególności wykluczyć wypowiedzenia typu:

- (14) *\*Przyszli głupi profesory.*

Formy deprecjatywne rzeczowników m1 z punktu widzenia składni zachowują się bowiem jak formy rodzaju m2.

Innym rozwiązaniem mogłoby więc być wyniesienie form deprecjatywnych do osobnych leksemów (defektywnych), którym przypisany byłby rodzaj m2. Taki opis byłby jednak w sprzeczności z seryjnością występowania form deprecjatywnych dla rzeczowników m1.

Tabela 1. Klasy leksemów i ich rozbięcie na fleksy. W wypadku leksemów złożonych z tylko jednego fleksemu wspólna nazwa leksemu i fleksemu zajmuje dwie kolumny tabeli.

leksem	fleksy	ozn.
rzeczownik	rzeczownik	subst
	forma deprecjatywna	depr
przymiotnik	przymiotnik	adj
	przymiotnik przyprzymiotnikowy	adja
	przymiotnik poprzyimkowy	adjp
przysłówek odprzymiotnikowy i/lub stopniowalny		adv
liczebnik		num
zaimek nietrzeciosobowy		ppron12
zaimek trzeciosobowy		ppron3
zaimek SIEBIE		siebie
czasownik	forma nieprzeszła	fin
	forma przyszła czasownika być	bedzie
	aglutynant czasownika być	aglt
	pseudoimiesłów	praet
	rozkaznik	impt
	bezosobnik	imps
	bezokolicznik	inf
	imiesłów przys. współczesny	pcon
	imiesłów przys. uprzedni	pant
	odsłownik	ger
	imiesłów przym. czynny	pact
imiesłów przym. bierny	ppas	
czasownik typu WINIEN (forma terażniejsza)		winien
predykatyw		pred
przyimek		prep
spójnik		conj
kublik (partykuło-przysłówek)		qub
ciało obce nominalne		xxs
ciało obce luźne		xxx

Bień (1991) proponuje wprowadzenie w obrębie leksemu rzeczownikowego dwóch fleksów różniących się rodzajem. Oznacza to dopuszczenie namiastki odmienności przez rodzaj w obrębie leksemu rzeczownikowego, na które z pewnego punktu widzenia można przystać, biorąc pod uwagę, że dla fleksów rodzaj jest ustalony.

To ostatnie rozstrzygnięcie zostało przyjęte w projekcie. Mianowicie w prezentowanym opisie leksem rzeczownikowy zawiera fleks **rzeczownikowy** oraz fleks **deprecjatywny** (dla rzeczowników m1).

Źródłem pewnych problemów są również rzeczowniki *plurale tantum*. Gdyby traktować jednorodność fleksyjną jako ściśle kryterium przynależności do klas fleksów, powinno się utworzyć co najmniej 2 klasy fleksyjne dla rzeczowników: jedną, zawierającą fleksy odmienne przez liczbę, i drugą, zawierającą fleksy o wartości liczby ustalonej słowni-

Tabela 2. Kategorie przysługujące poszczególnym klasom fleksemów. Symbol ⊕ oznacza, że dla danej klasy dana kategoria jest morfologiczna. Symbol ⊙ oznacza, że pewna wartość ustalona kategorii przysługuje wszystkim formom danego fleksemu.

	liczba	przypadek	rodzaj	osoba	stopień	aspekt	zanegowanie	akcentowość	poprzyimkowość	akomodacyjność	aglutynacyjność	wokaliczność
rzeczownik	⊕	⊕	⊙									
forma deprecjatywna	⊙	⊕	⊙									
przymiotnik	⊕	⊕	⊕		⊕							
przymiotnik przyprzym.												
przymiotnik poprzyim.												
przysłówek					⊕							
liczebnik	⊙	⊕	⊕							⊕		
zaimek nietrzecioosobowy	⊙	⊕	⊕	⊙				⊕				
zaimek trzecioosobowy	⊙	⊕	⊕	⊙				⊕	⊕			
zaimek <b>SIEBIE</b>		⊕										
czasownik nieprzeszły	⊕			⊕		⊙						
czas przyszły <b>BYĆ</b>	⊕			⊕		⊙						
aglutynant <b>BYĆ</b>	⊕			⊕		⊙						
pseudoimiesłów	⊕		⊕			⊙					⊕	⊕
rozkaźnik	⊕			⊕		⊙						
bezosobnik						⊙						
bezokolicznik						⊙						
im. przysł. współczesny						⊙						
im. przysł. uprzedni						⊙						
odstownik	⊕	⊕	⊙			⊙	⊕					
im. przym. czynny	⊕	⊕	⊕			⊙	⊕					
im. przym. bierny	⊕	⊕	⊕			⊙	⊕					
winien	⊕		⊕			⊙						
predykatyw												
przyimek		⊙										
spójnik												
kublik												
ciało obce nominalne	⊕	⊕	⊙									
ciało obce luźne												

kowo. Nie widać specjalnych zalet takiego rozstrzygnięcia, dlatego też w niniejszym opisie rzeczowniki *plurale tantum* są traktowane jako leksemy rzeczownikowe defektywne, pozbawione wszystkich form liczby pojedynczej. Identyfikatorami tych leksemów są wykładniki form mianownika liczby mnogiej. Analogicznie jako defektywne potraktowano nieliczne rzeczowniki *singulare tantum*, takie jak *WHO(S)*, *CO(S)*, *WSZYSTKO*.

## 6.2. Przymiotniki

Leksem przymiotnikowy składa się z fleksu przymiotnikowego, zawierającego tradycyjnie rozumiane formy odmiany przymiotnika i dwóch jednoelementowych fleksów zawierających formy nietypowe.

Fleksem **przymiotnikowy** jest odmienny przez liczbę, przypadek, rodzaj i stopień. Przymiotniki niestopniowalne traktowane są jako defektywne, posiadające jedynie formy o wartości pos stopnia.

**Przymiotnik przyprzymiotnikowy** to fleksem zawierający formy typu *biało-* pojawiające się w pierwszych członach złożeń typu *biało-czerwony*. **Przymiotnik poprzyimkowy** to fleksem utworzony dla opisanego form typu *polsku* występujących wyłącznie we frazach typu *po polsku*.

## 6.3. Przysłówki

Fleksem **przysłówkowy** jest odmienny tylko przez stopień. Głównym kryterium wyróżnienia tej klasy jest odprzymiotnikowość. Formy przysłówkowe pochodzące od niestopniowalnych przymiotników uznajemy również za przysłówkowe i przyznajemy im wartość stopnia pos. Do klasy tej należy również przysłówek *bardzo*, który jest stopniowalny, ale dla którego trudno byłoby wskazać przymiotnik. Saloni i Świdziński (2001, s. 101) proponują uznać go za pochodny od przymiotnika **WIELKI**.

Rozważamy włączenie tak rozumianych fleksów przysłówkowych do odpowiednich leksemów przymiotnikowych.

## 6.4. Liczebniki

Fleksemy **liczebnikowe** mają ustaloną wartość liczby, odmieniają się natomiast przez przypadek, rodzaj i akomodacyjność. Tak zdefiniowany fleksem liczebnikowy obejmuje tradycyjnie rozumiane liczebniki główne (formy rodzajów m1, m2, m3, f, n2) i zbiorowe (formy rodzajów n1, p1 i p2) (por. Saloni, 1977).

Kategoria **akomodacyjności** odróżnia formy liczebnika wiążące się z formami rzeczownika o tej samej wartości przypadku (np. *dwaj*) od wiążących formy o wartości przypadku równej dopełniaczowi (np. *dwóch, dwu*). Porównaj:

- (15) *Przyszli dwaj chłopcy.* (congr)
- (16) *Przyszło dwóch chłopców.* (rec)
- (17) *Przyszło dwu chłopców.* (rec)

Formy *dwóch/dwu* w wypowiedzeniach (16)–(17) interpretowane są jako mianownikowe. Odmiennie niż we wspomnianym artykule (Bień i Saloni, 1982) wartość akomodacyjności przypisywana jest wszystkim formom liczebników.

## 6.5. Zaimki

Tak zwane zaimki osobowe zostały opisane za pomocą dwóch klas leksemów: **zaimków nietrzeciosobowych i trzeciosobowych**. Pierwsza z klas zawiera cztery leksemy: *JA*, *TY*, *MY* i *WY*. Przyjęto, że leksemy tej klasy są odmienne przez przypadek, rodzaj (mimo że zachodzi pełna neutralizacja tej kategorii) i akcentowość. Klasa zaimków trzeciosobowych

zawiera tylko jeden leksem o postaci hasłowej ON, odmienny przez przypadek, rodzaj, akcentowość i poprzymkowość. Leksemy i fleksemy obu tych klas mają przypisaną słownikową wartość liczby i osoby.

Jeszcze jedną klasę specjalną, stworzoną dla opisanego wysoce nietypowego leksemu, stanowi **zaimek** SIEBIE. Klasa ta zawiera tylko jeden leksem zawierający jeden fleksem odmienny tylko przez przypadek i w dodatku defektywny (pozbawiony mianownika i wołacza). Należą do niego mianowicie formy o wykładnikach *siebie, sobie, sobą*. Segment *się* ze względu na trudności interpretacyjne jest traktowany zawsze jako wykładnik samodzielnego nieodmiennego leksemu *się* (kublikowego).

### 6.6. Czasowniki

Leksemy czasownikowe są w niniejszym opisie rozumiane dość szeroko. Formy wchodzące w ich skład zostały poklasyfikowane następująco.

Klasę nazwaną **nieprzeszłmi formami finitywnymi czasownika** tworzą formy, które samodzielnie występują w funkcji czasu teraźniejszego (dla czasowników niedokonanych) lub czasu przyszłego prostego (dla czasowników dokonanych).

Tradycyjnie rozumiane formy czasu przeszłego rozbijane są na formę **pseudoimiesłowu** i **aglutynacyjną formę czasownika** BYĆ.

(18)	<i>łgał-em</i>	<i>łgał-es</i>	<i>łgał</i>
	<i>łgała-m</i>	<i>łgała-s</i>	<i>łgała</i>
	<i>łgało-m</i>	<i>łgało-s</i>	<i>łgało</i>
	<i>łgali-śmy</i>	<i>łgali-ście</i>	<i>łgali</i>
	<i>łgały-śmy</i>	<i>łgały-ście</i>	<i>łgały</i>

Opis składniowy bazujący na przedstawionym tu znakowaniu powinien zdawać sprawę z tego, że w funkcji czasu przeszłego czasownika może wystąpić sam pseudoimiesłów (wtedy konstrukcja ma wartość ter kategorii osoby) lub konstrukcja złożona z pseudoimiesłowu i formy aglutynacyjnej (wtedy wartość osoby konstrukcji jest równa wartości osoby aglutynantu).

Pseudoimiesłów może także wystąpić jako składowa czasu przyszłego i trybu warunkowego. Te jego funkcje nie są w żaden sposób oznaczane, jako że ich uwzględnienie wymagałoby rozpatrywania fragmentów tekstu dłuższych niż słowo.

Wymienione segmenty po łączniku w powyższej tabelce nie wyczerpują wszystkich form aglutynacyjnych. W nieciągłym wariacie szyku form czasu przeszłego (typu *-(e)m łgał*) pojawić się mogą inne formy aglutynantu różniące się wokalicznością (na przykład formy *-eśmy* i *-eście*). Formy aglutynacyjne nie muszą stanowić składowej czasu przeszłego, mogą też wystąpić samodzielnie w zdaniach typu:

(19) *Głupiś.*

Czasownik *być* jest źródłem dodatkowej komplikacji, jako bowiem jedyny czasownik niedokonany ma zarówno formy czasu teraźniejszego, jak i przyszłego prostego. Formy czasu teraźniejszego opisujemy za pomocą klasy form nieprzeszłych. Dla form czasu przyszłego prostego określamy specjalną klasę **przyszłych form finitywnych czasownika** BYĆ.

Klasa **rozkazników** obejmuje syntetyczne formy trybu rozkazującego, czyli formy drugiej osoby liczby pojedynczej oraz pierwszej i drugiej osoby liczby mnogiej.

Jest też kilka czasownikowych fleksemów nieodmiennych. Są to: **bezosobnik** (forma bezosobowa typu *łgano*), **bezokolicznik**, **imiesłów przysłówkowy współczesny** i **uprzedni**. Ostatnie odmienne fleksemy czasownikowe stanowią **odśłownik** czyli rzeczownik odśłowny czyli gerundium (w wąskim znaczeniu), **imiesłów przymiotnikowy czynny** oraz **bierny**.

Czasowniki **POWINIEN**, **WINIEN** i **RAD** (być może, nieliczne inne) odmieniają się nietypowo. Mają one mianowicie syntetyczne formy jedynie dla czasu teraźniejszego w trybie oznajmującym. Formy te są przy tym zbudowane w sposób zbliżony do form czasu przeszłego innych czasowników. Dlatego konieczne okazało się wprowadzenie dla nich osobnego fleksemu i leksemu nazwanego winien. Formy tego fleksemu mają charakterystykę taką jak pseudoimiesłów, tak więc np. słowo *powinieneś* stanowi wykładnik dwóch form: formy winien i aglutynantu *-eś*.

Klasa leksemów i fleksemów nazwana **predykatywem** reprezentuje czasowniki niewłaściwe o odmianie czysto analitycznej takie jak **MOŻNA**, **TRZEBA**, **WIDAĆ**.

### 6.7. Leksemy nieodmienne

Kryteria czysto fleksyjne nie pozwalają na podzielenie leksemów nieodmiennych na takie klasy jak przyimki, spójniki i partykuło-przysłówki. Rozróżnienie między klasami nieodmiennymi odbywa się więc na zasadzie różnic własności składniowych.

Jako **przyimki** wyróżniono leksemy pełniące w wypowiedzeniu funkcję łączącą i wymagające określonego przypadku. Za **spójniki** uznano leksemy pełniące funkcję łączącą, ale nie wymagające określonego przypadku. Pozostałe leksemy nieodmienne należące do systemu leksykalnego polszczyzny zaliczono do resztkowej klasy **kublików** (odpowiadającej mniej więcej partykuło-przysłówkom).

### 6.8. Ciała obce

Ciała obce to elementy nie należące do systemu fleksyjnego polszczyzny, pojawiające się w polskich tekstach. Należą do tej klasy w szczególności elementy pochodzące z innych języków, także nazwy własne. Postanowiono, że w korpusie będzie się dzielić ciała obce na dwie podklasy. Mianowicie te ciała obce, które występują w kontekście wymagającym przypadku, będą oznaczane tak, jak rzeczowniki (zwłaszcza jeśli wykazują szczątkową odmienność). Są to **ciała obce nominalne**. Pozostałe ciała obce stanowią **ciała obce luźne**. Celem tego postępowania jest zdanie sprawy z uwikłań składniowych niektórych ciał obcych.

Identyfikator leksemu dla ciał obcych nominalnych stanowi wykładnik formy, która byłaby użyta w kontekście wymagającym mianownika. Identyfikator leksemu dla ciał obcych luźnych stanowi ten sam element, który jest wykładnikiem samego ciała obcego.

## 7. Struktura znaczników morfosyntaktycznych

Technicznym szczegółem notacji, który wszakże warto dopowiedzieć, jest sposób zapisywania znaczników. W podstawowej formie znacznik morfosyntaktyczny jest ciągiem wartości rozdzielonych dwukropkami, np.: subst:sg:nom:m1 dla segmentu *chłopiec*. Pierwsza wartość określa **klasę fleksemów** (za pomocą oznaczeń z tabeli 1), następne — wartości **kate-**

**gorii gramatycznych** przysługujących danej formie. Wartości są wymieniane w kolejności zgodnej z zestawieniem w punkcie 5 (i tym samym z porządkiem kolumn w Tabeli 2).

W wypadku, gdy dla danego segmentu występuje niejednoznaczność wartości jakiejś kategorii, podawane jest kilka znaczników, np.: subst:sg:nom:m3 i subst:sg:acc:m3 dla segmentu *stół*. Dopuszczalna jest skrócona notacja takich znaczników. Polega ona na wymienieniu alternatywnych wartości danej kategorii rozdzielonych kropką. Na przykład powyższe dwa znaczniki można zapisać łącznie: subst:sg:nom.acc:m3. Ta notacja może być stosowana również do skrótowego opisanego form kilku leksemów na przykład rzeczownikowych różniących się jedynie rodzajem.

Należy zwrócić uwagę, że alternatywy na więcej niż jednej pozycji znacznika oznaczają wszystkie możliwe kombinacje wartości. Dlatego na przykład nie da się zapisać łącznie znaczników adj:sg:gen:m1.m2.m3.n1.n2:pos i adj:sg:acc:m1.m2:pos.

Wartości kategorii, które są istotne tylko dla niektórych fleksemów danej klasy lub niektórych form, są notowane tylko dla tych fleksemów/form.

## 8. Podsumowanie

Jak starałem się wykazać, znakowanie morfosyntaktyczne jest koniecznością w przypadku korpusu języka polskiego. Przygotowanie korpusu tak, aby możliwe było automatyczne wyszukiwanie, wymaga opracowania sposobu znakowania tekstów, w którym klasyfikacja form jest prowadzona znacznie bardziej rygorystycznie i szczegółowo niż w zastosowaniach niekomputerowych.

Celem prac prowadzonych w IPI PAN jest opracowanie sposobu znakowania, który byłby klarowny i konsekwentny lingwistycznie, a zarazem spełniał wymogi ścisłości zastosowań komputerowych.

## Bibliografia

- J. S. Bień (1991) *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa Uniwersytetu Warszawskiego.
- J. S. Bień, Z. Saloni (1982) *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*, Prace Filologiczne, t. XXXI, s. 31–45.
- J. S. Bień, W. Łukaszewicz, S. Szpakowicz (1973) *Opis systemu MARYSIA. I. Zasady pisania scenariusza i scenopisu*, Sprawozdania Instytutu Maszyn Matematycznych i Zakładu Obliczeń Numerycznych Uniwersytetu Warszawskiego Nr 41, Wydawnictwa Uniwersytetu Warszawskiego.
- Z. Saloni (1976) *Kategoria rodzaju we współczesnym języku polskim*, w: *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, s. 41–75, Ossolineum, Wrocław.
- (1977) *Kategorie gramatyczne liczebników we współczesnym języku polskim*, w: *Studia gramatyczne I*, s. 145–173, Wrocław.
- (1981) *Uwagi o opisie fleksyjnym tzw. zaimków rzeczownych*, „Acta Universitatis Lodziensis”, Folia Linguistica 2, s. 265–271.

- (1988) *O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie*, Biuletyn Polskiego Towarzystwa Językoznawczego, t. XLI, s. 155–166.
- (2001) *Czasownik polski. Odmiana, słownik*, Wiedza Powszechna, Warszawa.
- Z. Saloni, M. Świdziński (2001) *Składnia współczesnego języka polskiego*, Wydawnictwo Naukowe PWN, Warszawa, wyd. piąte.
- J. Tokarski (2001) *Fleksja polska*, Klasyka Językoznawstwa Polskiego, Wydawnictwo Naukowe PWN, wyd. III z uzupełnieniami.
- M. Świdziński (1992) *Gramatyka formalna języka polskiego*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

## SUMMARY

### Morphosyntactic tagset in the IPI PAN corpus

The aim of this paper is to describe a morphosyntactic tagset used to mark up a large-scale corpus of Polish being developed at the Institute of Computer Science, Polish Academy of Sciences.

The key point of this tagset is the notion of *flexeme* — a set of wordforms which inflect in a uniform way. Lexemes are seen as sets of flexemes.

The paper discusses segmentation and lemmatization of Polish text, grammatical categories and lexemic and flexemic classes which were distinguished in this tagset.