Marcin Woliński, Adam Przepiórkowski

Projekt anotacji morfosyntaktycznej korpusu języka polskiego

Nr 938

Warszawa, grudzień 2001

Streszczenie

Niniejszy raport zawiera propozycję zestawu znaczników morfosyntaktycznych do anotacji korpusu tekstów języka polskiego. Zestaw ten opiera się wyłącznie na kryteriach morfologicznych i składniowych, w tych terminach zdefiniowane zostało też pojęcie klasy gramatycznej, zwykle utożsamiane z semantycznym pojęciem leksemu. Raport zawiera także ogólne wskazówki praktyczne dotyczące anotowania korpusu z wykorzystaniem zaproponowanego tu zestawu znaczników, a także porównanie tego zestawu znaczników z innymi systemami znakowania morfosyntaktycznego zaproponowanymi dla języka polskiego i innych języków słowiańskich.

Słowa kluczowe: zestaw znaczników morfosyntaktycznych (tagset), korpus tekstów, przetwarzanie języka naturalnego, morfologia i składnia.

Abstract

Project of a morphosyntactic tagset for Polish

This report presents a morphosyntactic tagset for Polish based solely on morphological and syntactic criteria. In particular, the notion of part-of-speech, often equated with the essentially semantic notion of lexeme, is defined in purely morphosyntactic terms. The report also contains general guidelines for text annotation using the proposed tagset, as well as a comparison with other tagsets proposed for Polish and other Slavic languages.

Keywords: tagset, text corpus, NLP, morphosyntax.

Spis treści

1. Zestaw znaczników morfosyntaktycznych

1.1. Uwagi ogólne

Opisywany tutaj sposób znakowania morfosyntaktycznego ma zostać użyty do anotowania obszernego korpusu tekstów polskich.[1] Korpus ten jest przygotowywany z myślą o zastosowaniach dotyczących

- automatycznego przetwarzania języka naturalnego,
- badania struktury składniowej polszczyzny,
- metod automatycznej analizy składniowej,
- prac leksykograficznych,
- innych związanych np. z wyszukiwaniem pełnotekstowym.

Biorąc pod uwagę, że poszczególni badacze reprezentujący te dziedziny posługują się odmiennym aparatem pojęciowym (nawet w obrębie samej składni), trudno zaproponować sposób anotacji, który odpowiadałby wszystkim. Nie czujemy się ani kompetentni ani władni narzucić tutaj jakiegoś jednego rozwiązania.

Dlatego też proponujemy sposób znakowania, który jest w pewnym sensie minimalny. Znakowanie ma mianowicie służyć jedynie dezambiguacji morfologicznej tekstu. Nie ma natomiast nieść w sobie kompletu informacji, które mógłby zawierać słownik morfologiczno-składniowy.

Mamy nadzieję, że dzięki temu przyjęte przez nas rozwiązania nie mają zbyt drastycznych implikacji dla systemu pojęciowego stosowanego przez jego użytkowników. Mówiąc inaczej, chcielibyśmy żeby proponowane tutaj znakowanie dało się łatwo odwzorować w różne rozwiązania szczegółowe.

Wynika z tego, że pomijamy cechy czysto słownikowe — przysługujące wszystkim formom danego leksemu i nie powodujące niejednoznaczności przy znakowaniu. W taki też sposób traktujemy przynależność do „części mowy". Nie widzimy na przykład potrzeby wyróżnienia klasy liczebników porządkowych. Wyrazy takie opisujemy jako przymiotniki, ze względu na sposób, w jaki się one odmieniają. Jeżeli do jakichś celów potrzebne jest wyróżnienie liczebników porządkowych, należy zestawić dane ze znakowanego korpusu ze słownikiem leksemowym czyniącym odpowiednie rozróżnienie. Zatem korzystanie ze znakowanego korpusu zamiast czystego tekstu ma uwolnić użytkownika od konieczności interpretowania słów jako realizacji tekstowych form pewnych leksemów i tylko tyle. W korpusie bywają jednak notowane pewne cechy przysługujące całym leksemom, na przykład aspekt czasowników. Dzieje się tak dlatego, że zdarzają się czasowniki dwuaspektowe, dla form których można sensownie mówić o ujednoznacznianiu aspektu. Podobnie notujemy rodzaj dla rzeczowników ze względu na rzeczowniki wielorodzajowe (a więc zbiory homonimicznych leksemów różniących się wartością rodzaju, np. *pływak*, por. niżej).

W dalszym opisie staramy się nie operować pojęciem leksemu, bo jest ono zwykle definiowane na gruncie semantyki. Tymczasem zasadniczym przedmiotem przedstawionego tutaj opisu są własności morfologiczne i składniowe definiowane bez odwołań semantycznych. Gdy pojęcie leksemu przewija się w omówieniu przykładów poniżej należy je rozumieć w sensie pracy Saloni i Świdziński, 2001.

1.2. Segmentacja tekstu

Nie jest rzeczą oczywistą, jaki jest optymalny sposób rozbicia maszynowego przetwarzania języka naturalnego na fazy (i czy w ogóle jest to możliwe i pożądane). Narzuca się sposób działania polegający na wyraźnym oddzieleniu etapu segmentacji od analizy morfologicznej i dalej od ewentualnego ujednoznaczniania wyników analizy morfologicznej albo od analizy składniowej, semantycznej itd. Większość dotychczasowych badań była prowadzona w taki sposób i wydaje się, że jest on akceptowalny.

[1] Autorzy dziękują Łukaszowi Dębowskiemu za owocną dyskusję i Monice Korczakowskiej za wnikliwe uwagi do wcześniejszej wersji raportu. Niniejsza publikacja powstała w ramach projektu badawczego Nr 7 T11C 043 20 finansowanego przez Komitet Badań Naukowych w latach 2001–2004. Prace omawiające inne aspekty anotowania korpusu tekstów polskich w ramach tego samego projektu to: Bański, 2001, Dębowski, 2001 oraz Hajnicz i Kupść, 2001.

Użyteczną definicję zakresu automatycznej analizy morfologicznej daje przyjęcie kategorycznej zasady nierozpatrywania żadnych segmentów zawierających odstęp. Znaczy to, że wszelkie zależności między jednostkami oddzielonymi odstępem przenosi się niejako na poziom składniowy. Przyjmujemy tę zasadę, ponieważ sądzimy, że nie przeszkadza ona w uzyskaniu opisu sensownego lingwistycznie.

Niektóre segmenty nie zawierające odstępu podlegają jednak podziałowi. Znaki interpunkcyjne uznajemy za osobne segmenty, z wyjątkiem kropki stojącej po skrótach. Znaki interpunkcyjne nie podlegają znakowaniu morfosyntaktycznemu.

Tak więc tekst polski traktujemy jako ciąg segmentów literowych bez odstępu (słów) i znaków interpunkcyjnych (każdy stanowi osobny segment). Tekst anotowany (znakowany) dodatkowo przy każdym słowie zawiera znacznik morfosyntaktyczny i formę podstawową. Te dwa elementy pozwalają zinterpretować słowo jako wykładnik formy pewnej abstrakcyjnej jednostki należącej do określonej klasy gramatycznej.

Formy czasu przeszłego czasowników traktujemy jako złożone z formy pseudoimiesłowu (o charakterystyce liczbowo-rodzajowej) i formy aglutynacyjnej leksemu BYĆ (o charakterystyce liczbowo-osobowej). Formy trybu warunkowego uznajemy za złożone z formy pseudoimiesłowu, partykuło-przysłówka warunkowego BY i formy aglutynacyjnej.

Rozbijamy formy pisane z łącznikiem (na 3 segmenty). Za osobne formy uznajemy aglutynacyjną postać *-ń* zaimka osobowego *on*, słowa *-że*, *-ż* i *-ć* stanowiące wykładniki leksemów wzmacniających ŻE i Ć oraz słowo *-li* stanowiące wykładnik leksemu pytającego LI.

1.3. Struktura znaczników

Materiał językowy dzielimy na klasy gramatyczne wyróżnione ze względu na to, czy przysługują im pewne charakterystyczne kategorie fleksyjne. Pomocniczo posługujemy się zróżnicowaniem ze względu na własności składniowe.

Kategorie fleksyjne należy tu rozumieć w sensie pracy Saloni i Świdziński, 2001, s. 87, a więc jako regularne opozycje w zbiorze form, które jesteśmy skłonni uznać za przynależne do tej samej jednostki słownikowej.

Znaczniki służą do notowania przynależności form do klas gramatycznych oraz notowania wartości kategorii gramatycznych przysługujących formom. Łącznie te cechy będziemy nazywać atrybutami formy.

Z technicznego punktu widzenia znakowanie ma dwa warianty: pozycyjny i skrócony. W wariancie pozycyjnym dla każdej formy notowane są wartości wszystkich atrybutów (klasy gramatycznej i wszystkich wymienionych poniżej kategorii gramatycznych). Jeżeli dany atrybut nie przysługuje jednostkom danej klasy gramatycznej na odpowiedniej pozycji stoi wartość `[]`, którą należy rozumieć jako „nie dotyczy". W wariancie skróconym podawane są tylko atrybuty, które przysługują danej klasie. Wariant pozycyjny jest wygodniejszy do automatycznego przetwarzania, wariant skrócony wydaje się łatwiejszy do opanowania dla człowieka. Oczywiście są one informacyjnie równoważne.

Ze względu na liczne systemowe synkretyzmy na każdej pozycji notowany jest zbiór wartości, które mogą być przypisane danemu słowu. Zapis ten można interpretować jako skrót zbioru alternatywnych znaczników powstałego przez wymnożenie kartezjańskie niejednoznaczności na poszczególnych pozycjach. Na przykład przy słowie *obmierzłym* może stać znacznik (skupmy się na kategoriach liczby, przypadka i rodzaju, wyjaśnienie symboli w następnym punkcie):

(1) `[sg],[inst,loc],[m1,m2,m3,n1,n2]`

Zapis ten oznacza, że w istocie słowo *obmierzłym* może być interpretowane jako wykładnik jednej z dziesięciu form o następujących wartościach kategorii gramatycznych:

(2)
```
sg,inst,m1  sg,loc,m1
sg,inst,m2  sg,loc,m2
sg,inst,m3  sg,loc,m3
sg,inst,n1  sg,loc,n1
sg,inst,n2  sg,loc,n2
```

Oczywiście nawet przy stosowaniu zapisów typu (1) zdarza się, że dane słowo zaopatrzone jest w kilka wariantywnych znaczników. Na przykład dla słowa *obmierzłym* kompletny opis (tylko ze względu na liczbę, przypadek i rodzaj) wygląda następująco:

(3) `obmierzłym[[sg],[inst,loc],[m1,m2,m3,n1,n2];`
`[pl],[dat],[m1,m2,m3,f,n1,n2,p1,p2,p3]]`

czyli istnieje 19 form o wykładniku *obmierzłym*.

1.4. Kategorie gramatyczne

Kategorie gramatyczne rozumiemy w duchu dystrybucyjnym, przyjmujemy więc, że wartości dopuszczalne dla danej kategorii muszą być takie same niezależnie od tego, do jakiej klasy gramatycznej należy rozpatrywana w danym momencie forma. Inaczej mówiąc, uznajemy, że uzgodnienia kategorii gramatycznych polegają na równości ich wartości dla określonych form w wypowiedzeniu. Przyjęcie, że np. rzeczowniki i przymiotniki mają różne zbiory wartości kategorii rodzaju mogłoby pozwolić na zmniejszenie liczby dopuszczalnych wartości. Wtedy trzeba by jednak w nietradycyjny sposób opisać uzgodnienie, wprowadzając nietrywialny operator uzgodnienia, zdający sprawę z tego, jakie wartości do siebie „pasują". Pozostajemy zatem przy rozwiązaniu tradycyjnym, czystszym z punktu widzenia teoretycznego.

Oto lista kategorii gramatycznych, które uznaliśmy za istotne dla opisu polskich form wyrazowych. Dla każdej kategorii podajemy zestaw wartości, które może ona przyjmować. Większość stanowią kategorie fleksyjne — uwidaczniające się jako opozycje w obrębie paradygmatów poszczególnych jednostek. Niektóre, jak na przykład aspekt, są kategoriami czysto słownikowymi interweniującymi tylko na poziomie składniowym.

- Liczba: pojedyncza sg (*oko*), mnoga pl (*oczy*).
- Przypadek: mianownik nom (*woda*), dopełniacz gen (*wody*), celownik dat (*wodzie*), biernik acc (*wodę*), narzędnik inst (*wodą*), miejscownik loc (*wodzie*), wołacz voc (*wodo*).
- Rodzaj: męski osobowy m1 (*papież, kto*), męski zwierzęcy m2 (*baranek*), męski rzeczowy m3 (*stół*), żeński f (*stuła*), nijaki zbiorowy n1 (*dziecko*), nijaki zwykły n2 (*okno, co*), przymnogi osobowy p1 (*wujostwo*), przymnogi zwykły p2 (*skrzypce*), przymnogi opisowy p3 (*spodnie*).
- Osoba: pierwsza pri (*bredzę*), druga sec (*bredzisz*), trzecia ter (*bredzi*).
- Stopień: równy pos (*cudny*), wyższy comp (*cudniejszy*), najwyższy sup (*najcudniejszy*).
- Aspekt: niedokonany imperf (*iść*), dokonany perf (*zajść*).
- Negacja: niezanegowana aff (*pisanie, czytanego*), zanegowana neg (*niepisanie, nieczytanego*).
- Deprecjatywność: niedeprecjatywna ndepr (*chłopi*), deprecjatywna depr (*chłopy*).
- Akcentowość: akcentowana akc (*jego, niego*), nieakcentowana nakc (*go, -ń*).
- Poprzyimkowość: poprzyimkowa praep (*niego, -ń*), niepoprzyimkowa npraep (*jego, go*).
- Akomodacyjność: uzgadniająca congr (*dwaj*), rządząca rec (*dwóch, dwu*).
- Aglutynacyjność: nieaglutynacyjna nagl (*niósł, dlaczego*), aglutynacyjna agl (*niosł-, dlaczegó-*).
- Wokaliczność: wokaliczna wok (*-em, -eś, ze*), niewokaliczna nwok (*-m, -ś, z*).

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.

Również dla kategorii **przypadka** przyjmujemy tradycyjny zestaw wartości (mimo że niektórzy badacze nie uznają wołacza za przypadek twierdząc, że występuje on poza strukturą zdania).

Zestaw kategorii **rodzaju** przejmujemy z pracy Saloniego (1976b). 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 frazami z udziałem form liczebnikowych, może operować prostym rodzajem nijakim stanowiącym sumę klas n1 i n2.

Wbrew badaczom, którzy uznają stopniowanie za zjawisko słowotwórcze, a nie fleksyjne, wprowadzamy kategorię fleksyjną **stopnia** przysługującą (niektórym) 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. Kategoria ta została wprowadzona ze względu na istnienie czasowników dwuaspektowych a więc występowanie w tekstach słów, które mogą być zinterpretowane jako formy czasowników różnego aspektu. Ponadto, jak się okaże dalej, jawne znakowanie aspektu pozwala uprościć notowanie własności czasowych form czasownikowych.

Kategoria fleksyjna **negacji** przysługuje w naszym opisie odsłownikom i imiesłowom przymiotnikowym czynnym i biernym. Uznajemy mianowicie, że np. formy *celebrowanie* i *niecelebrowanie* należą do tej samej jednostki i różnią się wartością negacji. W wypadku tych trzech klas gramatycznych obecność cząstki *nie-* ma konsekwencje składniowe. Kategoria negacji nie dotyczy przymiotników, np. *niełady*, w ich wypadku bowiem nie ma podobnych konsekwencji.

Zaproponowana w artykule Saloni, 1988 kategoria **deprecjatywności** odróżnia formy typu *chłopi* (niedeprecjatywna) od *chłopy* (deprecjatywna). Przysługuje ona wyłącznie formom mianownika i wołacza liczby mnogiej rzeczowników rodzaju m1.

(4) Przyszli uroczy profesorowie.
(5) Przyszły głupie profesory.

Zgodnie z tezami artykułu Saloniego przyjmujemy, że takie pary form istnieją dla każdego rzeczownika rodzaju m1, choć czasami dochodzi do neutralizacji. Mamy świadomość, że opis dystrybucyjny, w którym formy deprecjatywne mają przypisany rodzaj m1 musi rozciągać kategorię deprecjatywności również na inne klasy gramatyczne (co najmniej przymiotniki, liczebniki i czasowniki), aby wykluczyć wypowiedzenia typu:

(6) *Przyszli głupi profesory.

Świadomie nie czynimy tego uznając, że zjawisko o takiej skali nie zasługuje na zastosowanie tak drastycznych środków. Skłaniamy się raczej do naruszenia dystrybucyjności opisu i przemycenia na poziomie składniowym informacji, że formy deprecjatywne rzeczowników rodzaju m1 zachowują się jakby były rodzaju m2. Opis morfologiczny jest więc czyniony przy takim założeniu.

Kategoria **akcentowości** została wprowadzona dla odróżnienia form zaimka ON występujących w zdaniu na pozycji akcentowanej (*jego*) od form występujących na pozycji nieakcentowanej (*go*).

Poprzyimkowość różnicuje formy zaimka ON które mogą wystąpić wyłącznie po przyimku (*przez niego, do-ń*) od występujących w innych kontekstach (*jego, go*).

Kategoria **akomodacyjności**, zaproponowana w artykule Bień i Saloni, 1982, przysługuje wyłącznie formom liczebnikowym w mianowniku rodzaju m1. Odróżnia ona formy liczebnika wiążące się z formami rzeczownika o tej samej wartości przypadka (np. *dwaj*) od wiążących formy o wartości przypadka różnej od własnej (np. *dwóch, dwu*). Porównaj:

(7) Przyszli dwaj chłopcy. (congr)
(8) Przyszło dwóch chłopców. (rec)
(9) Przyszło dwu chłopców. (rec)

Formy *dwóch/dwu* w wypowiedzeniach (8)–(9) interpretujemy jako mianownikowe. Odmiennie niż we wspomnianym artykule wartość akomodacyjności przypisujemy formom nom m1 wszystkich liczebników. Większość liczebników ma jedynie formę o wartości rec.

Aglutynacyjność to kategoria odróżniająca formy niesamodzielne obligatoryjnie dołączające jakąś formę aglutynacyjną od form niedopuszczających aglutynantu. Zróżnicowanie to ujawnia się tylko dla pseudoimiesłowów niektórych czasowników (np. GNIEŚĆ: *gniótł* ale *gniotł-em*), zaimka ON (*niego* ale *do-ń*) i partykuło-przysłówków (np. DLACZEGO: *dlaczego* ale *dlaczegó-ż*, dotyczy to jedynie łączliwości z -ŻE).

W celu odróżnienia wariantu aglutynantu, który występuje po formie kończącej się spółgłoską (zawierającego *e*), od wariantu posamogłoskowego, wprowadzono kategorię **wokaliczności**. Podobne rozróżnienie dotyczy też niektórych przyimków i partykuło-przysłówków, w tym wypadku jednak *e* pojawia się na końcu formy i uwikłanie dotyczy formy występującej po danej.

Powyższa lista kategorii nie uwzględnia czasu, trybu ani strony czasowników. Jak się okaże dalej, kategoria czasu będzie wyrażana pośrednio poprzez specyficzne pogrupowanie form czasownikowych w klasy gramatyczne. W podobny sposób traktowany jest tryb rozkazujący. Tryb warunkowy jest dla nas funkcją składniową (por. p. 1.2), a nie kategorią gramatyczną, podobnie jak kategoria strony.

1.5. Klasy gramatyczne

Przedstawiony niżej system klas gramatycznych ma w założeniu być na tyle szczegółowy, aby można z niego było przez proste transformacje uzyskiwać opis potrzebny w danym zastosowaniu. Wśród lingwistów nie ma na przykład zgody, czy imiesłowy przymiotnikowe uznawać za formy czasownika, czy też traktować je jako zwykłe leksemy przymiotnikowe. W poniższym opisie wydzielamy imiesłowy przymiotnikowe w osobną klasę odróżnialną od przymiotników. Dzięki temu w zależności od potrzeb można albo uznać je za przymiotniki albo włączyć do odpowiednich leksemów czasownikowych. Podobnie postępujemy z innymi „kontrowersyjnymi" klasami.

Trudno powiedzieć, żeby klasy gramatyczne odpowiadały tradycyjnie rozumianym „częściom mowy". Stosowany przez nas podział jest bardziej szczegółowy. Zaproponowane tutaj jednostki przypominają swoim zakresem pojęcie fleksemu wprowadzone w pracy Bień, 1991. Mimo że nie jest to odpowiedniość dokładna, pozwalamy sobie zapożyczyć ten termin.

Zasadniczym kryterium wyróżnienia poniższych klas są kategorie morfologiczne przysługujące poszczególnym fleksemom. Przyjęliśmy, że wszystkie formy należące do danego fleksemu muszą być jednolicie (lub niemal jednolicie) zróżnicowane ze względu na właściwe im kategorie gramatyczne.

Dlatego na przykład tradycyjne leksemy czasownikowe zostały rozbite na kilka osobnych jednostek. Mamy więc fleksemy „nieodmienne" takie jak bezokolicznik i bezosobnik, odmienny przez liczbę i rodzaj pseudoimiesłów (formy osobowe mogące pełnić funkcję 3. os. czasu przeszłego, ale również stanowiące składowe form 1. i 2. os. oraz czasu przyszłego złożonego i trybu warunkowego), odmienne przez liczbę i osobę formy czasu teraźniejszego/przyszłego prostego itd.

Dwie klasy wprowadzono wyłącznie dla czasownika BYĆ. Pierwszą dla opisania czasu przyszłego prostego tego czasownika, ponieważ jako jedyny czasownik niedokonany ma on formy tego czasu. (Klasa form nieprzeszłych dla BYĆ opisuje czas teraźniejszy.) Druga klasa to formy aglutynacyjne (niesamodzielne) czasu teraźniejszego czasownika BYĆ (czyli *-(e)m*, *-(e)ś*, itd.). Formy wymienionych klas specjalnych są wykorzystywane do tworzenia form analitycznych innych czasowników.

Dokładniej kryteria przynależności do poszczególnych klas omówiono w p. 1.6.

- rzeczownik subst
- przymiotnik adj
- niesamodzielna forma przymiotnika (np. *polsko*) adj0
- przysłówek stopniowalny adv
- przysłówek przyprzyimkowy adv0
- liczebnik num
- zaimek osobowy *on* ppron
- nieprzeszła forma finitywna czasownika fin
- przyszła forma finitywna czasownika BYĆ bedzie
- forma aglutynacyjna czasownika BYĆ agl
- pseudoimiesłów praet

- rozkaźnik impt
- bezosobnik imps
- bezokolicznik inf
- imiesłów przysłówkowy współczesny pcon
- imiesłów przysłówkowy uprzedni pant
- odsłownik ger
- imiesłów przymiotnikowy czynny pact
- imiesłów przymiotnikowy bierny ppas
- przyimek prep
- spójnik conj
- partykuło-przysłówek part
- ciało obce ($2\pi r$, H_2O, http://www.ipipan.waw.pl) xxx

Niektóre z powyższych klas wymagają zapewne komentarza. **Niesamodzielna forma przymiotnika** to forma typu *biało-* pojawiająca się w pierwszych członach złożeń typu *biało-czerwony*. **Przysłówek przyprzyimkowy** to forma typu *polsku* występująca wyłącznie we frazach typu *po polsku*.

Klasa **liczebników** obejmuje tzw. liczebniki główne i zbiorowe traktowane łącznie jako formy tego samego fleksemu.

Osobną klasę stanowi odmienny przez osobę (!) zaimek ON. Ma on formy typu *ja, jej, ono, -ń, nam, wami, nie* itd.

Formy wchodzące w zakres szeroko rozumianego leksemu czasownikowego zostały poklasyfikowane następująco. Formy czasu teraźniejszego (dla czasowników niedokonanych) lub czasu przyszłego prostego (dla czasowników dokonanych) tworzą klasę **nieprzeszłych form finitywnych czasownika**. Tradycyjnie rozumiane formy czasu przeszłego rozbijamy na formę **pseudoimiesłowu** i **aglutynacyjną formę czasownika** BYĆ.

(10)

łgał-em	łgał-eś	łgał
łgała-m	łgała-ś	łgała
łgało-m	łgało-ś	łgało
łgali-śmy	łgali-ście	łgali
łgały-śmy	łgały-ście	łgały

Ewentualny opis składniowy bazujący na przedstawionej tu anotacji 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).

Słowa po łączniku w powyższej tabelce nie wyczerpują wszystkich form aglutynacyjnych. W nieciągłym wariancie 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:

(11) 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 przy pomocy zwykłej klasy form nieprzeszłych. Dla form czasu przyszłego prostego określamy specjalną klasę **przyszłych form finitywnych czasownika** BYĆ.

Klasa **rozkaźników** 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 klasy czasownikowe (lub odczasownikowe) stanowią **odsłownik** czyli rzeczownik odsłowny lub gerundium (w wąskim znaczeniu), **imiesłów przymiotnikowy czynny** oraz **bierny**.

Pozostałe klasy gramatyczne obejmują fleksemy nieodmienne: **przyimek**, **spójnik**, **partykuło-**
-przysłówek i **ciało obce**. Do klasy partykuło-przysłówków są zaliczane wszelkie jednostki trudne do sklasyfikowania.

1.6. Zestawienie kategorii przysługujących poszczególnym klasom gramatycznym

W poniższej tabeli użyto symbolu ⊕, jeżeli dla danej klasy dana kategoria jest morfologiczna (czyli fleksem „odmienia się" przez tę kategorię). Symbol ⊙ oznacza, że pewna wartość ustalona kategorii przysługuje wszystkim formom danego fleksemu odróżniając go od pewnych innych fleksemów tej samej klasy.

	liczba	przypadek	rodzaj	osoba	stopień	aspekt	negacja	deprecjatywność	akcentowość	poprzyimkowość	akomodacyjność	aglutynacyjność	wokaliczność
rzeczownik	⊕	⊕	⊙					⊕					
przymiotnik	⊕	⊕	⊕		⊕								
przysłówek stopn.					⊕								
liczebnik	⊙	⊕	⊕								⊕		
zaimek	⊕	⊕	⊕	⊕					⊕	⊕		⊕	
czas. nieprzeszły	⊕			⊕		⊙							
czas. przyszły *być*	⊕			⊕		⊙							
aglutynant	⊕			⊕		⊙						⊙	⊕
pseudoimiesłów	⊕		⊕			⊙						⊕	
rozkaźnik	⊕			⊕		⊙							
bezosobnik						⊙							
bezokolicznik						⊙							
im. przys. współczesny						⊙							
im. przys. uprzedni						⊙							
odsłownik	⊕	⊕	⊙			⊙	⊕						
im. przym. czynny	⊕	⊕	⊕			⊙	⊕						
im. przym. bierny	⊕	⊕	⊕			⊙	⊕						
przyimek		⊙											⊕
spójnik													
partykuło-przysłówek												⊕	⊕

W powyższej tabeli wyraźnie wyróżnia się grupa fleksemów deklinacyjnych — o kategorii fleksyjnej przypadka, oraz grupa fleksemów 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:

(12) Rozpoczęcie spawania konstrukcji nastąpiło w marcu.
(13) *Rozpoczęcie zespawania konstrukcji nastąpiło w marcu.

Do klasy przysłówków stopniowalnych zaliczamy fleksemy odmienne tylko przez stopień, pozostałe z tradycyjnie rozumianych przysłówków trafiają do klasy partykuło-przysłówków. Dla przymiotników odmienność przez stopień uznajemy za fakultatywną, przymiotniki niestopniowalne mają dla wszystkich form wartość pos stopnia.

Pozostałe cztery klasy są w zasadzie nieodmienne, rozróżnienie między nimi odbywa się na zasadzie odmiennych własności składniowych. Wymaganie przypadka jest charakterystyczne dla przyimków. Odmienność przez wokaliczność i aglutynacyjność trudno oczywiście uznać za cechę wyróżniającą, dotyczy ona bowiem jedynie pojedynczych fleksemów.

1.7. Formy podstawowe (lematy)

Przez analizę morfologiczną rozumie się przypisanie słowu tekstowemu interpretacji jako wykładnika pewnej formy pewnego leksemu. Oznacza to, że oprócz podania typu leksemu i położenia formy w paradygmacie tego typu (a więc podania omawianych wyżej atrybutów formy) należy wskazać, o jaki chodzi leksem. Trzeba więc podać konwencjonalną nazwę tego leksemu (formę podstawową, formę hasłową, lemat).

W naszym opisie posługujemy się jednak pojęciem drobniejszych jednostek, fleksemów. Dlatego też w niniejszym punkcie wprowadzimy formy podstawowe dla fleksemów. Ich dobór jest konwencjonalny, staramy się więc, aby nie był on zbyt szokujący z punktu widzenia tradycyjnego. W szczególności dla niektórych fleksemów forma podstawowa nie należy do opisywanego fleksemu (np. dla form finitywnych — bezokolicznik).

Poniżej wykaz form podstawowych dla poszczególnych klas gramatycznych.

- rzeczownik subst — mianownik liczby pojedynczej
- przymiotnik adj — mianownik liczby pojedynczej rodzaju męskiego osobowego
- niesamodzielna forma przymiotnika adj0 — mianownik liczby pojedynczej rodzaju męskiego osobowego
- przysłówek stopniowalny adv — forma stopnia równego
- przysłówek przyprzyimkowy adv0 — mianownik liczby pojedynczej rodzaju męskiego osobowego przymiotnika
- liczebnik num — mianownik rodzaju męskiego osobowego o akomodacyjności rec
- zaimek osobowy *on* ppron — *on*
- nieprzeszła forma finitywna czasownika fin — bezokolicznik
- przyszła forma finitywna czasownika BYĆ bedzie — *być*
- forma aglutynacyjna czasownika BYĆ agl — *być*
- pseudoimiesłów praet — bezokolicznik
- rozkaźnik impt — bezokolicznik
- bezosobnik imps — bezokolicznik
- bezokolicznik inf — bezokolicznik
- imiesłów przysłówkowy współczesny pcon — bezokolicznik
- imiesłów przysłówkowy uprzedni pant — bezokolicznik
- odsłownik ger — mianownik liczby pojedynczej
- imiesłów przymiotnikowy czynny pact — bezokolicznik
- imiesłów przymiotnikowy bierny ppas — bezokolicznik
- przyimek prep — forma niewokaliczna
- spójnik conj — jedyna forma tego fleksemu
- partykuło-przysłówek part — forma niewokaliczna
- ciało obce xxx — jedyna forma tego fleksemu

2. Wskazówki dotyczące anotowania

2.1. Uwagi ogólne

Do eliminacji niejednoznaczności oznaczeń morfosyntaktycznych wykorzystywać należy jedynie informacje dostępne w obrębie wypowiedzenia (zdania). Tak więc np. nie należy ujednoznaczniać rodzaju formy *jego* w wypowiedzeniu

(14) To jest jego piłka.

nawet, jeśli z otaczających zdań wiadomo, że chodzi o chłopca (m1), a nie psa (m2).

Jak powiedziano wcześniej, w niektórych sytuacjach słowo rozpatrywane w oderwaniu musi być opisane alternatywą kilku osobnych znaczników. Może się również zdarzyć, że takiej niejednoznaczności nie da się wyeliminować na podstawie kontekstu.

2.2. Oznaczanie klasy gramatycznej

Aby ustalić, do jakiej klasy gramatycznej należy fleksem, którego wykładnik stanowi dane słowo tekstowe, należy najpierw zbadać, czy jest to jednostka odmienna. Wśród klas odmiennych wyróżniamy grupę klas deklinacyjnych, charakteryzujących się odmiennością przez przypadek i klas, które będziemy nazywać koniugacyjnymi, którym przysługuje wartość aspektu.

Klasy deklinacyjne dzielimy na rzeczowniki, przymiotniki, liczebniki i zaimki osobowe.

Fleksemy przymiotnikowe charakteryzują się odmiennością przez liczbę, przypadek i rodzaj. Oprócz tradycyjnych przymiotników (stopniowalnych i niestopniowalnych) do tej klasy należą też liczebniki porządkowe, zaimki przymiotne i fleksem JEDEN.

Za rzeczowniki uznajemy jednostki odmienne przez liczbę i przypadek, ale o ustalonej (dla danego fleksemu) wartości rodzaju. Tak zwane rzeczowniki dwurodzajowe (czy wielorodzajowe), np. *pływak* traktujemy jako osobne jednostki o homonimicznej postaci mianownika liczby pojedynczej, ale różnych wartościach rodzaju. Na przykład uznajemy, że w wypowiedzeniach

(15) Polski pływak pokonał dystans w niewiarygodnym czasie 31,4 sekundy.
(16) W kałuży uwijały się pływaki żółtobrzeżki.
(17) Pływak zanurzył się gwałtownie i po chwili Maurycy wyciągał z wody okazałego jesiotra.

mamy do czynienia z trzema osobnymi fleksemami rzeczownikowymi o wartościach rodzaju odpowiednio: męskiej osobowej (m1), zwierzęcej (m2) i rzeczowej (m3).

W wynikach analizy morfologicznej może wystąpić zbiór kilku rodzajów w opisie formy rzeczownikowej. Należy rozumieć ten zapis jako skrótowy zapis interpretacji słowa jako wykładnika różnych fleksemów.

Liczebniki to fleksemy o ustalonej wartości mnogiej liczby, odmienne przez przypadek i rodzaj. Definicja ta obejmuje tradycyjne liczebniki główne i zbiorowe.

Klasa zaimków obejmuje wyłącznie fleksem ON, odmienny przez liczbę, przypadek, rodzaj i osobę, czyli tzw. zaimki osobowe.

Klasy koniugacyjne Klasy fleksemów odmiennych przez liczbę i osobę obejmują klasę finitywnych form nieprzeszłych, form przyszłych BYĆ, aglutynantów i rozkaźników. Finitywne formy nieprzeszłe to formy tradycyjnie rozumianego czasu teraźniejszego czasowników niedokonanych i przyszłego czasowników dokonanych. Do klasy tej nie należą formy czasu przyszłego czasownika BYĆ (*będę, będziesz, będzie, będziemy, będziecie, będą*) ani jego formy aglutynacyjne (*-(e)m, -(e)ś, -(e)śmy, -(e)ście*). Obecność litery *e* w wymienionych słowach wyróżnia formy wokaliczne aglutynantu. Rozkaźniki mają paradygmat ograniczony do form 2. osoby liczby pojedynczej i 1. i 2. osoby liczby mnogiej.

Pseudoimiesłów, czyli formy „trzeciej osoby czasu przeszłego", wyróżnia się odmiennością przez liczbę i rodzaj. Dla nielicznych leksemów ujawnia się też zróżnicowanie ze względu na aglutynacyjność. Formy aglutynacyjna to np. *gniotł-* w *gniotł-em*, nieaglutynacyjna to *gniótł*.

Cztery z klas koniugacyjnych są nieodmienne. Są to tradycyjnie rozumiane bezokoliczniki, bezosobniki (finitywne formy bezosobowe na *-no, -to*), imiesłowy przysłówkowe współczesne i uprzednie.

Pozostałe trzy klasy, którym przysługuje kategoria aspektu — odsłowniki i imiesłowy przymiotnikowe — wykazują także cechy deklinacyjne, to znaczy odmieniają się jak rzeczowniki lub przymiotniki. Te klasy uznajemy za odmienne przez negację, to znaczy uważamy, że formy *czytanie* i *nieczytanie* należą do tego samego fleksemu.

Zdarzają się oczywiście sytuacje, kiedy dane słowo może być wykładnikiem fleksemów różnych typów. Rozstrzygnięcie takiej niejednoznaczności na podstawie kontekstu należy do anotatorów. Może to dotyczyć na przykład słów, które mogą reprezentować rzeczownik lub gerundium (por. poniżej (18) i (19)), rzeczownik lub przymiotnik (por. (20) i (21)), rzeczownik lub bezokolicznik (por. (22) i (23)) itd.

(18) Zadanie rozwiązano w godzinę.
(19) Zadanie pracy domowej nie było dobrym pomysłem.

(20) Chory był wyczerpany.
(21) Kotek był chory.
(22) Chciał brać znać.
(23) Chciał brać i wiać.

Klasy inne Przysłówki stopniowalne to klasa fleksemów odmiennych wyłącznie przez stopień.

Przyimki to fleksemy nieodmienne mające wymaganie określonego przypadka.

Spójniki to jednostki A, ALBO, ALE, ANI, BĄDŹ, I, JAK, LECZ, LUB, NATOMIAST, NI, ORAZ, TAK, ZARÓWNO, ZAŚ, ABY, ALBOWIEM, AŻ, AŻEBY, BO, BOWIEM, BY, CHOCIAŻ, CHOCIAŻBY, CHOĆ, CHOĆBY, CZY, DOPÓKI, DOPÓTY, GDY, GDYBY, GDYŻ, IŻ, IŻBY, JAK, JAKBY, JAKOBY, JEDNAK, JEŚLI, JEŚLIBY, JEŻELI, JEŻELIBY, KIEDY, NIM, PONIEWAŻ, PÓKI, PÓTY, PRZETO, SKORO, TOTEŻ, WIĘC, WTEDY, WÓWCZAS, ZANIM, ZATEM, ŻE, ŻEBY.

Pozostałe fleksemy nieodmienne uznajemy za partykuło-przysłówki.

2.3. Oznaczanie rodzaju

Oznaczanie rodzaju w wypowiedzeniu należy rozpocząć od ustalenia wartości tej kategorii dla występujących w nim form rzeczownikowych.

Do ustalania rodzaju rzeczowników pomocny może być następujący zestaw kontekstów testowych pochodzący z pracy Saloniego (1976b).

Widzę jednego albo dwóch spośród tych ____, których lubię. m1
Widzę jednego albo dwa spośród tych ____, które lubię. m2
Widzę jeden albo dwa spośród tych ____, które lubię. m3
Widzę jedno albo dwoje spośród tych ____, które lubię. n1
Widzę jedno albo dwa spośród tych ____, które lubię. n2
Widzę jedną albo dwie spośród tych ____, które lubię. f
Widzę jedno albo dwoje spośród tych ____, których lubię. p1
Widzę jedne albo dwoje spośród tych ____, które lubię. p2
Widzę (jedną albo dwie pary) spośród tych ____, które lubię. p3

Ten ostatni kontekst należy rozumieć w ten sposób, że pasują do niego rzeczowniki w ogóle nie dopuszczające połączeń z liczebnikami, których liczebność daje się niekiedy wyrazić przez konstrukcje jak wyżej.

Dla przeważającej większości rzeczowników powinno dać się ustalić dokładnie jedną wartość rodzaju z powyższego zestawienia.

Wyjątek stanowią tzw. zaimki rzeczowne KTO, CO, KTOŚ, COŚ, NIKT, NIC, które trudno dopasować do powyższych kontekstów. Przyjmujemy rozstrzygnięcie arbitralne, mianowicie uznajemy, że KTO, KTOŚ i NIKT są rodzaju m1; CO, COŚ, NIC — rodzaju n2.

Po ustaleniu rodzaju form rzeczownikowych należy przejść do ustalenia rodzaju czasowników, przymiotników i liczebników uzgadniających swój rodzaj z danymi rzeczownikami. Dla takich form należy uznać rodzaj za tożsamy z rodzajem odpowiedniego rzeczownika.

Dla form czasowników, przymiotników i liczebników nie wchodzących w uzgodnienie z żadnym rzeczownikiem zwykle dopuszczalnych będzie kilka wartości kategorii rodzaju.

2.4. Oznaczanie aglutynacyjności

Jako aglutynacyjne oznaczamy formy, które muszą występować jako jedno słowo ortograficzne z formą aglutynantu. Jako nieaglutynacyjne oznaczamy pozostałe formy, a więc takie, które nie muszą występować z aglutynantem. Na przykład dla partykuło-przysłówka DLACZEGO formę *dlaczegó* oznaczamy jako aglutynacyjną, ponieważ musi ona wystąpić w konstrukcji *dlaczegó-ż*. Forma *dlaczego* jest nieaglutynacyjna, bo może wystąpić zarówno z aglutynantem (*dlaczego-ś*), jak i bez niego (*dlaczego*).

Dla pseudoimiesłowów sytuacja jest prostsza: formy aglutynacyjne (np. *gniotł*) występują zawsze z aglutynantem, formy nieaglutynacyjne (np. *gniótł*) — zawsze bez aglutynantu.

3. Porównanie z innymi zestawami znaczników morfosyntaktycznych

3.1. Tagsety

W niniejszym punkcie porównamy zestaw znaczników morfosyntaktycznych przedstawiony powyżej z kilkoma tagsetami zaproponowanymi wcześniej dla języka polskiego i innych języków słowiańskich.[2] Skupimy się na następujących tagsetach:

- Zaproponowane na gruncie języka polskiego:
  - tagset SFPW (nowy tagset korpusu *Słownika frekwencyjnego polszczyzny współczesnej*; Kurcz *et al.*, 1990);
  - tagsety używane przez analizatory morfologiczne:
    - LEM;
    - SAM (Szafran, 1996);
  - XeLDA.
- Zaproponowane dla innych języków słowiańskich:
  - tagset Czeskiego Korpusu Narodowego (CzKN);
  - Multext-East, dla czeskiego i słoweńskiego, a także dla rosyjskiego (SFB 441, Tybinga).

Nie uwzględniamy w niniejszym porównaniu wielu nieudokumentowanych tagsetów, np. tagsetów używanych przez analizatory morfologiczne POMOR, Amor czy NeurosoftGram. Krótka charakterystyka i porównanie tych i innych analizatorów znajduje się w pracy Hajnicz i Kupść, 2001.

3.2. Kategorie morfosyntaktyczne

3.2.1. Tagset IPI PAN

Zacznijmy od sumarycznego przedstawienia własności zestawu znaczników opisanego w pierwszej części niniejszego raportu. Zgodnie z tabelą przedstawioną w p. 1.6, wyróżniamy 21 klas gramatycznych, od tak licznych jak rzeczownik czy czasownik nieprzeszły, do tak wąskich jak czasownik przyszły *być*, zawierający formy tylko jednego leksemu.

Tagset IPI PAN jest tagsetem pozycyjnym, co oznacza w istocie, że możliwe wartości kategorii morfosyntaktycznych nie zależą od poszczególnych klas gramatycznych. Na przykład kategoria rodzaju ma te same wartości dla form rzeczownikowych, przymiotnikowych, czasownikowych i różnych form odczasownikowych. Nie jest to jedyne możliwe podejście: można twierdzić, iż np. czasowniki nie odróżniają trzech rodzajów męskich, a zatem nie należy przypisywać im kategorii m1, m2 i m3, właściwych dla form przymiotnikowych, a jedynie jednoznaczną kategorię rodzaju męskiego. Takie podejście przyjęte zostało np. w analizatorze morfologicznym SAM (por. 3.2.4).

W p. 1.4 zaproponowaliśmy 13 kategorii gramatycznych szczególnie istotnych, naszym zdaniem, do opisu form wyrazowych języka polskiego. Zestaw ten, jak i możliwe wartości poszczególnych kategorii, oparte są w dużej mierze na pracach Zygmunta Saloniego (Saloni, 1976a,b, 1977, 1988).

Traktując klasę gramatyczną i kategorię gramatyczną łącznie jako **atrybuty**, powiemy, że tagset IPI PAN zawiera 14 atrybutów, którymi opisywane są poszczególne formy wyrazowe: klasę gramatyczną, o 21 możliwych wartościach, oraz 13 kategorii gramatycznych, posiadających 2 (liczba, aspekt, negacja, deprecjatywność, akcentowość, poprzyimkowość, akomodacyjność, aglutynacyjność, wokaliczność), 3 (osoba, stopień), 7 (przypadek) lub 9 (rodzaj) możliwych wartości.

3.2.2. Tagset SFPW

Przez tagset SFPW można rozumieć dwa różne acz silnie ze sobą związane systemy anotacji morfosyntaktycznej. Po pierwsze, jest to tagset wykorzystany do ręcznej anotacji tzw. korpusu *Słownika frekwencyjnego polszczyzny współczesnej* (SFPW; Kurcz *et al.*, 1990, 1974). Tagset ten nie został

[2] Świadomie zapożyczamy tutaj angielski termin *tagset* jako bardziej zwięzły niż *zestaw znaczników morfosyntaktycznych*. Choć terminu *tagset* można używać także w odniesieniu do nie-morfosyntaktycznego zestawu znaczników, takie jego użycie jest najczęściej spotykane w literaturze anglosaskiej, i tak też będziemy rozumieć ten termin w dalszej części niniejszego raportu.

zaprojektowany jako tagset pozycyjny, ale stosunkowo łatwo daje się przekształcić do takiej postaci. Zakłada on tradycyjne części mowy (rzeczownik, czasownik, przymiotnik, liczebnik, zaimek, przyimek, wykrzyknik, partykuła, spójnik) i tradycyjne zestawy wartości kategorii przypadka i liczby, a także pozwala na wyróżnienie kilku form i funkcji czasowników, np. czasownik niezwrotny w bezokoliczniku w funkcji formy czasu przyszłego, czasownik zwrotny jako składnik formy czasu przyszłego zakończonych na *-ł* itp.

Tagset ten nie operował kategorią rodzaju, osoby i innymi kategoriami morfologicznymi przyjmowanymi we współczesnym językoznawstwie. Dlatego też opracowane zostało (przez Katarzynę Głowińską) pozycyjne rozszerzenie pierwotnego tagsetu korpusu *Słownika frekwencyjnego polszczyzny współczesnej*. Odtąd przez *tagset SFPW* będziemy rozumieli właśnie tę rozszerzoną wersję tagsetu *Słownika frekwencyjnego*. Tagset ten nie przez przypadek posiada strukturę podobną do tagsetu IPI PAN: tagset SFPW stanowił punkt wyjścia dla naszych prac. W tagsecie tym występuje 14 atrybutów: część mowy (klasa gramatyczna) i 13 kategorii morfosyntaktycznych.

Atrybut **klasa gramatyczna** (część mowy) posiada 11 możliwych wartości: czasownik, rzeczownik, przymiotnik, liczebnik, zaimek, przysłówek, przyimek, spójnik, wykrzyknik, partykuła, kod nieznany. Ten ostatni nie jest oczywiście częścią mowy w tradycyjnym znaczeniu tego terminu, lecz oznacza formę nie zaklasyfikowaną jako żadna z tradycyjnych części mowy.

Kategorie gramatyczne **liczby**, **przypadka**, **stopnia**, **aspektu**, **poprzyimkowości** i **akcentowości** są takie same jak w tagsecie IPI PAN.

Kategoria **rodzaju** posiada 9 wartości: męski, męskoosobowy, męskozwierzęcy, męskorzeczowy, żeński, nijaki, męskoosobowy, niemęskoosobowy, *plurale tantum*. Różnią się one od wartości rodzaju w tagsecie IPI PAN dwojako. Po pierwsze, nie uwzględniają one podziału rodzajów nijakiego na n1 i n2 oraz *plurale tantum* na p1, p2 i p3, który przyjęliśmy za Salonim na podstawie łączliwości z formami liczebników. Po drugie, wartości te mają tę nietradycyjną cechę, że zawierają w istocie informację zarówno o rodzaju, jak i o liczbie (np. *biały (stół)* — rodzaj męski albo męskorzeczowy, *białe (stoły)* — rodzaj niemęskoosobowy).

Kategoria **osoby** posiada trzy tradycyjne wartości (pierwsza, druga i trzecia) oraz cztery wartości odpowiadające czterem różnym klasom gramatycznym w naszym podejściu: bezokolicznik, bezosobnik, imiesłów przysłówkowy uprzedni i imiesłów przysłówkowy współczesny. Wydaje się, że połączenie tych form czasownikowych w jedną klasę gramatyczną miałoby sens o tyle, że wszystkie cztery klasy form są nieodmienne (zakładając nieobecność kategorii morfologicznej trybu) i posiadają kategorię aspektu. Przyjmując jednak istnienie kategorii trybu, należałoby chyba uznać, że bezosobnik i bezokolicznik, które odmieniają się przez tryb (*kupiono by*, *kupić by*), należą do innej klasy niż imiesłowy przysłówkowe, które przez tryb się nie odmieniają (**kupując by*, **kupiwszy by*).

Pięć pozostałych klas gramatycznych postulowanych w tagsecie SFPW nie ma swojego odpowiednika w tagsecie IPI PAN. Są to: **czas** (teraźniejszy, przeszły, przyszły złożony), **tryb** (oznajmujący, przypuszczający, rozkazujący), **strona** (czynna, bierna, zwrotna), **oznaczenie dodatkowe form czasownikowych** (bezokolicznik jako forma składowa czasu przyszłego, forma na *-ł* jako składowa form czasu przyszłego itp.) i **nazwy własne** (nazwa pospolita, nazwa własna, skrótowiec).

Różnice te wynikają przede wszystkim z założenia metodologicznego, które przyjęliśmy powyżej, a mianowicie, że oznaczane morfosyntaktycznie będą wyłącznie ciągi nie przekraczające granicy słowa ortograficznego. Na przykład, atrybut **nazwy własne** jest atrybutem całego ciągu *Nowy Jork*, a nie tylko wyrazów *Nowy* i *Jork*. Podobnie, nie jest jasne, że bezokolicznik *podróżować* w konstrukcji *będą podróżować* posiada kategorię **czasu** o wartości przyszły złożony — kategoria ta przysługuje raczej całej konstrukcji. Podobne argumenty można wysunąć dla kategorii **trybu** (np. *przyszedł* + *-em*, *przyszedł* + *-bym*, *niech* + *przyjdzie*), **strony** (*mył* + *się*) czy **oznaczenie dodatkowe form czasownikowych**, które wskazuje na funkcję danej formy w konstrukcji przekraczającej ramy słowa ortograficznego. We wszystkich tych wypadkach, odpowiednie informacje mogą być uzyskane na podstawie danych słownikowych (np. zwrotność) i odpowiedniej analizy składniowej.

Zasygnalizujmy jeszcze istnienie kategorii morfologicznych, które nie zostały uwzględnione w tagsecie SFPW: negacja, deprecjatywność, akomodacyjność, aglutynacyjność, wokaliczność.

Niektóre z wyżej wymienionych różnic wydają się wynikać z innych założeń, na których oparte

zostały oba tagsety (przede wszystkim założenie dotyczące analizy segmentów zawierających odstępy i segmentów krótszych niż słowo rozumiane jako ciąg liter od spacji/znaku interpunkcyjnego do spacji/znaku interpunkcyjnego), inne wynikają z nieuwzględnienia w tagsecie SFPW kategorii morfologicznych dotyczących tylko niewielkich klas form wyrazowych (np. akcentowość, aglutynacyjność). Istotną różnicą pomiędzy tagsetami SFPW i IPI PAN jest repertuar wartości rodzajowych.

3.2.3. LEM

Tagset analizatora morfologicznego LEM wyszczególnia następujących 29 części mowy: rzeczownik, zaimek rzeczowny, odsłownik, czasownik, *być*, nieodmienna forma czasownika, czasownik modalny, przymiotnik, zaimek przymiotny, przymiotnik liczebnikowy (*dwojaki*, *pojedynczy*), cztery standardowe rodzaje imiesłowów (2 przymiotnikowe i 2 przysłówkowe), tzw. imiesłów przeszły, przysłówek, zaimek przysłowny, przysłówek liczebnikowy, przyimek, spójnik, wykrzyknik, zawołanie (oddzielnie od wykrzyknika), onomatopeja, partykuła, liczebnik główny, liczebnik porządkowy, liczebnik zbiorowy, liczebnik partytywny oraz kategorię oznaczoną jako „PPRO (prep-noun-pronoun)". Nie jest do końca jasne, na podstawie jakich kryteriów klasy te zostały wyodrębnione; wydaje się, że zastosowane zostały tu zarówno kryteria morfologiczne i składniowe, jak i semantyczne (np. wyróżnienie liczebnikowych podklas przymiotników), pragmatyczne (zawołanie i wykrzyknik) oraz diachroniczne (onomatopeja).

Znacznie mniejszy jest natomiast zakres kategorii morfologicznych. Standardowy jest zakres wartości kategorii **osoby**, **liczby przypadka**, **stopnia** i **aspektu**. Wartościami kategorii **rodzaju** jest pięć rodzajów zaproponowanych przez Mańczaka (1956), tj. męskoosobowy, męskozwierzęcy, męskorzeczowy, nijaki i żeński. Trzy pozostałe kategorie to **forma czasownika** (osobowa, bezosobowa i bezokolicznikowa), **tryb** (oznajmujący, warunkowy i rozkazujący) oraz **czas gramatyczny** (przeszły, teraźniejszy i przyszły)[3].

Trzy tagsety omówione powyżej są tagsetami pozycyjnymi. Następne dwa tagsety zaproponowane dla języka polskiego nie posiadają tej własności.

3.2.4. SAM

Zestaw oznaczeń morfosyntaktycznych wykorzystywany przez analizator morfologiczny SAM (Szafran, 1996[4]) oparty jest na zestawie znaczników zaproponowanym w *Schematycznym indeksie a tergo polskich form wyrazowych* Jana Tokarskiego w opracowaniu Zygmunta Saloniego (Tokarski, 1993). Zestaw ten ma charakter w zasadzie czysto morfologiczny, a nie morfosyntaktyczny, jak to miało miejsce w wypadku tagsetów omawianych powyżej, i stąd w dużej mierze wynikają różnice pomiędzy tagsetem analizatora SAM i innymi tagsetami rozważanymi w niniejszym raporcie.

Program SAM operuje zestawem dziewięciu klas leksemów (klas gramatycznych): leksemy nieodmienne, rzeczownikowe, czasownikowe, przymiotnikowe, liczebnikowe, zaimkowe (zaimki rzeczowne), pseudoprzymiotnikowe, przysłówkowe i przyimkowe odmienne. Oznaczenia morfosyntaktyczne nadawane poszczególnym formom wyrazowym zależą od tego, do jakiej klasy forma ta została zaklasyfikowana.[5]

Leksemy **nieodmienne** nie posiadają żadnej dodatkowej charakterystyki morfosyntaktycznej.

Formy leksemów **rzeczownikowych** posiadają informację słownikową o rodzaju morfologicznym (istotnym z punktu widzenia odmiany) i o rodzaju gramatycznym (składniowym). Wartościami kategorii rodzaju może być m (męski), f (żeński), n (nijaki) i blp (*plurale tantum*). Oprócz rodzaju, formy rzeczownikowe oznaczone są standardowo rozumianymi kategoriami morfologicznymi liczby i przypadka.

Formy leksemów **czasownikowych** posiadają informację słownikową o klasie koniugacyjnej. Podstawową niesłownikową informacją morfologiczną jest forma czasownika; SAM rozróżnia następujące wartości tej kategorii: forma czasu teraźniejszego, forma czasu przeszłego, bezokolicznik, bezosobnik,

[3] Jak zauważają Hajnicz i Kupść (2001), w praktyce różnica pomiędzy czasem przyszłym i teraźniejszym nie jest wykorzystywana jako redundantna w stosunku do kategorii aspektu.

[4] Opis poniższy oparty jest na raporcie Szafran, 1996 i nie uwzględnia późniejszych zmian dokonanych w programie SAM.

[5] Oznaczenia przyjęte w niniejszym opisie tagsetu SAM nie odzwierciedlają faktycznych oznaczeń produkowanych przez ten program, opartych na różnych wartościach domyślnych i skrótach notacyjnych, por. Szafran, 1996.

forma trybu rozkazującego, cztery standardowe formy imiesłowowe, tzw. imiesłów przymiotnikowy przeszły oraz odsłownik (gerundium). Dalsze oznaczenia zależą od formy czasownika:

- w wypadku bezokolicznika, bezosobnika i imiesłowów przysłówkowych, dalsze oznaczenia nie istnieją (są to formy nieodmienne);
- w wypadku odsłownika, dalsze oznaczenia są takie, jak dla form rzeczownikowych;
- w wypadku trzech imiesłowów przymiotnikowych, dalsze oznaczenia są takie, jak dla form przymiotnikowych;
- w wypadku formy trybu rozkazującego, dodawane są odpowiednio zawężone informacje o liczbie i osobie;
- formy czasu teraźniejszego i przeszłego posiadają informację o liczbie i osobie;
- w wypadku czasu teraźniejszego, dodatkowo odróżnia się formy leksemu BYĆ (np. *jest* i *będzie*);
- w wypadku czasu przeszłego, wyróżnia się także trzy rodzaje dla liczby pojedynczej (męski, żeński i nijaki) i dwa dla liczby mnogiej (męskoosobowy i niemęskoosobowy).

Formy leksemów **przymiotnikowych** oznaczone są standardowymi kategoriami liczby, stopnia (równy lub wyższy) i przypadka (bez wołacza), kategorią rodzaju o wartościach m1, m2, m3, n, ż, oraz kategorią deprecjatywności.

Leksemy **liczebnikowe** podzielone zostały na cztery podklasy:

1. liczebniki główne DWA, OBA, OBYDWA, TRZY, CZTERY;
2. pozostałe liczebniki główne,
3. liczebniki zbiorowe,
4. liczebniki ułamkowe.

Szczegółowe omówienie charakterystyki poszczególnych klas liczebnikowych można znaleźć w pracy Szafran, 1996.

Leksemy **zaimkowe** podzielone zostały na trzy podklasy:

1. zaimki o jednej mające jedynie kategorię fleksyjną przypadka (KTO, MY),
2. zaimki o dwóch kategoriach fleksyjnych, tj. przypadka i akcentowalności (JA, TY, SIĘ),
3. zaimki o pięciu kategoriach fleksyjnych, tj. przypadka, liczby, rodzaju, akcentowalności i poprzyimkowości (ON).

Formy leksemów **pseudoprzymiotnikowych** (*godzien*, *zdrów*) oznaczone są liczbą i rodzajem rozumianym jak w wypadku form czasu przeszłego czasowników.

Formy leksemów **przysłówkowych** oznaczone są kategorią stopnia (równego lub wyższego).

Formy leksemów **przyimkowych odmiennych** oznaczone są kategorią wokaliczności.

Cechą odróżniającą opisany powyżej system oznaczeń gramatycznych od tagsetów opisanych wcześniej jest to, że dotyczy on wyłącznie informacji morfologicznych. Na przykład różne są zbiory wartości kategorii rodzaju dla form rzeczownikowych, przymiotnikowych i czasownikowych. A zatem, *na poziomie takich oznaczeń gramatycznych*, nie można ująć faktu, że wszystkie trzy formy wyrazowe w zdaniu *Biały kot usnął* posiadają tę samą wartość kategorii rodzaju (tj., że ma tu miejsce uzgodnienie rodzaju).

3.2.5. XeLDA

Tagset opracowany przez firmę Xerox w ramach projektu XeLDA składa się z 85 binarnych kategorii morfologicznych (np. +Noun, +Neut, +Sg, +Poss itp.) i z 12 kategorii reprezentujących formy homonimiczne (np. +GenAccLoc przy słowie *wpadowych*).[6] Z powodu niedostępności dokumentacji oraz znacznej odmienności tego systemu oznaczeń morfoskładniowych od innych systemów opisanych w niniejszym raporcie, poprzestaniemy na podaniu kilku przykładów.

znacznik	**opis**	**przykłady**	
		wyraz	**analiza**
+ALoc	location (adverb)	tędy	tędy+Adv+ALoc
+Adv	adverb	dalej	daleki+Adv+Comp

cd. na następnej stronie

[6] `http://www.xrce.xerox.com/research/mltt/demos/doc/mor-pol-1.html`.

cd. z poprzedniej strony

znacznik	**opis**	**przykłady**	
		wyraz	**analiza**
+Card	cardinal (digit)	123	123+Num+Dig+Card
+Comma	punctuation comma	,	,+Punct+Comma
+Conj	conjunction	niźli	niźli+Conj
+Dat	dative case	klauzom	klauza+Noun+Fem+Pl+Dat
+Date	date expression (digit)	9.9.1999	9.9.1999+Num+Dig+Date
+Depr	depreciative	rabiny	rabin+Noun+M1+Pl+Depr
+Fem	feminine gender	zgodną	zgodny+Adj+Fem+Sg+Ins
+Gerund	verb gerund	nużąc	nużyć+Verb+Imperf+Gerund+Pres
+Infinit	verb infinitive	dodać	dodać+Verb+Perf+Infinit
+Inv	invariant	jego	jego+Pron+Poss+3P+Sg+MN+Inv
+M3	masculine, inanimate	sklepem	sklep+Noun+M3+Sg+Ins
+Noun	common noun	muzeum	muzeum+Noun+Neut+Sg+InvCase
+NumOrd	ordinal number	setną	setny+NumOrd+Fem+Sg+Acc
+Partcl	particle	samyż	sam+Pron+M3+Sg+Acc+Partcl
+Phras	phraseology	górsku	górsku+Phras
+Pl	plural	one	one+Pron+Pers+3P+Pl+M23+Nom
+Poss	possessive (pronoun)	mym	mój+Pron+Poss+MN+Sg+InsLoc
+Prep	preposition	od	od+Prep+Gen
+Punct	punctuation	?	?+Punct+Final
+QVerb	quasi verb	braknie	braknąć+QVerb+Perf+Ind+Pres
+Refl	reflexive (pronoun)	się	się+Pron+Refl+Acc
+Spec	special symbol	@	@+Spec
+1P	first person	my	my+Pron+Pers+1P+Pl+MFN+Nom
+GenAcc	genitive or accusative	szóstego	szósty+NumOrd+M12+Sg+GenAcc
+MFN	any gender	my	my+Pron+Pers+1P+Pl+MFN+Nom

3.2.6. Tagset CzKN

Tagset Czeskiego Korpusu Narodowego[7] jest tagsetem pozycyjnym zawierającym dwa atrybuty odpowiadające częściom mowy, tj. POS, o wartościach: rzeczownik, przymiotnik, liczebnik, przysłówek, wykrzyknik, spójnik, zaimek, przyimek, partykuła, czasownik, nieznany, interpunkcja, oraz SUBPOS zawierający 74 szczegółowe podkategorie części mowy takie jak typy zaimka (względny/pytajny, poprzyimkowy, zwrotny nieakcentowany, zwrotny akcentowany, zwrotny dzierżawczy itd., w sumie 20 podkategorii), typy liczebnika (np. nieokreślony, większy od 4 przymiotnikowy, większy od 4 rzeczownikowy itd.; w sumie 17) itp.

Oprócz tych dwóch atrybutów, tagset wyróżnia 11 kategorii morfologicznych: rodzaj (zestaw o charakterystyce zbliżonej do wartości kategorii rodzaju w tagsecie SFPW), liczba, przypadek, osoba, stopień, czas, negacja, rodzaj posesora, liczba posesora, wariant/styl. Możliwą wartością każdej kategorii jest „nie dotyczy", a prawie każdej — „nieznany".

3.2.7. Multext-East

Celem projektu Multext-East, finansowanego w ramach programu *Copernicus* Wspólnoty Europejskiej (1995–1997), było stworzenie anotowanego korpusu wielojęzycznego oraz „baz leksykalnych" dla siedmiu języków: bułgarskiego, czeskiego, estońskiego, węgierskiego, rumuńskiego, słoweńskiego i angielskiego.[8] W szczególności stworzony został zestaw atrybutów morfoskładniowych i ich możliwych wartości wystarczający do skonstruowania tagsetu dla każdego z tych języków. Choć tagsety te nie mają postaci tagsetów pozycyjnych, gdyż zbiory kategorii morfoskładniowych zależą od poszczególnych

[7] `http://ucnk.ff.cuni.cz/`.

[8] `http://nl.ijs.si/ME/`.

klas gramatycznych (np. przypadek jest odpowiedni dla rzeczownika, ale nie dla czasownika), można je jednak łatwo przekształcić do pełnej postaci pozycyjnej.

Tagsety dla wspomnianych wyżej języków przedstawione są w raporcie Erjavec, 2001. Dodatkowo, w ramach niemieckiego projektu SFB 441, opracowany został na Universität Tübingen podobny tagset dla języka rosyjskiego.[9] W niniejszym raporcie skupimy się na dwu z powyżej wymienionych tagsetów, a mianowicie na tagsecie dla języka czeskiego stworzonym przez Vladimira Petkeviča, oraz na tagsecie dla języka rosyjskiego autorstwa Rolanda Meyera.

Specyfikacja Multext-East przewiduje 14 klas gramatycznych (części mowy), spośród których tagsety dla języka czeskiego i rosyjskiego wykorzystują 12: rzeczownik (N), czasownik (V), przymiotnik (A), zaimek (P), przysłówek (R), przyimek (S), spójnik (C), liczebnik (M), wykrzyknik (I), skrót (X), partykuła (Q) oraz element obcy (*residual*; Y).

Oprócz klasy gramatycznej, tagset dla języka czeskiego definiuje następujące atrybuty:

atrybut	**liczba wartości**	**zdefiniowany dla**
typ	2,4,5,9,1,1,2,4,0,0,0,0	NVAPRSCMIXYQ
rodzaj	3	NVAPM
liczba	3	NVAPCM
osoba	3	VPC
przypadek	7	NAPSM
żywotność	2	NVAPM
forma czas.	6	V
czas	3	V
strona	2	V
negacja	2	V
akcentowość	2	P
klityka *s*	2	VP
stopień	3	AR
złożoność	2	AS
liczba posesora	2	P
rodzaj posesora	3	P
typ referencji	2	P
typ składniowy	2	P
forma	3	M
klasa	8	M

Zgodnie z tabelką powyżej, tagset Multext-East dla języka czeskiego obejmuje 20 kategorii. Wartości prawie wszystkich z nich nie zależą w zasadzie od części mowy (choć zbiór tych wartości może być ograniczony dla niektórych części mowy, np. wśród wartości przypadka dla klasy gramatycznej P nie ma wołacza), z wyjątkiem „kategorii" *typ*, której wartości i ich liczba silnie zależą od klasy gramatycznej (2 możliwe wartości dla rzeczowników, 4 — dla czasowników, 5 — dla przymiotników itd.). W zasadzie należałoby tę „kategorię" rozbić na kilka atrybutów takich jak *typ rzeczownika* (o 2 możliwych wartościach), *typ czasownika* (o 4 możliwych wartościach), *typ przymiotnika* (o 5 możliwych wartościach) itd., dla uproszczenia założymy jednak, że jest to jeden atrybut o $2+4+5+\ldots = 28$ wartościach.[10]

Podobną tabelkę można skonstruować dla tagsetu języka rosyjskiego. W tabelce podanej poniżej znaczenie atrybutów *typ* i *forma* zależy od danej klasy gramatycznej.

[9] `http://www.sfb441.uni-tuebingen.de/c1/tagset.html`.

[10] W tabelce w p. 3.3 wyodrębnimy jednak dla ułatwienia porównania „typ rzeczownikowy", który w istotcie jest kategorią pospolitości o dwóch możliwych wartościach (nazwa pospolita i nazwa własna). Dlatego też, w tabelce tej liczba wartości atrybutu *typ* będzie odpowiednio mniejsza.

atrybut	liczba wartości	zdefiniowany dla
typ	2,3,2,9,1,2,2,3,0,0,0,0	NVAPRSCMIXYQ
rodzaj	4	NVAPM
liczba	2	NVAPM
osoba	3	VP
przypadek	7	NAPSM
podprzypadek	2	N
żywotność	2	NVAPM
forma czas.	6	V
czas	3	V
strona	3	V
aspekt	3	V
stopień	2	AR
złożoność	2	A
liczba posesora	2	P
typ referencji	2	P
typ składniowy	2	P
forma	2,3	VM
klasa	8	M

3.2.8. Podsumowanie

Tabelka poniżej podsumowuje podstawowe własności tagsetów omawianych w poprzednich punktach.

tagset	poz.?	# atr.	przykładowe atrybuty (oprócz POS)
IPI PAN	+	14	przypadek, deprecjatywność, negacja itp.
SFPW	+	14	przypadek, strona, tryb, czas itp.
LEM	−/+	10	przypadek, aspekt, czas itp.
SAM	−	$\leqslant 14$	do 13 kat. w zależności od POS
XeLDA	−	wiele	wiele kat. binarnych (np. +Fem, +Gen itp.)
CzKN	+	13	subPOS, przypadek, rodzaj itp.
Multext-East-cz.	−/+	21	typ, przypadek, strona, żywotność itp.
Multext-East-r.	−/+	19	typ, (pod)przypadek, strona, aspekt itp.

Spośród 8 tagsetów branych przez nas pod uwagę (łącznie z zestawem znaczników morfosyntaktycznych zaproponowanym w p. 1), 3 są tagsetami pozycyjnymi (IPI PAN, SFPW i CzKN), dalsze 3 można stosunkowo łatwo przekształcić do postaci tagsetu pozycyjnego (LEM, Multext-East-cz., Multext-East-r.), zaś w wypadku dwóch pozostałych takie przekształcenie byłoby znacznie mniej trywialne (SAM, XeLDA). Większość tych tagsetów wprowadza od 10 do 14 atrybutów (klas gramatycznych i kategorii morfoskładniowych), tagsety Multext-East przewidują około 20 takich atrybutów, zaś tagset XeLDA wprowadza 86 różnorodnych „kategorii" binarnych i dodatkowo 12 kategorii oddających różne synkretyzmy.

Wobec znacznych różnic metodologicznych leżących u podstaw tagsetów przedstawionych powyżej, dalsze ich porównanie wydaje się bezcelowe. W następnym punkcie scharakteryzujemy jedynie podobieństwa i różnice pomiędzy tagsetami pozycyjnymi, włączając do tego porównania także te tagsety, które dają się łatwo sprowadzić do postaci tagsetów pozycyjnych.

3.3. Porównanie tagsetów pozycyjnych

Poniższa tabela przedstawia dane dotyczące repertuaru kategorii morfoskładniowych wprowadzanych przez tagsety IPI PAN, SFPW, LEM, CzKN, Multext-East-cz. i Multext-East-r., oraz liczbę wartości tych kategorii w różnych tagsetach.

kategoria	liczba wartości					
	IPI PAN	**SFPW**	**LEM**	**CzKN**	**MtE-cz.**	**MtE-r.**
POS	21	11	29	12	12	12
subPOS / typ	–	–	–	74	26	22
przypadek	7	7	7	8	7	7
podprzypadek	–	–	–	–	–	2
liczba	2	2	2	5	3	2
liczba posesora	–	–	–	2	2	2
rodzaj	9	9	5	10	3	4
rodzaj posesora	–	–	–	4	3	–
stopień	3	3	3	3	3	2
osoba	3	7[a]	3	4	3	3
negacja	2	–	–	2	2	–
formy czasownikowe	12[b]	7	3	–	6	6
czas	–	3	3	5	3	3
tryb	–	3	3	–	–[c]	–
aspekt	2	2	2	–	–	3
strona	–	3	–	2	2	3
żywotność	–	–	–	–	2	2
pospolitość[d]	–	3	–	–	2	2
wariant/styl	–	–	–	9	–	–
deprecjatywność	2	–	–	–	–	–
akcentowość	2	2	–	–	2	–
poprzyimkowość	2	2	–	–	–	–
aglutynacyjność	2	–	–	–	–	–
wokaliczność	2	–	–	–	–	–
						etc.

[a] Jest to łączna liczba wartości kategorii osoby (3) i form bezosobowych czasownika (4).

[b] Tagset IPI PAN nie przewiduje atrybutu *formy czasownikowe*; podana tu liczba dotyczy liczebności klas czasownikowych i odczasowników wśród ogółu 21 klas gramatycznych.

[c] W wypadku tagsetów Multext-East, informacja o trybie zawarta jest w wartościach atrybutu *formy czasownikowe*.

[d] Por. przypis 10.

Należy zauważyć, że zestaw znaczników morfosyntaktycznych IPI PAN nie odbiega znacząco od innych tagsetów pod względem liczby klas gramatycznych i kategorii morfosyntaktycznych. Zawiera on większą liczbę klas gramatycznych (*POS*) niż tagset SFPW, ale mniejszą, niż tagsety LEM, Czeskiego Korpusu Narodowego, czy tagsety Multext-East (biorąc pod uwagę liczność kategorii *subPOS / typ*). Poza tym informacje zakodowane w innych tagsetach w wartościach atrybutu *formy czasownikowe* obecne są w tagsecie IPI PAN w postaci dodatkowych klas gramatycznych takich jak bezokolicznik, bezosobnik itp. Także pod względem liczby kategorii morfosyntaktycznych i liczby wartości poszczególnych kategorii tagset przyjęty w niniejszym raporcie nie odbiega istotnie od innych tagsetów.

Podstawowa różnica między tagsetem IPI PAN, a innymi tagsetami rozważanymi powyżej jest jakościowa, a nie ilościowa, i polega na pryncypialnym potraktowaniu pojęcia *klasa gramatyczna* jako zbliżonego do pojęcia *fleksemu* Bienia (1991). Dzięki temu, że klasy gramatyczne definiujemy morfoskładniowo, a nie semantycznie (jako leksemy), jak ma to miejsce w wypadku innych tagsetów, możemy w sposób konsekwentny mówić o kategoriach gramatycznych poszczególnych klas (np. czasowniki nieprzeszłe, ale nie bezokoliczniki, odmieniają się przez liczbę).

Literatura

P. Bański (2001) *The proposed annotation scheme for the IPI PAN corpus*, Prace IPI PAN 936, Instytut Podstaw Informatyki, Polska Akademia Nauk.

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.

Ł. Dębowski (2001) *Tagowanie i dezambiguacja morfologiczna*, Prace IPI PAN 934, Instytut Podstaw Informatyki, Polska Akademia Nauk.

T. Erjavec (red.) (2001) *Specifications and Notation for MULTEXT-East Lexicon Encoding*, Ljubljana.

E. Hajnicz, A. Kupść (2001) *Przegląd analizatorów morfologicznych dla języka polskiego*, Prace IPI PAN 937, Instytut Podstaw Informatyki, Polska Akademia Nauk.

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

I. Kurcz, A. Lewicki, J. Sambor, J. Woronczak (1974) *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Maszynopis, Uniwersytet Warszawski.

W. Mańczak (1956) *Ile jest rodzajów w polskim?*, Język Polski, t. XXXVI, nr 2, s. 116–121.

Z. Saloni (1976a) *Cechy składniowe polskiego czasownika*, Ossolineum, Wrocław.

— (1976b) *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*, Studia Gramatyczne, t. I, s. 145–173.

— (1988) *O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie*, Biuletyn Polskiego Towarzystwa Językoznawczego, t. XLI, s. 155–166.

Z. Saloni, M. Świdziński (2001) *Składnia współczesnego języka polskiego*, PWN, Warszawa, wyd. piąte.

K. Szafran (1996) *Analizator morfologiczny SAM-95: opis użytkowy*, TR 96-05 (226), Instytut Informatyki Uniwersytetu Warszawskiego, Warszawa.

J. Tokarski (1993) *Schematyczny indeks* a tergo *polskich form wyrazowych.* Opracowanie i redakcja Zygmunt Saloni, Wydawnictwo Naukowe PWN, Warszawa.

Pracę zgłosił: **prof. dr hab. Leonard Bolc**

Adresy autorów: **Adam Przepiórkowski, Marcin Woliński**
Instytut Podstaw Informatyki PAN
ul. Ordona 21
01-237 Warszawa
e-mail: {adamp,wolinski}@ipipan.waw.pl

Symbole klasyfikacji rzeczowej: I.2.7

Na prawach rękopisu
Printed as a manuscript

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