UNIWERSYTET WARSZAWSKI
WYDZIAŁ NEOFILOLOGII
KATEDRA LINGWISTYKI FORMALNEJ

Maciej Ogrodniczuk

Weryfikacja korpusu wypowiedników polskich (z wykorzystaniem gramatyki formalnej Świdzińskiego)

Rozprawa doktorska
przygotowana pod kierunkiem dr. hab. Janusza S. Bienia, prof. UW

Warszawa 2006

Spis treści

Wstęp

Cel, metoda i zakres pracy

Głównym celem pracy była weryfikacja danych tzw. korpusu wypowiedników polskich[1] (czyli zdań i oznajmień) stworzonego przez Marka Świdzińskiego i ręcznie anotowanego[2] strukturalnymi oznaczeniami gramatycznymi. Za narzędzie posłużyła do tego zadania gramatyka formalna języka polskiego — również autorstwa Świdzińskiego i wykorzystujący ją, niedawno powstały analizator składniowy Świgra Marcina Wolińskiego. Opisane niżej eksperymenty można traktować jako odpowiedź na sformułowaną w końcowym rozdziale książki [Świdziński, 1996] zachętę do *rozmaitych przedsięwzięć informatycznych*[3] na bazie omawianego zestawu danych.

Metoda weryfikacji — oprócz analizy danych zastanych — polegała w znacznej mierze na konfrontacji dwóch opisów danego zbioru wypowiedzeń polskich: oryginalnych strukturalizacji dodanych ręcznie zgodnie z daną gramatyką języka polskiego oraz rozbiorów dokonanych automatycznie przy użyciu tej samej gramatyki. Zadanie to okazało się być pierwszą próbą automatycznej weryfikacji korpusowej gramatyki Świdzińskiego.

Wynikiem pracy jest zoptymalizowana wersja samej gramatyki przygotowana na potrzeby jak najszerszej akceptacji wypowiedników zdaniowych oraz równoległy XML-owy korpus analiz strukturalnych. Do zakresu pracy nie należy natomiast modyfikacja gramatyki w stopniu wykraczającym poza wymagania stawiane przez korpus wypowiedników ani przetwarzanie za jej pomocą obszerniejszych zbiorów tekstów, co również mogłoby dostarczyć materiału do ciekawych badań.

Realizacja pracy nie byłaby możliwa bez udostępnienia przez Świdzińskiego zainteresowanym osobom zarówno samego korpusu, jak i omawiającej go książki w wersji elektronicznej, za co należą mu się serdeczne podziękowania.

[1] Pojęciem *korpusu* posługuję się w tekście zakładając, że nie wymaga ono chyba szerszej definicji — rozumiem pod nim utworzony na potrzeby badań zbiór tekstów, być może opatrzonych pewną dodatkową informacją jakościową lub ilościową. Jedyną cechą odróżniającą *korpus* od luźnej *kolekcji* tekstów wydaje się być właśnie cel jego utworzenia, którym są zazwyczaj badania lingwistyczne.

[2] Tłumaczenie angielskiego terminu *annotate* jako *anotować* zostało spopularyzowane przez projekt korpusowy IPI PAN, o którym wspominam dalej w rozdziale 5.4 (s. 51).

[3] [Świdziński, 1996], s. 155.

Zawartość pracy

Część I stanowi wprowadzenie do właściwej treści pracy. W rozdziale 1 podaję pojęcie wypowiednika i opisuję prace Świdzińskiego nad korpusem. W rozdziale 2 prezentuję przetwarzany zbiór w postaci oryginalnej. W rozdziale 3 przedstawiam konsekwencje zależności między wypowiednikami a wypowiedzeniami i opisuję konstrukcję korpusu wypowiedzeń w formacie XML-owym. W rozdziale 4 dokonuję interpretacji informacji dostępnej w korpusie niejawnie.

Część II opisuje wstępne etapy weryfikacji kolejnych warstw danych korpusu: rozdział 5 przedstawia wyniki dla warstwy typograficznej, rozdział 6 — morfologicznej.

Część III prezentuje weryfikację warstwy składniowej. W rozdziale 7 opisuję zwięźle narzędzia weryfikacji: formalizm analizy strukturalnej tekstów polskich — gramatykę formalną języka polskiego Marka Świdzińskiego oraz jej komputerową implementację — analizator składniowy Świgra. Rozdział 8 zajmuje się praktyczną weryfikacją składników frazowych oraz przedstawia wstępne wnioski z analizy składniowej pełnych tekstów wypowiedników. W rozdziałach 9, 10 i 11 opisuję kroki podjęte w celu podniesienia jakości analizy korpusu — kolejno poprzez korektę tekstową i morfologiczną próbek oraz rozszerzenie i usprawnienie gramatyki.

Część IV podsumowuje proces weryfikacji korpusu. W rozdziale 12 dyskutuję niektóre założenia przyjęte przy tworzeniu korpusu wypowiedników oraz poruszam kwestię wieloznaczności i obecności w korpusie elementów niezdaniowych. Rozdział 13 zwięźle omawia dane liczbowe uzyskane w procesie weryfikacji. W rozdziale 14 przedstawiam perspektywy przyszłego rozwoju poruszonych zagadnień.

Dodatki zawierają kolejno szczegółową charakterystykę opisową i liczbową wypowiedników, objaśnienie parametrów gramatyki Świdzińskiego, listę poprawionych w korpusie usterek z ich podziałem na kategorie oraz opis zawartości dołączonej do pracy płyty CD ze środowiskiem analizy składniowej i XML-ową wersją korpusu wypowiedników.

Oznaczenia przykładów

Przykłady wypowiedzeń ilustrujących omawiane własności językowe numeruję w sposób ciągły; podając je staram się używać oryginalnych tekstów wypowiedników (w postaci zawierającej oznaczenia znakowe wyodrębniające strukturę frazową lub bez nich), co znajduje odzwierciedlenie w oznaczeniu w nawiasach kwadratowych umieszczonych z prawej strony treści przykładu:

(0) *Następnie, po wygłoszeniu przemówienia, marszałek Marian Spychalski przekazuje na ręce ministra Czinege dla towarzyszy broni z Węgierskiej Armii Ludowej sztandar Ludowego Wojska Polskiego oraz portret Józefa Bema jako wyraz braterstwa i jedności we wspólnej służbie wielkim ideałom pokoju, socjalizmu i komunizmu.* *[6364]*

W razie potrzeby przytoczenia wariantu treści wypowiednika, która jednak ingeruje w oryginalny tekst (np. dla ilustracji pewnych szczególnych własności składniowych) dodaję do przytaczanego numeru oznaczenie literowe.

Tekst nielicznych przykładów zdań błędnych oznaczam zgodnie z konwencją symbolem gwiazdki.

Część I

Korpus wypowiedników

Rozdział 1

Wprowadzenie do korpusu wypowiedników

Korpus wypowiedników jest zestawem przeszło 6 700 próbek, których zasadniczym elementem jest fragment tekstu polskiego zaopatrzony w dodany ręcznie szczegółowy opis gramatyczny — morfologiczny i składniowy.

Zbiór powstał w wyniku projektu badawczego KBN **1 P104 030 04** *Ukierunkowana gramatycznie tekstowa baza danych: korpus wypowiedzeń współczesnej polszczyzny pisanej* realizowanego w latach 1993–1996 pod kierunkiem Marka Świdzińskiego w Instytucie Języka Polskiego UW przez grupę jego pracowników i studentów (których nazywam w dalszej części pracy *edytorami korpusu*). Założenia i wyniki projektu zostały opisane w książce [Świdziński, 1996].

1.1 Pojęcie wypowiednika

Wypowiednik to jednostka składniowa będąca bezpośrednią składową wypowiedzenia, realizowana jako zdanie lub oznajmienie. W sprawozdaniu z projektu Marek Świdziński definiuje ją jako „— intuicyjnie — zdanie lub składnik funkcjonalnie zdaniopodobny". Wypowiednikiem jest więc w szczególności całe wypowiedzenie oraz każdy jego składnik zdaniowy lub równoważnikowy — niekoniecznie bezpośredni. W ogólnym przypadku dane wypowiedzenie złożone jest źródłem wielu wypowiedników, z których każdy posiada własną, odrębną charakterystykę.

1.2 Pochodzenie wypowiedników korpusu

Materiałem źródłowym dla korpusu wypowiedników były teksty korpusu słownika frekwencyjnego polszczyzny współczesnej [Kurcz i in., 1990][1] — dane zebrane w latach 1963–1967 na potrzeby badań nad częstością występowania wyrazów w języku polskim. Składa się na nie 10 000 próbek po około 50 słów każda, czyli ogółem

[1] Patrz także [Ogrodniczuk, 2003b] i [Ogrodniczuk, 2003a].

ok. 500 000 słów z tekstów języka pisanego, zgromadzonych w pięciu transzach odpowiadających najważniejszym stylom polszczyzny pisanej. Korpus istnieje w różnych formach, różniących się sposobem reprezentacji polskich liter i innymi szczegółami. Jak wynika z opisu przedmiotu badań[2], Świdziński korzystał z jednej z jego najstarszych wersji, zapisanej w następującym formacie[3]:

```
Widac~ przypatrywanie[111] nie najgorzej dla[62] ciebie[42] wypadl~o,
Walik[/][171]. Teraz nikt ci[43] nie staje[5] w[+] poprzek, a wszyscy
chca~ pomagac~... Ty, bracie[171], uwaz~aj, bo za[+] duz~o
pomocniko~w[122], to[9] wiesz... O[7]! Ten[211] jak[9] strzeli, to[9]
nie wiadomo do[62] czego[42] mierzyl~... Naprawde~ nie wiadomo, moja
ty ,,sprzedana narzeczono''? Do[62] celu[121], Jan~cia[/][171],
a jakz~e. A cyl[141], to[8] Edus~[/][111] ma. We[64] mnie[4] sie~
wrodzil~[501].
```

Obecnie z inicjatywy J. S. Bienia dane i dokumentacja korpusu — dzięki zgodzie autorów na udostępnienie ich na zasadach licencji GNU — zamieszczone są w Internecie na utrzymywanej przeze mnie witrynie pod nazwą *Polszczyzna lat sześćdziesiątych XX wieku*[4]. Wersja korpusu zgodna z rekomendacjami konsorcjum TEI w XML-owym wariancie oznaczonym jako P4[5] jest też zamieszczona w zasobach *Oxford Text Archive*[6].

Wyjaśniając wybór korpusu słownika frekwencyjnego jako źródła danych dla badań nad składnią polską Świdziński odwołuje się do jego reprezentatywności lingwistycznej i statystycznej oraz adekwatności składniowej mimo nieaktualności części materiału leksykalnego („gramatyka ewoluuje wolniej niż leksyka")[7].

Z powodu ograniczeń czasowych opisowi poddano dziesiątą część danych korpusu frekwencyjnego. Autor uzasadnia to ograniczenie wynikami doświadczenia pilotowego z podkorpusem o długości ok. 5 000 słów — opisane w pracy [Świdziński, 1992b] badania wykazały dużą różnorodność danych ujawniającą większość faktów składniowych już w podkorpusie testowym (m. in. fakt występowania już w tak niewielkiej próbce 25 spośród wszystkich 33 schematów zdaniowych[8]), co pozwoliło na uznanie korpusu złożonego z 1000 próbek (dziesięciokrotnie większego) za reprezentatywny lingwistycznie.

1.3 Parametry wypowiedników

Wybrane próbki korpusu słownika frekwencyjnego (co dziesiąta poczynając od pierwszej, czyli próbki nr 1, 11 itd., z wyjątkiem miejsc uszkodzonych, kiedy wybierano próbki zastępcze[9]) dzielono na wypowiedzenia, które stanowiły jednocześnie

[2]Patrz [Świdziński, 1996], s. 1.
[3]Próbka nr 1821 stylu E — dramat (W Jezioranach, 17.05.1964, odc. 212, s. 21).
[4]Patrz `http://www.mimuw.edu.pl/polszczyzna/pl196x/`.
[5]Por. [TEIP4, 2001] oraz najnowszą wersję standardu [TEIP5, 2005].
[6]Patrz `http://ota.ahds.ac.uk/textinfo/2482.html`.
[7][Świdziński, 1996], s. 10–11.
[8]Patrz [Świdziński, 1996], s. 16.
[9]Patrz [Świdziński, 1996], s. 17.

wypowiedniki najwyższego poziomu. W każdym z nich (i następnie w każdej z jednostek podrzędnych, aż do ujawnienia się nie zawierającej wypowiedników struktury frazowej[10]) wyodrębniano wypowiedniki składowe. Każdy z wypowiedników poddawano opisowi polegającym na dodaniu szczegółowej informacji gramatycznej trzech rodzajów:

- o segmentacji tekstu,
- o strukturze tekstu,
- o parametryzacji segmentów.

Dla pojedynczego wypowiednika w pierwszej kolejności ustalano jego cechy dystrybucyjne (typ — zdanie lub oznajmienie, koordynacja — wypowiednik złożony lub elementarny, status — wypowiednik samodzielny lub składowy). Następnie w zależności od typu wypowiednika podejmowano decyzję w kwestii analizy składniowej, dokonywanej wyłącznie dla wypowiedników zdaniowych. Wynikiem tej analizy było wyróżnienie jednostek składowych poprzez nawiasowanie; rodzaj nawiasów odpowiadał typowi składnika. Sam wypowiednik był następnie kategoryzowany zgodnie z przyjętą taksonomią.

Dodatkowo próbki korpusowe opatrywano szczegółową informacją lokalizacyjną (w rodzaju oryginalnego numeru próbki w źródłowym korpusie słownika frekwencyjnego, numeru zdania w próbce, długości wypowiednika itp.) nieistotną z punktu naszych dalszych rozważań.

Szczegóły opisu pojedynczego wypowiednika zawiera rozdział A.1 (s. 181).

1.3.1 Klasyfikacja wypowiedników

Wypowiedniki złożone i elementarne

Nadrzędnym sposobem klasyfikacji wypowiedników jest ich podział na złożone i elementarne. *Wypowiedniki złożone* są oparte o centrum spójnikowe współrzędne (które stanowi spójnik lub szereg spójników równorzędnych, traktowanych wówczas również jako pojedynczy spójnik) i zawierają co najmniej dwa wypowiedniki składowe; konstrukcje nie spełniające tego warunku są *wypowiednikami elementarnymi*.

Współrzędne wypowiedniki składowe, których wyodrębnienie (zgodnie z założeniem, że centrum spójnikowe góruje nad finitywnym) rozpoczyna proces analizy składnikowej, ujmowane są niekiedy w nawiasy klamrowe ({...}). Wypowiedniki złożone nie są poddawane analizie frazowej (patrz rozdział 1.3.2, s. 19), która odbywa się dopiero na poziomie każdego z elementarnych wypowiedników składowych.

Składnikowość wypowiednika znajduje odzwierciedlenie w wartości parametru koordynacji o nazwie **WSP**. Wartością **K** oznaczone są wypowiedniki złożone, wartością pustą — elementarne. Liczba wypowiedników złożonych wynosi 700, co stanowi 10,42% ogólnej liczby wypowiedników. Warto w tym miejscu zaznaczyć, że przy badaniu elementarności wypowiednika wartość koordynacji należy traktować nadrzędnie, gdyż nawiasy klamrowe nie były stosowane konsekwentnie (lub zostały usunięte

[10]W pracy używam terminu *fraza* w rozumieniu Świdzińskiego, a zatem na oznaczenie członu składniowego poziomu niższej niż zdanie, nie zaś złożonej jednostki frazeologicznej.

z tekstów próbek podczas burzliwej historii korpusu) i są użyte w tekstach jedynie 225 wypowiedników złożonych o numerach niższych niż 3316, a zatem należących do stylu dramatycznego.

Wypowiedniki zdaniowe i oznajmieniowe

Podział wypowiedników na zdaniowe i oznajmieniowe dokonuje się ze względu na obecność frazy finitywnej. *Zdaniami* są w korpusie wypowiedniki elementarne o centrum finitywnym albo wypowiedniki złożone, których pierwszy niespójnikowy składnik bezpośredni jest wypowiednikiem zdaniowym. *Oznajmieniami* są natomiast wypowiedniki nie mające centrum finitywnego ani spójnikowego oraz dodatkowo takie wypowiedniki złożone, których pierwszy niespójnikowy składnik bezpośredni jest właśnie oznajmieniem.

Oto przykłady wypowiedników poszczególnych rodzajów:

(1) *Maku ja nie mam.* *[2684]*

(2) *Rosa bardzo słaba.* *[3896]*

(3) *Na fryzjerstwo się przerzuciłeś, czy co...* *[2869]*

(4) *Także dlatego, że obowiązuje nas stworzenie równych szans dla każdego obywatela każdego terenu.* *[5857]*

Przykłady 1 i 2 prezentują odpowiednio typowy wypowiednik zdaniowy i oznajmieniowy. Wypowiedzenie z przykładu 3 zawiera dwa współrzędne wypowiedniki składowe: pierwszy z nich jest wypowiednikiem zdaniowym, drugi — eliptycznym. Zgodnie z przyjętą regułą wypowiednik ten zostanie sklasyfikowany jako zdaniowy. Przykład 4 przedstawia sytuację symetryczną dla oznajmień (zdaniowy składnik podrzędny następuje po składniku oznajmieniowym, zatem wypowiednik złożony otrzymuje charakterystykę oznajmieniową).

Klasyfikacja ze względu na funkcję wewnątrz jednostki nadrzędnej

Dodatkową klasyfikację wypowiedników stanowi ich podział ze względu na funkcję, jaką pełnią względem wypowiednika macierzystego. Wyróżniamy:

- *wypowiedniki samodzielne* — stanowiące realizację wypowiedzenia,
- *wypowiedniki współrzędne* (*początkowe*, *środkowe* lub *końcowe*) — składniki wypowiednika złożonego o centrum spójnikowym współrzędnym,
- *wypowiedniki podrzędne* — składniki podrzędne wypowiednika złożonego,
- *wypowiedniki dostawione* — składniki luźne w rodzaju wtrąceń, wołaczy, wykrzyknień, nie tworzące związków z pozostałymi składnikami wypowiednika macierzystego,
- *zdania złożone – reszty* — składniki zdania elementarnego z „orzeczeniem szeregowym".

Oto przykłady wypowiedników poszczególnych rodzajów:

(5) *Dotąd jakby półmartwa, ożywiła się naraz, stanęła między Serabem a braćmi i dumnie zeznała: Już się nie zapieram, przyznaję się do tego, co zrobiłam.* *[5448]*

(6) *Dotąd jakby półmartwa, ożywiła się naraz.* *[5449]*

(7) *Stanęła między Serabem a braćmi.* *[5450]*

(8) *Dumnie zeznała: Już się nie zapieram, przyznaję się do tego, co zrobiłam.* *[5451]*

(9) *Obmyślono, że zostaną w nich posadzeni duchowni.* *[5316]*

(10) *Zostaną w nich posadzeni duchowni.* *[5317]*

(11) *Tak jest, panie dyrektorze...* *[5187]*

(12) *Panie dyrektorze...* *[5188]*

(13) *Chyba mi się przywidziało albo przyśniło.* *[1952]*

(14) *Się przywidziało albo przyśniło.* *[1953]*

Grupa wypowiedników (5)–(8) reprezentuje odpowienio wypowiednik samodzielny oraz składowe wypowiedniki początkowy, środkowy i końcowy; samodzielne wypowiedniki (9), (11) i (13) są źródłami kolejno wypowiednika podrzędnego (10), dostawionego (12) i zdania złożonego–reszty (14).

1.3.2 Oznaczenia składniowe

Jak już wspomniano, analizę składnikową zmierzającą do wyodrębnienia struktury frazowej prowadzono dla wypowiedników realizowanych jako zdania elementarne, czyli wypowiedników zdaniowych nie zawierających wypowiedników składowych. Szczegółowy opis procedury analitycznej zawiera instrukcja opisu wypowiedników [Świdziński, 1994a] na podstawie której redagowano próbki korpusu (jej duża część została bezpośrednio włączona do podsumowującej projekt pracy [Świdziński, 1996]).

W tekście wypowiedników elementarnych (korpus zawiera ich 4810, czyli ponad 71%) wyróżniano parami nawiasów następujące rodzaje fraz:

Rodzaj frazy	Symbol	Opis
finitywna	<...>	rozumiana tradycyjnie, być może poprzedzona partykułą *nie*, która stanowi wówczas jej integralną część,
podmiotowa	[...]	fraza nominalna w mianowniku lub jej równoważny dystrybucyjnie odpowiednik,
wymagana	(...)	fraza realizująca wymaganie czasownika — 4483 wystąpień,
luźna	/.../	składnik pomijalny strukturalnie (okolicznik lub zdanie okolicznikowe),
człon inny	\...\	część nie związana składniowo (w rodzaju wtrąceń, wykrzyknień, wołaczy).

Oto przykładowy wypowiednik zawierający wszystkie wymienione rodzaje składników[11]:

(15) *(Co) \więc\ [pan] /w końcu/ <postanowił>?* *[738]*

W przypadku wystąpienia w tekście więcej niż jednej frazy danego typu, oznaczenie znakowe frazy uzupełniane jest o jej numer, np. [1 ... 1] (w praktyce często numery dodawane są nadmiarowo, także w przypadku fraz występujących pojedynczo):

(16) */Przez całe życie/ <stawiają> (1 mi 1) (2 go 2) (3 za przykład 3).* *[752]*

Frazy nieciągłe oznaczane są z wykorzystaniem znaku wielokropka umieszczanego po nawiasie otwierającym i przed nawiasem zamykającym oznaczającymi daną część frazy:

(17) *<Nie trzeba> (ich...) \zresztą\ (...długo namawiać, żeby poszli do kina).* *[5898]*

Teksty wypowiedników podrzędnych mogą ponadto zawierać pochodzące z jednostki wyższego poziomu fragmenty nie należące do danego wypowiednika (dodane zapewne jako ułatwienie w czytaniu dla przyszłych użytkowników korpusu). W instrukcji dla edytorów[12] pojawia się uwaga o konwencji zapisu wypowiednika–reszty z ujętym w nawiasy)(...)(członem wspólnym; oznaczenie to, już w zmienionej postaci z parą pojedynczych nawiasów)...(występuje też w jednym z przykładów w pracy sprawozdającej projekt[13], nie jest w niej natomiast opisane. Tekst zawiera za to objaśnienie innego symbolu o podobnym charakterze, służącego do sygnalizacji spójników nie należących do wypowiednika (oznaczenie =...=, nie pojawiające się z kolei w instrukcji). W praktyce oba oznaczenia używane są w tekstach wymiennie.

Analizę wypowiedzenia przykładowego wyjaśniająca szczegółowo sposób wykorzystania wymienionych oznaczeń można znaleźć w pracy [Świdziński, 1996][14]; jej skrótowy wariant dla wybranego wypowiedzenia korpusowego zamieszczam dla celów poglądowych w rozdziale 3 (s. 31).

[11] W dalszej części pracy w przypadku wystąpienia w treści cytowanego przykładu symboli strukturyzujących tekst wypowiednika decyduję się ich dodatkowo nie komentować, odsyłając do wyjaśnienia w bieżącym rozdziale.

[12] [Świdziński, 1994a], s. 20.

[13] [Świdziński, 1996], s. 146.

[14] Patrz s. 64–69.

1.3.3 Zapis próbek

Korpus został wprowadzony do komputera w formie bazy danych (patrz rozdział 2, s. 25), w której każdemu wypowiednikowi odpowiada jeden rekord zawierający oprócz specjalnie oznaczonej treści wypowiednika także zestaw jego cech jakościowych i ilościowych oraz pomocniczych informacji lokalizacyjnych.

Oto jedna z próbek w zapisie stosowanym w pracy [Świdziński, 1996][15] — po umieszczonej w nawiasach nazwie pola następuje wartość parametru[16]:

```
[STYL] DR   [PR] 1821   [WYP] 2   [ZD] 2

[TEKST] /Teraz/ [nikt] (1 ci 1) <nie staje> (2 w poprzek 2),...

[TW] Z   [WSP] K    [ST] Wp      [TYP] A

[CEN] nie staje     [HAS] stawać      [NEG] N

[KL] V   [ASP] i    [CHAR] 3p,te

[SCH] 27            [OPIS] C+PS\$

[DL] 7

[VF] 2   [SU] 1   [OB1] 1   [OB2] 2   [LU1] 1   [LU2] 0   [IN] 0

         [TPSU]                                 [TPI]

    [SZYK] SOVO
```

1.4 Grupy wypowiedników

Zapisana w korpusie wypowiedników informacja gramatyczna sprawia, że tekstową bazę danych Świdzińskiego można traktować jako *korpus rozbiorów gramatycznych* (ang. *treebank*), definiowany znów intuicyjnie jako zbiór tekstów zawierający informację składniową ustalonego poziomu. Szerokie możliwości wykorzystania korpusów takiego rodzaju w badaniach nad składnią, dziś zazwyczaj wspieranych komputerowo, to m. in. możliwość konstrukcji modeli statystycznych dla fragmentów gramatyki, porównywania gramatyk dla różnych stylów języka, automatycznego generowania gramatyki czy weryfikacji i porównywania technik analizy składniowej. Anotacja składniowa będąca podstawą budowy korpusu rozbiorów jest także często uważana za etap pośredni między anotacją morfosyntaktyczną a anotacją semantyczną lub wyróżniającą strukturę wypowiedzi.

[15] Patrz s. 46 i 65–69.

[16] Por. 2.1 (s. 26). Szczegółowe wyjaśnienie znaczenia pól i dopuszczalnych wartości zawiera rozdział A.1 (s. 181).

Jednym z najbardziej znanych korpusów rozbiorów gramatycznych jest Penn Treebank[17] tworzony na Uniwersytecie Pensylwanii, którego podkorpus o wielkości około 3 mln słów amerykańskiej odmiany języka angielskiego zawiera oprócz dostępnej dla całego, półtorakrotnie większego korpusu informacji o częściach mowy (ang. *POS*) także informację składniową, generowaną „półautomatycznie" (uzyskane przy użyciu parsera o nazwie Fidditch wyniki automatyczne były następnie korygowane ręcznie).

Korpus wypowiedników, choć przechowywany w niekanonicznej postaci, jest także korpusem rozbiorów gramatycznych zdań złożonych — drzewu rozbioru odpowiada grupa wypowiedników powstałych z pojedynczego wypowiedzenia wraz z podstrukturą frazową ujawnioną w danym wypowiedniku.

Oto zestaw informacji jakościowej zawartej w korpusie dla cytowanego zestawu wypowiedników:

(18) *Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...* *[3040]*

(19) */Teraz/ [nikt] (1 ci 1) <nie staje> (2 w poprzek 2),...* *[3041]*

(20) *...=a= [wszyscy] <chcą> (pomagać)...* *[3042]*

Nr wypowiednika	`3040`	`3041`	`3042`
Typ wypowiednika	`Z` (zdaniowy)	`Z` (zdaniowy)	`Z` (zdaniowy)
Koordynacja	`K` (wypowiednik złożony)	brak wartości (wypowiednik elementarny)	brak wartości (wypowiednik elementarny)
Status	`S` (wypowiednik samodzielny)	`Wp` (wypowiednik współrzędny początkowy)	`Wk` (wypowiednik współrzędny końcowy)
Charakterystyka kontekstowa	brak	`A` (wypowiedniki współrzędne połączone spójnikiem *a*)	`A` (wypowiedniki współrzędne połączone spójnikiem *a*)
Centrum struktury	`a` (centrum spójnikowe)	`nie staje` (centrum finitywne)	`chcą` (centrum finitywne)
Klasa gramatyczna centrum	`C` (spójnik)	`V` (czasownik)	`V` (czasownik)
Charakterystyka fleksyjna centrum	brak	`3p,te` (3 os. l.p., czas ter.)	`3m,te` (3 os. l.mn., czas ter.)
Schemat zdaniowy[18]	`11` (1.1 — jednomiejscowy z frazą nominalną)	`27` (2.7 — dwumiejscowy z frazą nominalną i przysłówkową)	`16` (1.6 — jednomiejscowy z frazą werbalną)
Charakterystyka frazy wymaganej	brak	`C+PS$` (fraza nominalna w celowniku, fraza przysłówkowa)	`BOK` (bezokolicznik)

[17]Patrz `http://www.cis.upenn.edu/~treebank/`.

[18]Patrz rozdział 4.1, s. 39.

Przykład drzewa struktury gramatycznej zbudowanego w oparciu o tę informację (w wariancie bez kompletu parametrów składniowych, jedynie z identyfikacją numerów wypowiedników podrzędnych) prezentuje np. fragment zrzutu ekranu przygotowanej strony umożliwiającej przeglądanie korpusu (patrz rys. D.3, s. 210).

Rozdział 2

Korpus wypowiedników jako baza danych

Korpus wypowiedników w postaci oryginalnej jest tzw. relacyjną bazą danych, ograniczoną właściwie do jednej dużej tabeli, której wierszami są poszczególne próbki (wypowiedniki), zaś kolumnami — parametry ich opisu. Tabelaryczny model danych wpłynął także sposób ich przechowywania i obróbki — od początku wykorzystujący formaty i aplikacje baz danych.

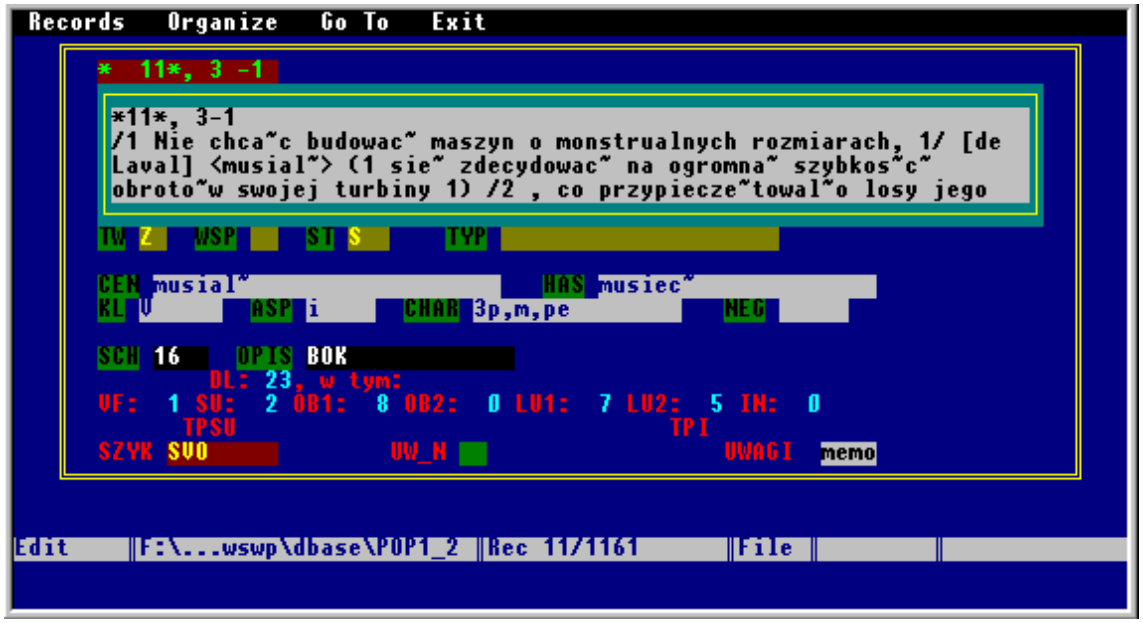
2.1 Baza danych Świdzińskiego

Zbiór wypowiedników utrzymywany był przez Świdzińskiego w postaci pliku programu dBASE IV[1], zawierającego złożone z 30 pól rekordy bazy danych z charakterystyką składniową i tekstami kolejnych wypowiedników. Polskie litery zapisano w nich w specyficznym kodowaniu: bezpośrednio po znaku bez diakrytu następuje znak tyldy lub pionowej kreski (w przypadku „*ź*") — dla zapewnienia poprawności ewentualnego sortowania.

Do przeglądania i uzupełniania danych służył program w języku Clipper z formularzem umożliwiającym wyświetlanie pojedynczych rekordów; jego oryginalną postać przedstawia rys. 2.1 (s. 26).

Warto zauważyć, że taka konstrukcja aplikacji ograniczała pracę z wypowiedzeniami złożonymi do ich poszczególnych składników, bez możliwości prezentacji całego wypowiedzenia, złożonego z grupy rekordów (patrz rozdział 1.4, s. 21). W ten sposób znacznie skomplikowano uzyskiwanie informacji o składni wypowiedników złożonych, łatwe dla człowieka przeglądającego ekrany z poszczególnymi rekordami, trudne natomiast dla automatów przetwarzających próbki.

[1] Świdziński wspomina także — patrz [Świdziński, 1996], s. 17 i 35 — formaty pliku tekstowego ASCII i tekstu w formacie edytora WordPerfect 5.1, materiałem końcowym jest jednak właśnie plik dBase, zaś pozostałe formaty są produktami ubocznym procesu obróbki danych źródłowego korpusu frekwencyjnego.



Rysunek 2.1: Zrzut ekranu z programu Świdzińskiego obsługującego oryginalną bazę danych wypowiedników

2.2 Baza danych Skibickiego

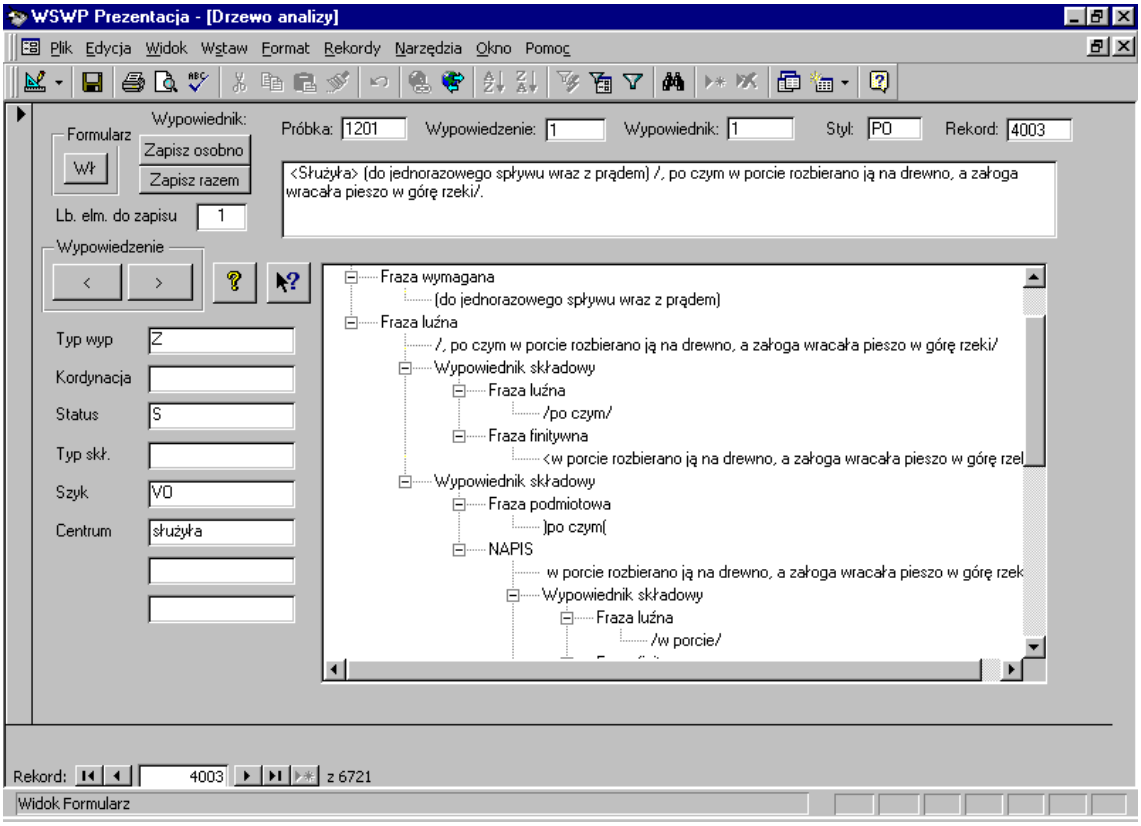
Na danych korpusu w wersji dostarczonej przez Świdzińskiego powstała wykonana pod kierunkiem Janusza S. Bienia praca magisterska Krzysztofa Skibickiego *Komputerowa weryfikacja wybranych zasobów lingwistycznych* [Skibicki, 2000], w ramach której dokonana została konwersja bazy wypowiedników z formatu dBase IV na Microsoft Access oraz opracowano program do analizy, prezentacji i edycji danych korpusu wypowiedników. Zrzut ekranu powstałego programu przedstawia rys. 2.2 (s. 27):

Oprócz czysto technicznego zadania konwersji formatu Skibicki dokonał także wstępnej weryfikacji danych korpusu:

- porównania przekazanych przez Świdzińskiego różnych wersji plików (korpusu zapisanego łącznie z podkorpusami pomocnicznymi zawierającymi dane poszczególnych stylów polszczyzny),
- zamiany kodowania polskich liter na format Windows (strona kodowa 1250),
- korekty licznych błędów w opisie jakościowym próbek (brak zapisu informacji składniowej, kolidująca z oznaczeniami gramatycznymi reprezentacja nawiasów, oznaczenia niezgodne z instrukcją dla edytorów, błędna klasyfikacja składników, błędny opis informacji o szyku),
- automatycznego sprawdzenia danych ilościowych i ich poprawienia (w rodzaju błędnie podawanej długości fraz).

W ramach analizy struktury frazowej wykryto także oznaczenia dodatkowe niejasnego pochodzenia (nie udokumentowane ani w instrukcji, ani w książce omawiającej projekt). Skibicki dopatruje się ich źródeł w sposobie prowadzenia prac edycyjnych[2]:

[2]Patrz [Skibicki, 2000], rozdział 4, s. 18.



Rysunek 2.2: Zrzut ekranu z programu Skibickiego umożliwiającego przeglądanie i korektę wypowiedników

> Osoby wykonujące analizę często wprowadzały rozszerzenia do konwencji oznaczeń, aby zaznaczyć nieprzewidziane w niej zjawisko.

Jest to jednak wyjaśnienie niepewne, skoro wprowadzane symbole nie zostały włączone przez Świdzińskiego do opisu wyników prac, a jednocześnie pozostały w bazie.

Oznaczeniem tego rodzaju są dodatkowe znaki dodawane do numerów fraz w ich ogranicznikach w przypadku łącznego wystąpienia oznaczenia frazy nieciągłej i numeru frazy; dla pierwszego członu jest to apostrof, dla drugiego — cudzysłów:

(21) *[Ucho] (1' czułe... 1') <jest> \bowiem\ (1" ...na bardzo ograniczony zakres częstotliwości 1")...* *[3372]*

Sam Skibicki wprowadza natomiast do tekstu[3] dla ułatwienia automatycznego przetwarzania oznaczenie #...# odpowiadające wystąpieniu w materiale źródłowym fragmentu ujętego w nawiasy okrągłe (co w tekście korpusu edytorzy oznaczali ponoć podwójnymi nawiasami okrągłymi — również bez dokumentacji).

Jednym z celów pracy Skibickiego było też umożliwienie prezentacji pełnych wypowiedzeń, zatem odmiennie niż w przypadku bazy danych Świdzińskiego, korpus

[3]Patrz [Skibicki, 2000], rozdział 7, s. 29.

w wersji stworzonej przez Skibickiego zawiera już szczątkową, dedukowaną półautomatycznie informację o powiązaniach między grupami wypowiedników tworzących jedno wypowiedzenie złożone (w postaci drzewa struktury przedstawionego na zrzucie ekranu z rys. 2.2 na poprzedniej stronie). Mechanizm łączenia wypowiedników oparty na dopasowywaniu tekstów ograniczono jednak wyłącznie do badania długości elementów składowych na mocy przeświadczenia, że ze względu na brak informacji o powiązaniach w ramach grupy nie jest możliwe automatyczne połączenie wypowiedników elementarnych w drzewa reprezentujące struktury złożone. W przypadku trudności w dopasowaniu pozostawiano decyzję o sposobie połączenia wypowiedników użytkownikowi programu — poprzez wyświetlenie okna umożliwiającego ręczne dokonanie wyboru. Praktyczne znaczenie tego rozwiązania było jednak ograniczone ze względu na pracochłonność procesu łączenia wypowiedników, który nie został nigdy wykonany. Jak pokazuję w następnym rozdziale, automatyczna budowa drzew poprzez dokładniejszą analizę tekstowej postaci wypowiedników jest jednak możliwa i została wykonana w ramach niniejszej pracy.

Z punktu widzenia dalszych rozważań najważniejszym wynikiem pracy Skibickiego było więc zebranie danych korpusu wypowiedników i ich zapis w spójnej postaci (numery rekordów w bazie stały się unikalnymi numerami wypowiedników w wersji bieżącej) oraz ręczna korekta dostrzeżonych błędów[4].

2.3 Bieżąca wersja korpusu

Korpus w wersji Skibickiego zawiera 6721 wypowiedników, która to liczba odbiega znacznie od podawanej przez Świdzińskiego[5] wartości 8907 próbek. Różnica ta ujawniła się już na wstępnym etapie pracy Skibickiego, który podaje szczegóły uzgadniania wersji korpusu[6]:

> W trakcie pracy nad programem WSWPP udostępniony został pełen korpus WSWP. W zawierającej 7170 próbek bazie zawarto rekordy z wszystkich pięciu stylów wypowiedzi. (...) Przed przekazaniem korpus został przejrzany i ręcznie oczyszczony. Niestety, weryfikacja wykazała, że w korpusie znalazły się grupy rekordów wkopiowane dwukrotnie oraz rekordy podrzędne, dla których nie ma wypowiedników nadrzędnych. Rzeczywista liczba poprawnych wypowiedników w bazie wynosi 6721. Uzyskany plik stał się podstawą dalszej pracy.

Zgodnie z powyższym zakładam, że wersja korpusu pochodząca z pracy Skibickiego jest wersją najbardziej spójną z dostępnych. Weryfikacja rozkładu wypowiedników ze względu na oryginalne transze korpusu słownika frekwencyjnego wykazała, że wersja bieżąca zawiera zgodną z wartościami podawanymi przez Świdzińskiego liczbę

[4] Pełną listę poprawionych błędów zawiera dodatek A (s. 34–44) do pracy [Skibicki, 2000].

[5] Patrz [Świdziński, 1996], rozdział 6.3, s. 72–73.

[6] [Skibicki, 2000], rozdział 7, s. 29. Liczba rekordów w przekazanym materiale również odbiega od wartości bliskiej 9000 próbek, podawanej zarówno w cytowanej pracy, jak i w raporcie końcowym projektu [Świdziński, 1997].

próbek stylu popularnonaukowego i dramatu oraz około połowy liczby próbek stylu wiadomości prasowych, publicystycznego i prozy artystycznej — łącznie przeszło 75% zawartości oryginalnej.

Dla wygody dalszego przetwarzania korpusu niezbędne było zapisanie go w pliku tekstowym, gdzie każdemu rekordowi odpowiada jeden wiersz. Przykładowy wypowiednik

(22) *Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...* *[3040]*

ma w tej formie postać 3 wierszy (rekordów), niżej podzielonych ze względów typograficznych:

```
3040;;1821;;2;;1;;Z;;K;;S;;;;a;;a;;C
    ;;;;;;;;;;;;11;;0;;0;;;;0;;0;;0;;0;;11;;\%;;;;DR
    ;;Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...

3041;;1821;;2;;2;;Z;;;;Wp;;A;;nie staje;;stawać;;V
    ;;i;;3p,te;;N;;27;;C+PS\$;;7;;2;;1;;;;1;;2;;1;;0;;0;;;;SOVO;;DR
    ;;/Teraz/ [nikt] (1 ci 1) <nie staje> (2 w poprzek 2),...

3042;;1821;;2;;3;;Z;;;;Wk;;A;;chcą;;chcieć;;V
    ;;i;;3m,te;;T;;16;;BOK;;3;;1;;1;;;;1;;0;;0;;0;;0;;;;SVO;;DR
    ;;...=a= [wszyscy] <chcą> (pomagać)...
```

Zaprezentowany fragment zawiera wypowiednik złożony oraz dwa pochodne wypowiedniki elementarne; poszczególne pola z opisem zostały oddzielone znakiem podwójnego średnika.

Rozdział 3

Korpus wypowiedzeń w formacie XML-owym

Korpus wypowiedników w postaci bazy danych udostępnia ograniczone środki reprezentacji informacji o wypowiedzeniach, którym odpowiadają grupy rekordów. Rozwiązaniem jest opisana poniżej zmiana formatu korpusu, której celem jest ujawnienie drzewiastej struktury wypowiedników złożonych, reprezentowanej zarówno w bazie Świdzińskiego, jak i Skibickiego w sposób ukryty. Jej wynikiem jest zbiór nazywany dla odróżnienia *korpusem wypowiedzeń*, zawierający dla wypowiedników elementarnych informację bezpośrednio odpowiadającą danym z bazy, zaś dla wypowiedników złożonych — zapis informacji składniowej dla całej grupy rekordów.

Zmiana reprezentacji korpusu jest dobrą okazją do ulepszenia formatu danych. Korpus wypowiedników, jak wiemy, wprowadza własny, specyficzny zestaw oznaczeń[1]. Może się on wydawać wystarczający do edycji i przeglądania wyników analizy przez człowieka, jest jednak z całą pewnością za mało uniwersalny dla potrzeb neutralnej reprezentacji danych rozbioru gramatycznego, której głównym celem jest z jednej strony ogólność i kompletność zapisu, z drugiej — łatwość przetwarzania danych i ich przenośność. Więcej: format ten wydaje się nie spełniać żadnego z tych warunków — błędnie zakłada on jednoznaczność rozbioru, jego dalsze automatyczne przetwarzanie jest trudne, a format — niestandardowy[2].

Wad tych nie posiadają formaty oparte o *języki adiustacyjne* (ang.*markup languages*), takie jak SGML (ang. *Standard Generalized Markup Language* — Standardowy Uogólniony Język Adiustacyjny[3]) i jego prostsza, aczkolwiek wystarczająca do większości zastosowań wersja XML (ang. *eXtensible Markup Language* — Rozszerzalny Język Adiustacyjny[4]). Ich idea opiera się na pojęciu tzw. *generycznej adiustacji tekstu* (ang. *generic markup*), czyli takiego jego opisu, który przyporządkowuje fragmentom tekstu określoną strukturę logiczną (co umożliwia spełnienie

[1]Patrz rozdział 1.3.2, s. 19.

[2]Jak pamiętamy, oryginalny format dBase IV został przez Skibickiego zamieniony na format programu Microsoft Access; oba są niestety związane z określonymi produktami firm komercyjnych, co wymusza konieczność stosunkowo częstej konwersji danych i uzależnia zainteresowanych od producenta konkretnego narzędzia, co należy uznać za sytuację co najmniej mało komfortową.

[3]Patrz [SGML, 1986].

[4]P. [XML, 2004].

pierwszego założenia o maksymalnej pełności zapisu, gdyż stopień ogólności reprezentacji nadawanej fragmentom analizowanego tekstu może być dowolny). Innym ważnym założeniem takiego sposobu reprezentacji informacji jest zapewnienie jej maksymalnej trwałości i przenośności poprzez drastyczne ograniczenie środków reprezentacji informacji (format tekstowy oraz bodaj najprostszy z możliwych sposób jego anotacji), co spełnia drugi warunek naszego założenia, sprawiając że przetwarzanie tak przygotowanych zasobów jest względnie proste i tanie.

3.1 Wykorzystanie języków adiustacyjnych do reprezentacji danych korpusowych

Przykład wykorzystania języka SGML do reprezentacji polskiej informacji lingwistycznej został przeze mnie opisany w pracy [Ogrodniczuk, 2000], natomiast możliwość automatycznej konwersji tego opisu do zdobywającego coraz silniejszą pozycję języka XML — w artykule [Ogrodniczuk, 2004]. Oba formaty mają już za sobą długą tradycję wykorzystania w zapisie korpusów; dwa najbardziej znane z ich zastosowań to TEI i CES, udostępniające zestawy znaczników o określonym przeznaczeniu.

TEI (ang. *Text Encoding Initiative*) to nazwa rozpoczętego w 1987 roku wspólnego projektu stowarzyszeń informatycznych i humanistycznych mającego na celu określenie spójnego, rozszerzalnego standardu zapisu informacji tekstowej dla celów lingwistycznych. Jego wynikiem jest stale uaktualniany zestaw reguł zapisu i wymiany danych tekstowych, publikowany regularnie jako „wytyczne" (ang. *Guidelines for Electronic Text Encoding and Interchange*[5]). Do bardzo znanej wersji trzeciej standardu z 1994 roku formatem reprezentacji danych był SGML; wersje późniejsze, w tym obecna (piąta) dopuszczają na równych prawach stosowanie formatów SGML i XML.

Format CES (ang. *Corpus Encoding Standard*) powstał w 1996 roku jako oparty na TEI SGML-owy schemat reprezentacji szczególnie przydatny do opisu korpusów lingwistycznych; jego XML-owa wersja otrzymała akronim XCES. Format ten został wykorzystany m. in. do zapisu korpusu IPI PAN; szczegóły jego zastosowania opisuje np. praca [Bański, 2001]. Przykłady fragmentów próbek tego korpusu zawierają rozdziały 6.3 (s. 59) i 9.2.2 (s. 96).

3.2 Wynikowy format danych

Do zapisu korpusu wypowiedników wybrałem format XML-owy ze względu na intensywny rozwój standardów pokrewnych i dostępność wielu darmowych narzędzi umożliwiających obróbkę tak reprezentowanych danych.

Jak już wspomniałem, XML jest formatem tekstowym (ang. *plain text*) używającym do zapisu reprezentowanej informacji logicznej określonych *znaczników* (ang. *tags*) — sparametryzowanych jednostek tekstowych o ustalonej składni, łatwo odróżnialnych od anotowanego (czyli tego rodzaju dodatkową informacją oznaczanego)

[5]Patrz [TEIP5, 2005].

tekstu właściwego. Dokument XML-owy w rozważanym przez nas zakresie zawiera zatem dokładnie trzy rodzaje jednostek: tekst anotowany ujmowany w klamry oznaczeń — *elementy*, mogących posiadać dodatkowe parametry — *atrybuty*. Poprzez zawieranie w sobie fragmentów tekstu (lub innych elementów) elementy tworzą usystematyzowaną, drzewiastą strukturę dokumentu, zaś atrybuty pozwalają nadawać składnikom tej struktury nazwane własności (również w postaci ciągów znaków).

Zapis elementu wymaga użycia specjalnych znaczników z nazwą elementu — początkowego, postaci <`nazwa_elementu`> i końcowego, postaci </`nazwa_elementu`>. Nazwy i wartości atrybutów umieszczane są „wewnątrz" znacznika początkowego (przed zamykającym go nawiasem kątowym) i mają postać tekstu `nazwa_atrybutu="wartosc_atrybutu"`. Nazwa elementu i definicje wartości atrybutów rozdzielone są spacjami. Poprawność składniowa formatu wymaga ponadto występowania *deklaracji XML-owej* (rozpoczynającej dokument specjalnej instrukcji zawierającej informację o wersji standardu i kodowaniu znaków w dokumencie) oraz dokładnie jednego *elementu głównego*, nadrzędnego dla pozostałych elementów struktury.

Oto przykład najprostszego dokumentu XML-owego prezentującego opisane wyżej własności:

```
<?xml version="1.0"?>
<wypowiednik id="0766" tw="Z">
  <ff>
    <term>Prześpij</term>
    <term>się</term>
  </ff>
  <fl>
    <term>lepiej</term>
  </fl>
  <term>...</term>
</wypowiednik>
```

Rozpoczynający przykładowy dokument tekst <`?xml version="1.0"?`> to właśnie wymagana deklaracja XML-owa; <`wypowiednik`> jest elementem głównym rozciągającym się na cały dokument; <`ff`>, <`fl`> i <`term`> to pozostałe elementy, przy czym zawartość elementu <`term`> jest czysto tekstowa, zaś innych elementów — złożona. Element <`wypowiednik`> posiada dwa atrybuty — `typ` o wartości `Z` i `id` o wartości `0776`.

Powyższy przegląd jest oczywiście uproszczony do niezbędnego minimum cech standardu XML, wystarczającego do zrozumienia zapisu próbek korpusu; więcej informacji o standardzie zawiera np. rozdział *A Gentle Introduction to XML* wytycznych TEI [TEIP5, 2005][6] lub książka [Prinke, 2000].

Warto też zaznaczyć, że świadomie pominąłem ważne z punktu widzenia standardu zagadnienie tworzenia XML-owych języków opisu dokumentów (przeanalizowane dokładnie np. w artykule [Ogrodniczuk, 2001]), czyli reprezentujących zależności pomiędzy elementami i ich atrybutami sformalizowanych wzorców hierarchii części

[6] Patrz `http://www.tei-c.org/release/doc/tei-p5-doc/html/SG.html`.

składowych przyszłych dokumentów-egzemplarzy. W naszym korpusie projekt takiego języka jest w dużej mierze implikowany składnią gramatyki, a jego tworzenie nie wydaje się celowe (definicję ograniczam do podania listy dopuszczalnych elementów, bez określania reguł ich wzajemnego zawierania się).

3.3 Konstrukcja drzew wypowiedzeń

Korpus drzew rozbioru dla wypowiedników analizowanych ręcznie został stworzony jako wariant oryginalnego korpusu wypowiedników, tj. z założeniem przechowania kompletu informacji jakościowych i ilościowych zawartych w wersji źródłowej oraz z zachowaniem oryginalnego podziału na próbki.

Jednostkę opisu korpusu XML-owego stanowi, jak w korpusie Świdzińskiego, pojedynczy wypowiednik. Próbka zawiera oryginalne parametry wypowiednika zapisane w dokładnie jednym drzewie rozbioru ręcznego, stworzonym na bazie oryginalnych oznaczeń. W przypadku wypowiedników niezdaniowych drzewa rozbioru ręcznego występują w postaci zdegenerowanej i zawierają wyłącznie tekst i komplet informacji korpusowych poziomu wypowiednika, bez dalszego rozbicia jego struktury.

Węzłom drzewa analizy odpowiadają elementy XML-owe o nazwach odpowiadających nazwom wyodrębnionych przez Świdzińskiego jednostek składniowych. Atrybuty służą z kolei do przechowania cech składniowych węzłów (parametrów i ich wartości). Tak utworzone drzewa dokumentów XML-owych są bezpośrednio drzewami analizy składniowej wypowiedzenia reprezentowanego w danym dokumencie.

Lista składników drzew analizy ogranicza się do następujących elementów:

- <`wypowiednik`> — element zawierający cały wypowiednik: element główny dokumentu zawierającego drzewo rozbioru lub składnik wypowiednika nadrzędnego,
- <`ff`>, <`fp`>, <`fl`>, <`fw`>, <`in`> — odpowiednio fraza finitywna, podmiotowa, luźna, wymagana, i człon inny; dla fraz nieciągłych element zawiera fragment frazy i atrybut `czesc` o wartości odpowiadającej numerowi fragmentu,
- <`term`> — element terminalny: rozumiane intuicyjnie słowo lub znak interpunkcyjny.

Zaproponowany format nie bazuje na żadnym z formalnych opisów ze względu na ich ekstensywność, nieadekwatną do bieżącego zastosowania. Konwersja na inne formaty reprezentacji danych korpusowych jest jednak oczywiście możliwa i prosta, gdyż każdy z nich zawiera co najmniej elementy służące neutralnej reprezentacji arbitralnych danych (TEI udostępnia do tego celu np. element <`w`>, XCES — element <`chunk`>).

W przypadku wypowiedników złożonych, zgodnie z założeniem budowy korpusu wypowiedników, konstrukcja drzewa rozbioru wymaga przetworzenia nie tylko treści danego wypowiednika, ale i wypowiedników zależnych (podrzędnych), dlatego elementy <`wypowiednik`> mogą być zagnieżdżane. Każde zagnieżdżenie reprezentuje użycie wypowiednika podrzędnego w treści nadrzędnego — wykryte automatycznie, gdyż parametry wypowiednika nie zawierają informacji o wzajemnych zależnościach

wypowiedników danej grupy (utworzonej przez wypowiedniki należące do tego samego zdania — numer próbki i zdania w obrębie próbki słownika frekwencyjnego należy do parametrów lokalizacyjnych każdego wypowiednika).

Jak pamiętamy, nie jest również regułą nawiasowanie struktury wypowiednika złożonego, zatem proces analizy grupy polegał na dopasowaniu tekstowym fragmentów treści wypowiedników zależnych, w czym sporą trudność sprawiał specyficzny, wskazujący niesamodzielność reprezentowanych członów zapis interpunkcji wypowiedników z początkowym i końcowym wielokropkiem, które, co więcej, nie były stosowane konsekwentnie. Dla wypowiedników sklasyfikowanych jako niesamodzielne tego rodzaju znaki, o ile nie miały odpowiedników w wypowiednikach nadrzędnych, nie zostały uwzględnione w wypowiednikach składowych.

Mimo początkowych obaw co do jakości reprezentacji grup (patrz rozdział 2.3, s. 28) ich automatyczna analiza nie przysporzyła większych problemów, co dowodzi, że analiza składnikowa została przeprowadzona z dużą starannością. Z punktu widzenia łączenia opisów przydatne okazało się spostrzeżenie, że treść wypowiednika podrzędnego po usunięciu oznaczeń frazowych ma w prawie każdym przypadku bezpośredni odpowiednik we fragmencie treści wypowiednika nadrzędnego, co znacznie ułatwiło dopasowanie tekstów. Podczas konstrukcji wynikowej postaci każdego wypowiednika wziąłem pod uwagę dostępną w ramach grupy analizę o maksymalnej długości, co sprawia, że wypowiednikom odpowiadają analizy o maksymalnej dostępnej złożoności.

Jeden z najprostszych przykładów wynikowych drzew rozbioru dla wypowiednika samodzielnego został już przedstawiony w rozdziale 3.2 (s. 32), oto przykład dla przywoływanego już w postaciach oryginalnej i drzewiastej wypowiednika złożonego:

```
<wypowiednik id="3040">
  <wypowiednik id="3041">
    <fl>
      <term>Teraz</term>
    </fl>
    <fp>
      <term>nikt</term>
    </fp>
    <fw>
      <term>ci</term>
    </fw>
    <ff>
      <term>nie</term>
      <term>staje</term>
    </ff>
    <fw>
      <term>w</term>
      <term>poprzek</term>
    </fw>
  </wypowiednik>
  <term>,</term>
```

```
  <term>a</term>
  <wypowiednik id="3042">
    <fp>
      <term>wszyscy</term>
    </fp>
    <ff>
      <term>chcą</term>
    </ff>
    <fw>
      <term>pomagać</term>
    </fw>
  </wypowiednik>
  <term>...</term>
</wypowiednik>
```

W przeciwieństwie do oryginalnego korpusu wypowiedników w korpusie XML-owym nie umieszczam:

- członów oznaczonych jako nie należące do wypowiednika,
- fragmentów wypowiednika nadrzędnego włączonych do wypowiednika podrzędnego dla ułatwienia czytania tekstu wypowiednika,
- znaków interpunkcyjnych z początku wypowiednika.

Przyjęte założenie niereprezentowania członów oznaczonych jako nie należące do wypowiednika skutkuje niekiedy brakiem jego treści:

(23) *...=ale= 0.* *[2449]*

(24) *=Bo= 0?* *[2455]*

(25) *=że= 0...* *[5179]*

(nie opisane oznaczenie `0` reprezentuje zapewne człon pusty; wszystkie powyższe wypowiedniki są elipsami). W tego rodzaju przypadku plik XML-owy stanowiący próbkę korpusu wypowiedzeń zawiera wyłącznie pusty element <`wypowiednik`> przechowujący charakterystykę wypowiednika.

Z powyższego wykazu stosunkowo największe konsekwencje niesie rezygnacja ze znaku interpunkcyjnego z początku wypowiednika, przede wszystkim ze względu na korpusową niekonsekwencję umieszczania znaków interpunkcyjnych na początku i końcu wypowiednika (niejednokrotnie znaki interpunkcyjne, nie oznaczone w żaden szczególny sposób, wydają się być używane w korpusie jedynie dla podkreślenia niesamodzielności wypowiednika, a nie jako jego właściwe składniki, które to przypadki są od siebie nieodróżnialne). Podobnie zapis wypowiednika jako rekordu, któremu odpowiada wypowiedzenie, implikuje (być może tylko w mniemaniu edytorów, gdyż reguła ta również nie jest stosowana konsekwentnie) jego zakończenie znakiem interpunkcyjnym równoważnym kropce, podczas gdy użycie wypowiednika jako składowej struktury wyższego poziomu wymaga usunięcia końcowego znaku interpunkcyjnego. Z tego względu proces sklejania wypowiedników nie jest kwestią trywialną i wymaga

konsekwentnej obsługi interpunkcji. W przyjętym zapisie przyjmuję, że znaki interpunkcyjne otaczające wypowiednik podrzędny nie są jego częścią. Takie rozwiązanie wydaje się poprawne i zgodne z dużą autonomią wypowiedników, jaka została im nadana przez twórcę korpusu.

Rozdział 4

Korpus wypowiedników jako źródło informacji pochodnych

Oprócz interesującej mnie w szczególny sposób informacji gramatycznej korpus wypowiedników jest także cennym źródłem innych danych lingwistycznych. Niektóre z nich uwypuklił sam Świdziński, prezentując aspekt statystyczny niektórych zjawisk w zakresie występowania poszczególnych schematów zdaniowych oraz charakterystyki fraz realizujących je typów, a także notując obserwacje z zakresu porządku składników zdania elementarnego czy typologii oznajmień.

Poniższy opis ma na celu jedynie zasygnalizowanie bogactwa lingwistycznego korpusu w zakresie wpływającym bezpośrednio na składnię, nie wkraczam zatem na teren wykorzystania danych korpusu np. w pracach leksykograficznych, również wymienianych jako jedna z motywacji jego powstania, poruszam tu natomiast problemy syntaktyczne nie omawiane szerzej w dalszej części pracy.

4.1 Schematy zdaniowe

Jednym z najbardziej znaczących elementów opisu rekordów korpusu jest ich klasyfikacja względem *schematów zdaniowych*, których symbole zostały przypisane wszystkim zdaniom elementarnym. Ciekawym zadaniem wydaje się możliwość zbadania ich rozkładu i reprezentowalności.

Koncepcja schematu zdaniowego datuje się na rok 1966; jej autorem jest Kazimierz Polański [Polański, 1966]. Schemat zdaniowy to model struktury zdania elementarnego zadany przez zbiór fraz wymaganych przez *verbum finitum*[1], czyli abstrakcja zdania empirycznego złożona z typów fraz wymaganych, nie uwzględniająca porządku składników ani członów luźnych.

Nadrzędna klasyfikacja schematów zdaniowych dzieli je na czasownikowe i quasi-czasownikowe, czyli odpowiednio obejmujące i nie obejmujące wymaganej frazy nominalnej w mianowniku („podmiotu") uzgadniającej z frazą finitywną osobę, liczbę

[1] Por. [Szpakowicz i Świdziński, 1981], [Świdziński, 1993c], [Świdziński, 1993b], [Świdziński i Szpakowicz, 1994].

i rodzaj. Klasyfikacja niższego poziomu dzieli schematy ze względu na liczbę fraz wymaganych i konkretne klasy wymagań.

Oto poprawiona na potrzeby podręcznika [Saloni i Świdziński, 1981] tabela schematów zdaniowych[2] wraz z zebraną informacją statystyczną dotyczącą rozkładu schematów w korpusie wypowiedników:

Schematy czasownikowe			
Zeromiejscowe			
V-0	0	Jan śpi.	691
Jednomiejscowe			
V-1.1	fno	Jan kupuje dom.	1313
V-1.2	fpn	Jan śmieje się z Marii.	475
V-1.3	fpm	Jan jest głupi.	362
V-1.4	fpp	Jan wygląda na zmęczonego.	6
V-1.5	fps	Jan zachowuje się niewłaściwie.	396
V-1.6	fwb	Jan chce spać.	402
V-1.7	fzd	Jan pyta, która godzina.	253
V-1@[3]	—	*A jednak wolę, niech patrzą, niż mieliby opuszczać oczy.*	61
Dwumiejscowe			
V-2.1	fno + fno	Jan pożycza Marii książkę.	128
V-2.2	fno + fpn	Jan pożycza książki od Marii.	222
V-2.3	fno + fpm (bez zgody rodzaju i liczby)	Jan wydaje się Marii miły.	3
V-2.4	fno + fpm (ze zgodą rodzaju i liczby)	Jan pamięta Marię młodą.	8
V-2.5	fno + fpp (bez zgody rodzaju i liczby)	Jan wygląda nam na zmęczonego.	0
V-2.6	fno + fpp (ze zgodą rodzaju i liczby)	Jan bierze ją za wykształconą.	3
V-2.7	fno + fps	Jan stawia kubek tutaj.	93
V-2.8	fno + fwb	Jan każe Marii czekać.	23
V-2.9	fno + fzd	Jan mówi Marii, że nie ma czasu.	28
V-2.10	fpn + fpn	Jan dowiaduje się o Piotrze od Marii.	19
V-2.11	fpn + fps	Jan mówi o Marii nieżyczliwie.	14
V-2.12	fpn + fzd	Jan wie od Marii, gdzie Piotr mieszka.	7
V-2.13	fps + fps	Jan leci stąd do Berlina.	8
V-2@	—	*Z tego, co mówisz, widać, że przeprowadza się na dobre...*	8

[2]Cytowana za [Świdziński, 1996], s. 29–30, z angielskimi skrótami oznaczeń fraz wymaganych zastąpionymi symbolami polskimi dla potrzeb późniejszej dyskusji.

[3]Jak się okazuje, przyjęty rejestr schematów zdaniowych okazał się niewystarczający — w korpusie pojawiły się schematy nietypowe, oznaczane specjalnie. W przypadku tego rodzaju schematów podaję odnośny przykład korpusowy i wyróżniam go kursywą.

Trzymiejscowe[4]			
V-3@	—	*Przez całe życie stawiają mi go za przykład.*	11
Czteromiejscowe			
V-4@	—	*Z funduszów SFBSII przekazano dotychczas dodatkowo województwom na budowę mieszkań nauczycieli wiejskich pięćset milionów złotych, co pozwoli uzyskać około piętnastu tysięcy nowych izb.*	1

Schematy quasi-czasownikowe			
Zeromiejscowe			
Q-0	0	Dnieje.	18
Jednomiejscowe			
Q-1.1	fno	Jana mdli.	45
Q-1.2	fpn	Czas na obiad.	15
Q-1.3	fps	Jest zimno.	23
Q-1.4	fwb	Trzeba pracować.	128
Q-1.5	fzd	Wiadomo, co Jan powie.	14
Dwumiejscowe			
Q-2.1	fno + fno	Janowi brak pieniędzy.	8
Q-2.2	fno + fpn	Jana ciągnie do picia.	13
Q-2.3	fno + fps	Janowi idzie dobrze.	12
Q-2.4	fno + fwb	Pracować jest niepodobieństwem.	5
Q-2.5	fno + fzd	Janowi żal, że Maria uciekła.	0
Q-2.6	fwb + fps	Wygrywać jest łatwo.	6
Q-2@	—	*Z tego, co mówisz, widać, że przeprowadza się na dobre...*	3

Najciekawsze przypadki stanowią oczywiście wystąpienia schematów nietypowych. Większość z nich (63) jest pochodną reprezentacji tekstu w mowie niezależnej (schemat uzupełniony o oznaczenie `@o`):

(26) *(Ale to nie matka), <mówiłam>.* *[1154]*

Nietypowość przypadków pozostałych (łącznie 21) wynika w całości z interpretacji użycia frazeologii — jako wymagane zostały oznaczone frazy ewidentnie luźne, jak w przykładach:

(27) [*Łącznik z placówki, przyprowadzony do obozu przez straże*], *<był> (1 święcie przekonany 1) (2, że go wciągnięto w zasadzkę 2).* *[4957]*

(28) */Teraz/ (1 sobie 1) \znowu\ <nabił> (2 głowę 2) (3 jakimiś nowymi planami 3)...* *[2820]*

[4]Schematy trzy- i czteromiejscowe są wyłącznie nietypowe!

4.2 Słownik czasowników z informacją składniową

Jednym z najciekawszych zagadnień „okołoskładniowych", dość dobrze reprezentowanych danymi korpusu, jest możliwość zbadania zakresu wymagania czasownika, zapisanego w parametrze `OPIS`. Opis korpusowy przytacza komplet wartości wymagań dla każdej z fraz składowych odpowiednich typów (dla fraz nominalnych w postaci oznaczenia przypadka, dla fraz przyimkowych — przyimka i symbolu przypadka itd.) wraz z ewentualnym indeksem frazeologicznym (symbol dolara przy oznaczeniu będącym realizacją frazeologizmu).

Solidną podbudową tej klasyfikacji stał się wcześniejszy projekt Świdzińskiego z lat 1988–90, zmierzający do stworzenia składniowego słownika czasowników polskich[5], nazywanego dalej SSCP. W wyniku ówczesnych prac powstała baza danych o rozmiarze 1494 rekordów zawierająca pełną informację syntaktyczną wybranych czasowników (ze słownika tego przejęta została do projektu korpusowego m. in. notacja i lista schematów zdaniowych). Osobno opisywane były czasowniki i quasi-czasowniki oraz czasowniki różniące się aspektem, np.

```
wychodzić 1 [Vi 6a]: 0; 11 C; 12 na B\$/do D/z D\$/poza B/za B\$/
                     od B/*w B; 15 Ps; 22\$ C+z D; 211 z D+Ps
wychodzić 2 [Vp 6a]: 11 B; 21 B+C
wychodzić 3 [Qi 6a]: 12 na B\$; 14 ZE&n B
...
wyjść 1     [Vi 12]: 0; 11 C; 12 na B\$/do D/z D\$/poza B/za B\$/
                     od B/*w B; 15 Ps; 22\$ C+z D; 211 z D+Ps
wyjść 2     [Qp 0]: 12\$ na B; 14 ZE&na B
```

W pracy o wypowiednikach[6] Świdziński analizuje dokładnie aspekt statystyczny korpusowego opisu wymagań w zakresie rządzących przyimków polskich oraz częstości wystąpień poszczególnych schematów zdaniowych, nie konfrontuje go jednak z wcześniejszą wersją składniowego słownika czasowników, stworzonego krańcowo odmienną metodą (poprzez ekstrakcję informacji słownikowej). W celu dokonania tego porównania stworzyłem minisłowniczek wymagań w kształcie zbliżonym do słownika składniowego, zawierający pełen wykaz wymagań zawartych w tekstach wypowiedników wraz z informacją o odpowiadającym danemu zestawowi wymagań schemacie zdaniowym.

Słownik wyekstrahowany z korpusu zawiera z oczywistych przyczyn znacznie uboższą informację niż rekordy SSCP:

```
wychodzić [Vi]: 0 m; 21 B+C
...
wyjść [Qp]: 22 C+z-N
wyjść [Vp]: 0 m; 12 na-B\$; 12 w-Ms
```

jednak z 1476 czasowników korpusowych aż 846 to czasowniki inne niż w SSCP, może więc on stanowić cenny materiał pozwalający na uzupełnienie słownika Świdzińskiego.

[5] Patrz [Świdziński, 1994b] oraz [Świdziński, 1996], s. 6–9.

[6] [Świdziński, 1996], s. 84–109.

4.3 Porządek linearny i ciągłość składników zdania elementarnego

Osobny rodzaj informacji dostępnej dla wypowiedników stanowią dane z zakresu uporządkowania linearnego składników zdania elementarnego[7] zapisane w parametrze `SZYK`. Na 6721 rekordów dane te zostały ustalone dla wszystkich 4824 zdaniowych wypowiedników elementarnych (71,8% ogólnej liczby wypowiedników) oraz dodatkowo dla jednego wypowiednika eliptycznego. Zapis wartości pola uwzględnia wzajemne położenie frazy finitywnej (oznaczanej symbolem `V`), podmiotowej (`S`) i fraz wymaganych (`O`), wraz z reprezentacją nieciągłości (części oznaczane są kolejno znakami apostrofu i cudzysłowu przy symbolu frazy). Obserwacje Świdzińskiego o przeważającej większości zdań „układu SVO", potwierdzające powszechne przekonanie o zaliczeniu polszczyzny do języków tego typu zostały wykorzystane m. in. w pracy [Derwojedowa, 2000].

Wnioski ze statystycznego rozkładu nieciągłości (354 zdania badanej wersji korpusu, czyli ponad 7%, zawierają nieciągłość) są równie ciekawe i wymagałyby dalszych badań. Przykładowo, dla fraz finitywnych często obserwowana jest nieciągłość „fleksyjna", wynikająca z rozdzielenia formy czasu przeszłego na aglutynant i pseudoimiesłów lub przerwania ciągłości czasownika z „się".

Zanotowany udział nieciągłości Świdziński uznaje za jeden z ciekawszych wyników. Trudno zgodzić się jednak z konkluzją[8]:

> Udało się stwierdzić, że konstrukcje nieciągłe, mimo że strukturalnie możliwe w językach z silnie rozbudowaną fleksją, są w polszczyźnie zjawiskiem tekstowo wyjątkowym.

gdyż osobiście nie uznałbym zjawiska występującego w 1/13 wyników za wyjątkowo rzadkie.

4.4 Próba ekstrakcji gramatyki z korpusu wypowiedników

Wątkiem pobocznym, ale stanowiącym dość ciekawy eksperyment, mogący sprawdzić dokładność opisu gramatycznego zawartego w korpusie wypowiedników, jest próba automatycznej konstrukcji „gramatyki korpusowej" — sformalizowanego opisu tekstów wypowiedników stworzonego przy maksymalnym wykorzystaniu zawartej w korpusie informacji gramatycznej. Próbę tę, której wynikiem jest zestaw klauzul Prologu[9], przeprowadziłem z założeniem akceptacji wszystkich (lub przeważającej większości) wypowiedników, lecz bez nakładania na konstrukcję gramatyki

[7] [Świdziński, 1996], s. 110–116.

[8] [Świdziński, 1993a], s. 23

[9] Języka programowania w logice, szczególnie przydatnego w zastosowaniach z dziedziny przetwarzania języka naturalnego. Przystępne wprowadzenie do języka Prolog zawiera np. pozycja [Gazdar i Mellish, 1989].

dodatkowych warunków nie wynikających z danych korpusowych, co miało na celu faktyczną weryfikację dokładności opisu.

W oczywisty sposób (poprzez konstrukcję jednostek terminalnych na bazie korpusowych jednostek leksykalnych o arbitralnej długości) tak wyekstrahowana gramatyka posługuje się jawnie przywołanym dalej pojęciem leksykonu w rozumieniu gramatyki formalnej języka polskiego Świdzińskiego (dalej GFJP, gramatyką tą zajmuję się w dalszej części pracy)[10]. Na leksykon ów składają się równoważniki dystrybucyjne terminali o stopniu złożoności wyznaczonym przez stopień szczegółowości opisu gramatycznego korpusu wypowiedników. Przydatność tak utworzonego opisu jest wyznaczona przez jego dokładność, przez co rozumiem zakres parametrów opisujących dany rekord, odróżniających go dystrybucyjnie od pozostałych rekordów leksykonu. Dla wykorzystania tak stworzonego opisu decydująca jest odpowiedź na pytanie, czy korpusowe definicje gramatyczne są wystarczające do zapewnienia rekordom owej dystrybucyjnej różnicy. Jest to warunek wstępny, którego weryfikacja warunkuje celowość dalszych prac nad leksykonem — i ekstrahowaną gramatyką.

Gdyby w ramach projektu korpusu wypowiedników przeprowadzono w sposób pełny weryfikację GFJP, odpowiedź na powyższe wątpliwości byłaby twierdząca, jednak, jak wspomniałem wcześniej, rodzaj informacji gramatycznej dostępnej w korpusie różni się od używanej w regułach GFJP. W szczególności, od opisu formalnie weryfikującego GFJP należałoby wymagać przede wszystkim informacji o numerach i parametrach reguł wykorzystywanych przy akceptacji danego wypowiednika. Mimo obszernego opisu każdej jednostki (patrz rozdział 1.3, s. 16) próbka korpusu nie zawiera np. informacji o uzgodnieniach parametrów frazowych ani nawet o ich zakresie, co uniemożliwia wykorzystanie tego opisu do konstrukcji pełnego opisu gramatycznego wykraczającego poza informację podstawową. Oczywiście nie oznacza to, że opis ten jest zupełnie nieprzydatny, gdyż bazową informację składniową możemy wydedukować z klasyfikacji schematów zdaniowych opisujących zakres i ogólne zasady uzgodnienia *verbum finitum* z frazą podmiotową i frazami wymaganymi, jest to jednak informacja niewystarczająca do uzyskania stopnia szczegółowości porównywalnego z tym z GFJP i aby uzyskać użyteczną gramatykę, należałoby wzbogacić ją o pewne warunki dodatkowe, tj. określone podstawy wiedzy o polskiej składni — przejęte choćby z gramatyki Świdzińskiego (np. co do szyku elementów zdania czy liczby fraz wymaganych).

W związku z założeniem o wykorzystaniu w eksperymencie wyłącznie informacji zawartej w korpusie, ograniczyłem konstrukcję gramatyki do jednostki, która wydaje się być opisana w sposób maksymalnie pełny, mianowicie zdaniowego wypowiednika elementarnego, czyli odpowiednika zdania elementarnego w terminologii GFJP. Model zdania elementarnego `ze` w powstałej gramatyce odpowiada liście schematów zdaniowych — zdanie zawiera frazę finitywną `ff` i odpowiadający danemu schematowi zestaw fraz wymaganych. Wstępnie ustalone parametry fraz wynikają z własności schematu: uzgodnienie frazy finitywnej z podmiotową (nominalną w mianowniku) w zakresie osoby, liczby i rodzaju implikuje co najmniej czteroparameterową definicję frazy nominalnej i co najmniej trójparametrową definicję frazy finitywnej (w rzeczywistości korpus zawiera jeszcze łączną informację o czasie, trybie, negacji i aspekcie frazy, która także została wykorzystana); opis schematów czasownikowych

[10]Patrz rozdział 9.2.2, s. 96.

2.3–2.6 w zakresie zgody lub braku uzgodnienia rodzaju i liczby wymaga dodania tych parametrów także do frazy przymiotnikowej i przyimkowej.

Oto przykładowy zestaw istotnych dla konstrukcji zdania elementarnego parametrów składniowych[11] jednego z najprostszych wypowiedników:

Parametr	Wartość parametru
ID	6721
TW	Z
WSP	*brak wartości*
ST	S
TYP	*brak wartości*
KL	V
ASP	i
CHAR	3p,te
NEG	T
SCH	16
TEKST	<Ma> [on] (być gościem prezydenta Saragata).

i odpowiadający mu fragment wygenerowanej gramatyki zapisany w formalizmie DCG:

```
ff(trzecia, pojedyncza, \_, terazniejszy, oznajmujacy, niedokonany,
   nie) --> ['Ma'].
fno(mianownik, trzecia, pojedyncza, \_) --> ['on'].
fwb --> ['być', 'gościem', 'prezydenta', 'Saragata'].
ze  --> ff(trzecia, pojedyncza, Osoba, Czas, Tryb, Aspekt, Negacja),
        fno(mianownik, trzecia, pojedyncza, Osoba), fwb, ['.'].
```

W powstałej gramatyce koncepcja leksykonu realizuje się na poziomie definicji fraz, gdyż taką granulację osiąga anotacja korpusowa. Liśćmi drzew rozbioru stają się wyekstrahowane z tekstów wypowiedników ciągi odpowiadające tekstowej wartości fraz odpowiednich rodzajów. Zawartość leksykonu można łatwo badać zadając zapytania w języku Prolog, np. zapytanie

```
ff(trzecia, pojedyncza, \_, terazniejszy, oznajmujacy, niedokonany,
   nie, X, []).
```

daje w wyniku wszystkie zarejestrowane frazy finitywne o podanej charakterystyce (uzależnionej ściśle od opisu korpusowego, o czym świadczy postać ostatniego przytoczonego przykładu — postać fraz poszczególnych typów dyskutuję ponadto w podrozdziałach rozdziału 8.1 na s. 75):

[11] Za nieistotny dla konstrukcji tak ustalonego podzbioru gramatyki uznaję np. parametr wskazujący centrum struktury oraz jego postać hasłową czy informację o szyku, uzyskiwaną poprzez analizę postaci tekstowej wypowiednika.

```
X = [ma] ;
X = [podaje] ;
X = [występuje] ;
X = [znajduje, się] ;
X = [ma, kształt, trójkąta, o, wypukłych, bokach, i, wiruje,
     wewnątrz, cylindra, o, kształcie, spłaszczonego, walca].
...
```

Posługując się leksykonem można także wykorzystać powstały mechanizm do generowania poprawnych zdań elementarnych zbudowanych z fraz korpusowych odpowiednich typów, np.

```
X = [ma, on, wprowadzić, na, wody, kanału, sueskiego, uzbrojoną,
     łódź, ',', holownik, i, zbiornikowiec, '.'].
```

Mimo zadowalającego wyniku próby konstrukcji gramatyki dla zdań elementarnych, zakres informacji korpusowej dla zdań złożonych jest już niewystarczający — brak m. in. informacji o pytajności składników czy dodatkowych warunkach składniowych, silnie reprezentowanych w GFJP i niezbędnych dla poprawnego opisu polszczyzny, wobec czego dalsze prace w tym kierunku wydawały się ślepą uliczką i nie były prowadzone.

Część II

Wstępne etapy weryfikacji

Rozdział 5

Weryfikacja warstwy typograficznej

Inicjalny etap weryfikacji stanowiło sprawdzenie typograficznej warstwy korpusu, tj. przede wszystkim jakości oryginalnego podziału tekstów próbek na jednostki zbliżone do wprowadzanych w artykule [Bień i Saloni, 1982] *wyrazów grafemicznych*: segmentów tekstu jednowymiarowego wyodrębnianych za pomocą spacji, znaków interpunkcyjnych i łączników. W związku z tym, że tekst wypowiedników zawiera oprócz jednostek wyrazowych także oznaczenia dodatkowe, do etapu weryfikacji typograficznej zaliczam także sprawdzenie ich kompletności i jakości. Zadaniem osobnym, choć zbliżonym pod względem logicznym i technicznym, było także sprawdzenie zgodności opisu całych próbek (nie tylko zawartych w nich tekstów) z przewidzianym zestawem wartości parametrów ilościowych i jakościowych.

Powodów wystąpienia przeważającej liczby problemów tego rodzaju należy upatrywać w swobodnym podejściu do tworzenia korpusu. Jak pisze Skibicki[1]:

> W trakcie analizy korpusu WSWP podstawowym problemem, który utrudniał automatyzację pracy, była duża liczba błędów. Ich źródło leży w sposobie wprowadzania informacji do bazy korpusu. Proces nie był wspierany programem, który proponowałby poprawne oznaczenia w oparciu o rozszerzalne słowniki dopuszczalnych oznaczeń czy właściwe oznaczanie fraz w tekście.

W świetle tego faktu konieczne wydawało się dokonanie weryfikacji własności typograficznych tekstów korpusu — jak się okazało, niebezpodstawne, gdyż stosunek liczby wypowiedników zawierających opisane niżej usterki (897) do łącznej liczby rekordów w korpusie okazał się bardzo wysoki i wyniósł aż 13,35%. Problemy typograficzne zostały wykryte przez napisane przeze mnie programy w języku Perl[2].

Dodatkowo w procesie weryfikacji wykorzystałem możliwość porównania zasobów korpusu wypowiedników z dwoma innymi korpusami utworzonymi w oparciu o ten sam zestaw danych.

[1] Patrz [Skibicki, 2000], s. 31.
[2] Patrz [Wall i in., 2001].

5.1 Poprawność typograficzna tekstu próbek

Eliminacja usterek typograficznych wydaje się podstawowym zadaniem warunkującym poprawne przetwarzanie korpusu. Do błędów typograficznych zaliczyłem przede wszystkim niestandardowe cechy tekstu utrudniające przetwarzanie, a zatem nieprawidłowe oznaczenia interpunkcyjne czy niestandardowe spacjowanie. Oto pełny wykaz rodzajów zaobserwowanych (i poprawionych) błędów wraz z liczbą wystąpień każdego z ich rodzajów:

- występowanie zbędnego tekstu (nietypowy znak na końcu próbki, informacje lokalizacyjne w treści próbki) — 10 wystąpień,
- brak oznaczenia znaku nie należącego do wypowiednika (występującego w treści samodzielnie, a zaburzającego postać tekstową — np. wiodącego dwukropka) — 34 wystąpienia,
- błędna postać wielokropka (dwie lub cztery kropki) — 46 wystąpień,
- występowanie znaków interpunkcyjnych nie należących do frazy w jej wnętrzu lub znaków przynależnych frazie poza jej obrębem — łącznie 113 wystąpień,
- brak lub błędny znak interpunkcyjny na końcu wypowiednika — 129 wystąpień,
- brak spacji — po znaku interpunkcyjnym (oddzielającej go od tekstu lub granicy frazy, oddzielającej numer frazy od jej treści, rozdzielającej frazy) — 104 wystąpienia,
- nadmiarowa spacja (przed znakiem interpunkcyjnym, po łączniku, między znakiem granicy frazy a jej numerem lub tekstem) — 105 wystąpień,
- błędnie umieszczona spacja (np. przed zamiast po znaku interpunkcyjnym) — 26 wystąpień.

Szczegółowy wykaz numerów próbek zawierających błędy wymienionych kategorii zamieszczam w rozdziale C.1 (s. 197).

5.2 Poprawność oznaczeń struktury frazowej

Jako błędne traktuję oznaczenia składników frazowych nie uwzględnione w instrukcji [Świdziński, 1994a] przekazanej zespołowi uczestniczącemu w pracach nad tworzeniem korpusu wypowiedników. Są to przede wszystkim nieprawidłowe nawiasowanie oraz niestandardowe oznaczenia fraz, uniemożliwiające przetwarzanie bez ręcznej ingerencji w tekst próbki; oto szczegółowa lista kategorii błędów:

- ujęcie znaku w nawiasy właściwe frazie — 40 wystąpień,
- błędne oznaczenie granic frazy (odwrócony lub błędny nawias, brak nawiasu, nawias nadmiarowy) — 35 wystąpień,
- brak lub niepełne oznaczenie nieciągłości frazy — 9 wystąpień,
- błędna numeracja frazy (brakujące oznaczenie numeru wymaganego, nadmiarowy numer frazy pojedynczej, niezgodność numerycznych oznaczeń początku i końca frazy, nieciągłość numeracji fraz, błędne oznaczenie frazy nieciągłej) — 16 wystąpień.

Wykaz numerów próbek zawierających błędy struktury frazowej zamieszczam w rozdziale C.2 (s. 199).

W oryginalnej wersji tekstów wypowiedników pojawia się także nie udokumentowane ani przez Świdzińskiego, ani przez Skibickiego oznaczenie `@`[3]. Jego nierównomierny rozkład (102 wystąpienia, od próbki 2251 do 3084) pozwala przypuszczać, że symbol ten mógł być również wprowadzony na etapie obróbki tekstów przez jednego z edytorów; na potrzeby dalszego przetwarzania został on usunięty z tekstów korpusu (z zachowaniem wersji archiwalnej).

5.3 Poprawność opisu parametrów

Opis parametrów jakościowych korpusu okazał się, poza charakterystyką fraz wymaganych, w dużej mierze zgodny z zestawem przewidzianych wartości. Zaobserwowane odstępstwa ograniczyły się do:

- niezgodności wartości parametru `OPIS` z faktyczną liczbą fraz wymaganych — 185 wypowiedników,
- błędnej klasyfikacji wypowiednika — 2 wypowiedniki,
- błędnego oznaczenia schematu — 1 wypowiednik.

Bardzo duża częstość błędów w opisie fraz wymaganych zdaje się pochodzić z jednej strony ze zwyczaju mechanicznego kopiowania odnośnego parametru z wypowiednika nadrzędnego do podrzędnego, z drugiej — z niezrozumienia sposobu zapisu informacji o wymaganiu (edytor podawał wymaganie dla schematu, a nie dla jego konkretnej realizacji w danym wypowiedniku). Przykładowo, dla wypowiednika z tekstem:

(29) *<Zastanów się>.* *[1167]*

na mocy ogólnego wymagania dla czasownika *zastanowić się* parametr `OPIS` przyjmuje wartość `nad-N`. Usterki tego rodzaju (jako jedyne) nie zostały poprawione.

Weryfikacji parametrów ilościowych, jako nie należących do przedmiotu badawczego niniejszej pracy, nie przeprowadzałem w ogóle.

Numery próbek zawierających niepoprawne parametry opisu zamieszczam w rozdziale C.3 na s. 200.

5.4 Konfrontacja korpusu wypowiedników z innymi wersjami korpusu źródłowego

Jak już wspomniałem w rozdziale 1.2 (s. 15), źródłem danych dla korpusu wypowiedników był bezpośrednio korpus słownika frekwencyjnego polszczyzny współczesnej,

[3] Oznaczenie to nie ma związku z wartością parametru `SCH`, w którego treści również może pojawić się symbol `@` na oznaczenie schematu nie przewidzianego w zestawie (patrz rozdział 4.1, s. 39) ani z wspominanym przez Skibickiego zwyczajem zapisywania liczb dwucyfrowych w wartościach parametrów `WYP` i `ZD` w postaci `@n` dla zapewnienia poprawności sortowania — patrz [Skibicki, 2000], s. 12.

obecnie dostępny w wersji rozszerzonej pod nazwą *korpusu polszczyzny lat sześćdziesiątych XX wieku*[4]. Niedawno jedna z wersji tego korpusu została także wykorzystana w pracach nad obszernym korpusem przygotowanym w Instytucie Podstaw Informatyki PAN z uwzględnieniem etapu ręcznej korekty (zwanym oficjalnie *korpusem IPI PAN*[5]). Dostępność trzech pokrewnych zbiorów danych, z których każdy zawiera dokładną informację lokalizacyjną w odniesieniu do źródła pozwoliła na konfrontację zasobów i porównanie ich postaci tekstowych w celu wykrycia i poprawienia nieoczywistych błędów w opisie wynikających z przeoczeń edytorów, a polegających na:

- zamienionej kolejności wyrazów — 1 wystąpienie,
- urwaniu tekstu wypowiednika — 3 wystąpienia,
- niezgodności tekstów w wypowiednikach zależnych — 2 wystąpienia.

Listę odnośnych numerów próbek zamieszczam w rozdziale C.4 (s.201).

[4]Patrz `http://www.mimuw.edu.pl/polszczyzna/pl196x/`.

[5]Patrz [Przepiórkowski, 2004] i `http://www.korpus.pl`. Udostępniana w IPI wersja korpusu słownika frekwencyjnego nosi nazwę *Korpusu Słownika Frekwencyjnego (IPI PAN)*.

Rozdział 6

Weryfikacja warstwy morfologicznej

Kolejnym krokiem weryfikacyjnym było sprawdzenie, czy stosowany analizator morfologiczny akceptuje materiał leksykalny korpusu wypowiedników, czego długofalowym celem było spełnienie warunku koniecznego do dokonania analizy składniowej wykorzystującej ów komponent morfologiczny. Dodatkowe możliwości stworzyła także analiza warstwy morfologicznej wspomnianych w poprzednim rozdziale korpusów pokrewnych.

6.1 Morfeusz — narzędzie analizy morfologicznej

W kontekście zadania weryfikacji warstwy morfologicznej korpusu interesująca może się wydawać choćby tylko odpowiedź na pytanie najbardziej podstawowe: czy dana forma jest w ogóle rozpoznawana przez analizator morfologiczny. Opisując dalej wykorzystane narzędzie staram się jednak nie tracić z oczu perspektywy jego wykorzystania w procesie analizy składniowej, dyskutuję zatem dość szczegółowo kwestię zadania analizy morfologicznej oraz sposobu jego realizacji w opisywanym narzędziu, skupiając się na koniec na rzeczywistej weryfikacji warstwy morfologicznej korpusu.

6.1.1 Zadanie analizy morfologicznej

Analizą morfologiczną nazywam proces przyporządkowania jednostkom tekstu (słowom, segmentom) pewnej *charakterystyki morfologicznej* rozumianej jako zestaw cech słownikowych wystarczających do opisu dystrybucji danej jednostki. Zakres tej charakterystyki nie jest zdefiniowany ściśle, zatem o ile proces analizy morfologicznej jest zadaniem zamkniętym, wyznaczają go zwykle tradycyjnie rozumiane kategorie fleksyjne. Jeśli jednak analiza morfologiczna jest wstępem do analizy składniowej, zakres ten winien być odzwierciedleniem stopnia szczegółowości zadanego przez gramatykę używaną w procesie analizy składni.

W przypadku wieloznaczności interpretacji danej jednostki w ramach przyjętych reguł opisu proces ma oczywiście podać wszystkie możliwe interpretacje. Często

zadaniem dodatkowym analizy morfologicznej jest także podanie postaci hasłowej analizowanego słowa, rozumianej zwykle jako słownikowy identyfikator zawierającego dane słowo *leksemu* — zbioru form o zbliżonej charakterystyce semantycznej.

Tak zdefiniowane zadanie może być wykonane przez program komputerowy — i właśnie w znaczeniu programu służącemu określaniu charakterystyki słów zgodnej z ustalonym opisem będę dalej używać wyrażenia *analizator morfologiczny*. Analizatorem takim jest w szczególności program Morfeusz Marcina Wolińskiego.

6.1.2 Jednostki analizy

Morfeusz definiuje *słowa* jako ciągi liter oddzielone spacjami, bez przylegających do nich ewentualnych znaków przestankowych (z wyłączeniem apostrofu i łącznika oraz kropki stanowiącej część skrótu, które również traktuje jako wchodzące w skład słowa).

Wewnątrz słów może dodatkowo istnieć podział na *segmenty*, definiowany na granicy form aglutynacyjnych:

- leksemu `być` dla czasu przeszłego czasowników (np. *znałeś* = *znał* + *-eś*),
- partykuło-przysłówka warunkowego `by` w formach trybu warunkowego (*miałby*),
- postaci zaimka osobowego `on` (*doń*, *zeń*),
- partykuły wzmacniającej `-że` i `-ż` (*chodźże*, *chodźcież*) lub `-ć` (*byłać*),
- leksemu pytającego `-li` (*znaszli*).

Jako osobne segmenty traktowane są także znaki interpunkcyjne oraz każdy człon form pisanych z łącznikiem (osobnym segmentem jest w tym przypadku także łącznik).

W przypadku słów złożonych z segmentów analizie morfologicznej poddawany jest każdy segment, a wynikiem analizy słowa jest ciąg analiz poszczególnych segmentów.

6.1.3 Źródło danych

Podstawą algorytmu działania analizatora Morfeusz (jak kilku innych analizatorów morfologicznych, także komercyjnych) jest *Schematyczny indeks* a tergo *polskich form wyrazowych* [Tokarski, 2002], powstały w wyniku prac nad morfologią haseł w 11-tomowym słowniku Doroszewskiego. Jego obecna wersja została uzupełniona już po śmierci autora przez Zygmunta Saloniego i wydana pod jego redakcją.

Indeks (patrz rys. 6.1) ma postać listy reguł – rekordów, z których każdy składa się (w przeważającej większości przypadków, gdyż indeks zawiera także pewne skróty notacyjne) z czterech pól:

- nagłówka stanowiącego wzorzec końcówki opisywanego danym rekordem zestawu słów,
- listy symboli kategorii morfologicznych odpowiadających danej analizie,
- końcówki postaci hasłowej odpowiadającej danemu słowu,

- przykładów ilustrujących daną regułę, często wraz z liczbą odpowiadających jej haseł słownika Doroszewskiego.

Reguły ułożone są w porządku *a tergo* ze względu na pole nagłówkowe.

-mam	*mIV N*		**imam, omam**
mam	*żIV lG*	**mama**	
-mam	*I 1*	**-mać**	**mniemam, imam, dumam, trzymam (70)**
mam	*VIa i*	**mamić**	**omam (4)**
(mam	*XII 1*	**mieć**	
(nam	*Za D*	**my**	
-nam	*mIV N*		**ignam, Uznam**
-nam	*nIII lG*	**-namo**	**dynam**
-nam	*żIV lG*	**-nama**	**panam, izodynam!**
-nam	*I 1*	**-nać**	**żegnam, zapinam, przekonam, zrzynam**
-nam	[2 formy]	**-na +)m**	**winnam, powinnam (*patrz:* Vm)**
-pam	*I 1*	**-pać**	**ćpam, stąpam (15)**
-ram	*mIV N*		**program, chram, kram, tram (85)**
ram	*żIV lG*	**rama**	**bram, dram, panoram, szram (13)**
-ram	*I 1*	**-rać**	**(się) staram, ścieram, otwieram, gram**
-ram	*VIa i*	**-ramić**	**obram, poszram**
(sam			**sam(f)**

Rysunek 6.1: Fragment indeksu Tokarskiego

Proces analizy polega (w wielkim skrócie) na odszukaniu w indeksie wiersza (wierszy) zawierającego w pierwszej kolumnie część końcową analizowanego słowa, weryfikacji słownikowej potencjalnej formy hasłowej utworzonej dzięki informacji odczytanej z trzeciej kolumny tego wiersza i po potwierdzeniu jej istnienia zwrócenia jako wyniku charakterystyki morfologicznej odczytanej z drugiej kolumny. Tak zdefiniowany wykaz reguł odmiany słów polskich nadaje się zatem świetnie do wykorzystania w algorytmie automatycznej analizy morfologicznej, co nastąpiło już w roku 1993 dzięki dostępności komputerowej wersji indeksu powstałej na potrzeby składu jego wydania książkowego.

6.1.4 Taksonomia

Przyjęta przez twórcę Morfeusza taksonomia stawia sobie za główne zadanie zapewnienie jednoznaczności morfologicznej tekstu, modeluje zatem wyłącznie własności służące temu celowi, a nie komplet informacji słownikowych. Konsekwencją tej koncepcji jest np. zaliczenie jednostek tradycyjnie określanych jako liczebniki porządkowe do kategorii przymiotników — ze względu na sposób ich odmiany. Taksonomia ta rozszerza znacznie opis Tokarskiego, m. in. z wykorzystaniem danych Saloniego [Saloni, 2001].

Poniżej podaję na podstawie [Woliński, 2003][1] zwięzły opis używanej przez analizator taksonomii:

[1]Por. rozdziały 5 i 6 s. 7–14.

Kategoria	Zbiór wartości
klasa gramatyczna (fleksem)	rzeczownik (`subst`) forma deprecjatywna rzeczownika (`depr`) przymiotnik (`adj`) przymiotnik przyprzymiotnikowy (`adja`), np. *polsko* w *polsko-suahilijski* przymiotnik poprzyimkowy (`adjp`), np. *polsku* w *po polsku* przysłówek stopniowalny (`adv`) zaimek przysłowny (`padv`), np. *dokąd*, *tam* liczebnik (`num`) — tylko liczebniki główne i zbiorowe zaimek nietrzecioosobowy (`ppron12`) — *ja*, *ty*, *my* i *wy* zaimek trzecioosobowy ON (`ppron3`) zaimek SIEBIE (`siebie`) nieprzeszła forma finitywna czasownika (`fin`) przyszła forma finitywna czasownika BYĆ (`bedzie`) forma aglutynacyjna czasownika BYĆ (`aglt`) pseudoimiesłów (`praet`) rozkaźnik (`impt`) bezosobnik (`imps`), np. *łgano* bezokolicznik (`inf`) imiesłów przysłówkowy współczesny (`pcon`) imiesłów przysłówkowy uprzedni (`pant`) odsłownik (`ger`) imiesłów przymiotnikowy czynny (`pact`) imiesłów przymiotnikowy bierny (`ppas`) czasownik typu WINIEN (`winien`) czasownik niewłaściwy (`pred`), np. *dość*, *warto* przyimek (`prep`) spójnik (`conj`) partykuło-przysłówek (`qub`) znak interpunkcyjny (`interp`)
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`), np. *dziecko* nijaki zwykły (`n2`), np. *okno*, *co* przymnogi osobowy (`p1`), np. *wujostwo*

Kategoria	Zbiór wartości
	przymnogi zwykły (`p2`), np. *skrzypce* przymnogi opisowy (`p3`), np. *spodnie*
osoba	pierwsza (`pri`) druga (`sec`) trzecia (`ter`)
stopień	równy (`pos`) wyższy (`comp`) najwyższy (`sup`)
aspekt	niedokonany (`imperf`) dokonany (`perf`)
negacja	niezanegowana (`aff`), np. *pisanie*, *czytanego* zanegowana (`neg`), np. *niepisanie*, *nieczytanego*
deprecjatywność	niedeprecjatywna (`ndepr`) deprecjatywna (`depr`)
akcentowość	akcentowana (`akc`), np. *jego*, *niego* nieakcentowana (`nakc`), np. *go*, *-ń*
poprzyimkowość	poprzyimkowa (`praep`), np. *niego*, *-ń* niepoprzyimkowa (`npraep`), np. *jego*, *go*
akomodacyjność	uzgadniająca (`congr`), np. *dwaj* rządząca (`rec`), np. *dwóch*, *dwu*
aglutynacyjność	nieaglutynacyjna (`nagl`), np. *niósł*, *dlaczego* aglutynacyjna (`agl`), np. *niosł-*, *dlaczegó-*
wokaliczność	wokaliczna (`wok`), np. *-em*, *ze* niewokaliczna (`nwok`), np. *-m*, *z*

6.1.5 Notacja

Wynik analizy w postaci charakterystyki morfologicznej pojedynczego segmentu jest w ogólnym przypadku zbiorem wartości kategorii odpowiadających danej formie. Dla wygody zbiory te prezentowane są w postaci listy par złożonych z postaci hasłowej analizowanej formy i listy pól kategorii gramatycznych. Zawartość pola kategorii stanowi z kolei lista wartości danej kategorii (w szczególności pojedyncza wartość), dla których dana postać hasłowa realizuje się jako analizowany napis.

Oto przykładowy wynik analizy formy *kurze* (pola kategorii oddzielone są dwukropkami, listy wartości — ujęte w nawiasy kwadratowe i oddzielone przecinkami):

Postać hasłowa	Analiza
kur	`subst:sg:[loc,voc]:m2`
kura	`subst:sg:[dat,loc]:f`
kurz	`subst:pl:[nom,acc]:m3`
kurzy	`adj:sg:[nom,acc]:[n1,n2]:pos,` `adj:pl:[nom,acc]:[m2,m3,f,n1,n2,p2,p3]:pos`

6.2 Analiza morfologiczna form wyrazowych

Korpus wypowiedników zawiera łącznie 13929 różnych form wyrazowych w 63103 wystąpieniach. Za pojedyncze wyrazy uznaję na potrzeby dodatkowej analizy morfologicznej także formy wieloczłonowe w rodzaju *cybernetyczno-geometryczno-kosmogoniczna*.

Form nie akceptowanych przez Morfeusza jest 811 (5,82% wszystkich form). Oto ich zgrubna klasyfikacja:

Rodzaj	Przykłady	Liczba	Udział %
nazwy własne	`Atlantyda`, `Świeradowa`, `Younga`	622	76,70 %
wyrazy z doklejonym aglutynantem	`byleś`, `doń`, `wyście`	9	1,11 %
wyrazy z doklejoną partykułą	`mówże`, `weźże`	6	0,74 %
formy złożone	`cukrowo-fosforanowe`, `blond-kawalerze`	10	1,23 %
wyrazy obcojęzyczne	`Danke`, `bezuggschein`	14	1,73 %
neologizmy	`informelu`, `bladawiec`	4	0,49 %
inne formy nie rozpoznawane przez Morfeusza	`encefalograf`, `ideologijki`	146	18,00 %

W tabeli nie uwzględniam wykrytych na etapie analizy morfologicznej i poprawionych niezależnie nie rozpoznanych form zawierających błędy czysto typograficzne. Do tego rodzaju usterek zaliczam:

- literówki[2] — 282 wystąpienia,
- użycie małej litery w miejscu wielkiej — 1 wystąpienie,
- sklejone wyrazy — 11 wystąpień.

Osobno sprawdzona została poprawność morfologiczna wartości parametrów reprezentujących cechy jakościowe z grupy „morfologicznej”, mianowicie tekstowe opisy centrum i hasła próbki.

Wykaz wystąpień tego rodzaju usterek zawiera rozdział C.5 (s.202).

[2]Za błąd nie uznaję specyficznej pisowni niektórych nazw (*Algerii* zamiast *Algierii*, *Kostarice* zamiast *Kostaryce*, ale także *Atylli* zamiast *Attyli*, *Fortynbasie* zamiast *Fortynbrasie*).

6.3 Uzupełnienie kodów morfologicznych

Etapem następującym bezpośrednio po wykryciu form poprawnych, a nie akceptowanych przez analizator morfologiczny, stało się uzupełnienie brakujących kodów morfologicznych. Zadanie to mogło zostać wykonane z ponownym wykorzystaniem danych korpusu słownika frekwencyjnego (IPI PAN) poprzez konfrontację wyników analizy morfologicznej dla tekstów próbek z opisem odnośnych rekordów w korpusie IPI. Mimo odmiennego formatu zapisu danych strukturalizacyjnych, taksonomie morfologiczne obu zbiorów są identyczne ze względu na wykorzystanie tego samego analizatora morfologicznego, a w związku z zastosowanym procesem statystycznej dezambiguacji jego wyników operacja ta mogła przyczynić się do pozyskania większej liczby kodów gotowych do bezpośredniego wykorzystania.

W opisany sposób pozyskano jednak tylko kody 67 form — okazało się, że większość „trudniejszych" wyrazów nie została ujednoznaczniona i posiadała kody ciał obcych (nominalnych — `xxs` lub luźnych — `xxx`). Uzyskane analizy wykorzystywałem w całości, przenosząc zawarte w nich wieloznaczności do docelowych opisów form. Warto także wspomnieć, że niekiedy — zazwyczaj już na etapie właściwej analizy składniowej — okazywało się, że także te formy wymagały poprawienia ze względu na błędny kod lub postać hasłową (dla ilustracji podaję zapis w oryginalnym formacie korpusowym)[3]:

```
<orth>Klonowskim</orth>
<lex disamb="1">
  <base>klonowski</base>
  <ctag>adj:sg:loc:m3:pos</ctag>
</lex>
<orth>McNamary</orth>
<lex disamb="1">
  <base>mcnamary</base>
  <ctag>subst:sg:gen:m1</ctag>
</lex>
```

Reszta kodów została przeze mnie dopisana ręcznie zgodnie z używanym przez analizator składniowy zestawem znaczników. Kategoryzacja form nie sprawiła większych trudności.

[3]Formatem reprezentacji korpusu IPI jest XCES [Ide i in., 1996] stanowiący XML-ową wersję SGML-owego standardu CES [Ide i in., 2000] opartego na wspominanym już wcześniej schemacie TEI.

Część III

Weryfikacja składniowa

Rozdział 7

Gramatyka formalna języka polskiego i Świgra

W ogólnym schemacie procesu analizy składniowej danego wypowiedzenia możemy wyróżnić trzy zasadnicze etapy: podział wypowiedzenia na składniki wchodzące ze sobą w relacje składniowe, analizę morfologiczną tych składników oraz właściwą analizę składniową na bazie wyników poprzednich etapów, z wykorzystaniem danej gramatyki formalnej. Dwie ostatnie operacje mogłyby się w zasadzie dokonywać na poziomie samej gramatyki skojarzonej z komponentem słownikowym[1], jednak wydzielenie analizy morfologicznej z procesu analizy składniowej umożliwia jej realizację za pomocą prostszych, a w zupełności wystarczających środków.

Opisywany analizator składniowy Świgra — wynik pracy doktorskiej Marcina Wolińskiego [Woliński, 2004] — korzysta z gramatyki Świdzińskiego oraz wspominanego już wcześniej modułu Morfeusz, odpowiedzialnego zarówno za segmentację, jak i za analizę morfologiczną.

7.1 Gramatyka formalna języka polskiego

Gramatyka formalna języka polskiego Marka Świdzińskiego (dalej GFJP lub gramatyka Świdzińskiego) to obecnie najobszerniejszy znany formalny opis polszczyzny. Ma on postać wielopoziomowego katalogu jednostek składniowych wraz z opisem ich realizacji, spajających konstrukcje zdaniowe, frazowe i jednostki odpowiadające szkolnym „częściom mowy".

7.1.1 Historia

Początek prac nad formalnym opisem składni polskiej w środowisku warszawskim datuje się na połowę lat 70-tych, a wyznaczają go artykuł Zygmunta Saloniego [Saloni, 1974] i jego praca habilitacyjna [Saloni, 1976]. W duchu tych prac Saloni

[1] W taki właśnie sposób, wobec braku działającego analizatora morfologicznego, definiowany jest styk jednostek funkcyjnych gramatyki z zawartością leksykonu słów zarówno u Szpakowicza, jak i u Świdzińskiego — poprzez wyliczenie form wraz z ich analizami.

prowadził także od 1977 roku w Instytucie Informatyki Uniwersytetu Warszawskiego seminarium „Formalny opis języka naturalnego", w którym brali udział zarówno lingwiści, jak i informatycy zainteresowani stworzeniem sformalizowanego opisu gramatycznego współczesnej polszczyzny.

Jednym z bardziej znaczących wydarzeń tego okresu, a także bezpośrednim źródłem inspiracji dla gramatyki Świdzińskiego, była praca doktorska jednego z uczestników seminarium – Stanisława Szpakowicza [Szpakowicz, 1978], opublikowana później w nieco zmienionym kształcie jako [Szpakowicz, 1986] i prezentująca opis obszernego podzbioru języka polskiego (a ponadto zawierająca prototyp programu komputerowego dokonującego analizy składniowej wypowiedzeń polskich zgodnie z zaproponowaną gramatyką).

W latach 80-tych praca nad formalnym modelem polszczyzny była kontynuowana przez Marka Świdzińskiego, początkowo wspólnie ze Szpakowiczem, następnie już bez jego udziału. Jej wynikiem jest nowy opis gramatyczny zawarty w pracy habilitacyjnej Świdzińskiego [Świdziński, 1987], opublikowanej później w wersji książkowej [Świdziński, 1992a]. Praca ta wyznacza zasadnicze ramy opisywanej tu gramatyki, gdyż można przyjąć, że od czasu jej powstania mimo wielu korekt ogólne założenia GFJP pozostały niezmienione.

7.1.2 Koncepcja i notacja

Gramatyka Świdzińskiego opiera się na koncepcji analizy na składniki bezpośrednie. Wynik analizy daje się zatem przedstawić w postaci sparametryzowanych (wartościami zebranymi podczas analizy i wykorzystywanymi w jej trakcie) drzew składników.

Do realizacji tego celu GFJP wykorzystuje formalizm gramatyki metamorficznej wprowadzony przez Alaina Colmerauera [Colmerauer, 1978]. Formalizm ten jest obecnie szerzej znany jako *Definite Clause Grammar (DCG)* i jest standardowo dostępny w implementacjach Prologu — języka programowania w logice. Gramatykę w postaci DCG tworzą reguły z lewą stroną w postaci pojedynczego symbolu nieterminalnego (i ew. dowolnej liczby symboli terminalnych) a prawą stroną pustą albo zawierającą dowolny ciąg terminali, nieterminali lub dodatkowych warunków.

Oto przykład reguły GFJP w używanej obecnie notacji[2]:

```
zr(Wf, A, C, T, Rl, O, Neg, I, Z)
--> zsz(Wf, A, C, T, Rl, O, Neg, I, Z1),
    { oblnp(Z1, Z) },
    przec,
    spoj(rc, Oz, ni),
    zsz(Wf1, A1, C1, T1, Rl1, O1, Neg1, ni, p).
```

[2]Tzw. edynburskiej, różniącej się od oryginalnie stosowanej przez Świdzińskiego notacji marsylskiej m. in. odwrotnym stosowaniem wielkich i małych liter na oznaczenie stałych i zmiennych; po szczegóły odsyłam do pracy [Woliński, 2004], s. 28–29.

Reguła definiuje zdanie równorzędne (`zr`) jako sekwencję dwu zdań szeregowych (`zsz`) rozdzielonych przecinkiem (`przec`) i spójnikiem równorzędnym centralnym (`spoj` o wartości parametru typu równej *rc*). Charakterystyka zdań opisana jest wartościami kategorii fleksyjnych i składniowych reprezentowanych przy użyciu parametrów danej jednostki[3].

W postaci oryginalnej, opisanej w pracy [Świdziński, 1992a], gramatyka Świdzińskiego zawiera 461 reguł, z czego 149 to reguły „quasi-słownikowe", definiujące jednostki funkcyjne („części mowy" i znaki interpunkcyjne).

7.1.3 Metoda

Gramatyka Świdzińskiego powstała jako model kompetencji jej autora, bez podbudowy korpusowej. Autor uzasadnia tę metodę pracy przypadkowością danych zawartych w każdym, nawet bardzo obszernym korpusie, problemami z oceną reprezentatywności tak zgromadzonych danych oraz przede wszystkim większą przydatnością do tego celu opisu słownikowego (pośrednio odwołującego się wprawdzie do korpusu stanowiącego bazę dla twórcy słownika). To właśnie słownik staje się dla autora źródłem informacji o własnościach składniowych leksemów, z których dają się wywnioskować cechy dystrybucyjne jednostek wyższego rzędu. Kompletność tak tworzonego opisu ma zapewnić intuicja autora[4].

Każdy rodzaj opisywanej konstrukcji składniowej ilustrowany jest przykładami zdań akceptowanych i nieakceptowanych przez podany opis. Wobec braku korpusu przykłady te są preparowane, co zgodnie z intencjami autora precyzyjnie ukazuje charakter problemu, z drugiej jednak strony skłania ku nadmiernej szczegółowości opisu, rejestrującej zjawiska w języku niszowe, przejawiające się generowaniem zdań, co do których nawet określenie ich poprawności okazuje się trudne.

Z racji rozmiarów gramatyki ilustracja konstrukcji jest silnie kontekstowa, a stwierdzenie, czy przykład oznaczony jako nieakceptowany jest faktycznie odrzucany (gdyż wyrażenia w oczywisty sposób nieakceptowane przez daną regułę mogą zostać zaakceptowane „kuchennymi drzwiami" — przez inny zestaw reguł) bez wsparcia komputerowego wydaje się prawie niemożliwe.

7.1.4 Zakres i stopień ogólności

GFJP jest gramatyką opisującą składnię zdań złożonych z centrum finitywnym lub spójnikowym. Zgodnie z opisem fraza finitywna może być realizowana jako *forma osobowa lub bezosobnik czasownika, forma finitywna czasownika niewłaściwego, forma bezokolicznikowa czasownika lub czasownika niewłaściwego, a także równoważnik dystrybucyjny wymienionych form*[5].

[3] Patrz także informacje w rozdziale 6.1.5, s. 57.

[4] Oczywiście jest to sytuacja modelowa; sam autor przyznaje się do konsultacji z opisami zawartymi w dostępnych gramatykach języka polskiego, jak również analizy materiału empirycznego z pracy Polańskiego o składni w języku górnołużyckim — patrz [Świdziński, 1992a], s. 26–27.

[5] Za [Świdziński, 1992a], s. 21.

Świdziński wymienia następujące konstrukcje nie należące do opisu[6]:

- zdania z centrum spójnikowym, których drugi składnik zdaniowy nie zawiera frazy finitywnej (w korpusie wypowiedników oznaczane jako równoważniki w znaczeniu frazy rekonstruowalnej na podstawie kontekstu):

  (30) *Jan czyta książkę i my też.*

- wyrażenia z frazami nieciągłymi,

  (31) *Wpłynąłem na suchego przestwór oceanu.*

- zdania zawierające mowę niezależną,

  (32) *Jan zapytał: - Czego chcecie?*

- konstrukcje porównawcze ze słowem typu *niż*,

  (33) *Jan jest mądrzejszy niż Piotr.*

  (34) *Jan jest mądrzejszy, niż się wydawało.*

- zdania ze spójnikiem *czyli*,

  (35) *Jan czyta książkę, czyli oni umieją czytać.*

- zdania ze spójnikiem *mianowicie*,

  (36) *Jan zapytał o przyjaciół, mianowicie o Piotra i Marię.*

- zdania z jednostką typu *ktokolwiek*,

  (37) *Ktokolwiek przyjdzie, ugościcie go.*

  (38) *Idź, dokądkolwiek chcesz.*

- zdania względne będące realizacją frazy nominalnej,

  (39) *Kto pyta, nie błądzi.*

  (40) *Kupili, co kupić mieli.*

- zdania z frazą przyimkowo-przymiotnikową,

  (41) *Jan wygląda na chorego.*

- zdania z frazą przymiotnikową luźną (tzw. przydawką orzekającą),

  (42) *Jan wrócił z pracy zmęczony.*

- zdania zawierające znaki przestankowe spoza zdefiniowanego zestawu (kropka, znak zapytania, wykrzyknik, przecinek, wielokropek przed wielką literą) lub niepoprawne interpunkcyjnie (bez znaku końca) — interpunkcja jest traktowana w GFJP jako zjawisko składniowe.

[6]Na podstawie [Świdziński, 1992a], s. 28–30.

Co ciekawe, podobnie jak w przypadku trudności z weryfikacją konstrukcji niepoprawnych w kontekście danej reguły, również część zdań z powyższych przykładów jest akceptowana przez oryginalną postać gramatyki. Fraza przyimkowo-przymiotnikowa *„na chorego"* jest w ogólności frazą luźną, zatem całe zdanie *„Jan wygląda na chorego."* zostanie przez gramatykę zaakceptowane. Podobnie *„z pracy"* i *„zmęczony"* są realizowane jako frazy luźne, toteż całe zdanie *„Jan wrócił z pracy zmęczony."* należy do języka generowanego przez gramatykę.

7.1.5 Typy konstrukcji składniowych i mechanizmy zapewnienia zgodności składniowej

Świdziński wyróżnia dwa główne typy konstrukcji składniowych: współrzędny, zakładający podobieństwo gramatyczne składników związanych spójnikiem równorzędnym lub szeregowym i podrzędny, polegający na związaniu konstrukcji potencjalnie różnych gramatycznie (choć zwykle zespolonych ścisłym wymaganiem).

Jednostki nieterminalne posiadają zestawy parametrów służących reprezentacji interesujących z punktu widzenia opisu składniowego cech jednostki, począwszy od kategorii gramatycznych poprzez sformalizowane przez autora gramatyki cechy składniowe (zależność, typ zdania podrzędnego) aż do własności słownikowych (aspekt czasownika, wymaganie przypadka).

Parametry te wykorzystywane są w dwóch zasadniczych mechanizmach, na których opiera się GFJP: wymaganiu składniowym (parametr konstrukcji musi przyjmować żądaną wartość) i uzgodnieniu (wartości parametrów składniowych różnych jednostek muszą być równe). Uzgodnienie może się dokonywać w pionie (między stronami reguły) lub poziomie (wewnątrz treści reguły, po jej prawej stronie)[7]. Oprócz reprezentacji interweniujących w uzgodnieniach tradycyjnych cech leksykograficznych parametry służą także wprowadzeniu dodatkowych ograniczeń gramatycznych. Poniżej przedstawiam krótki opis tego rodzaju niestandardowych parametrów[8], ograniczając się do własności nazywanych w artykule [Bień, 1997b][9] *zewnętrznymi*, czyli nie wynikających jednoznacznie z formy danego wyrazu czy konstrukcji.

Nazwa parametru	Opis i znaczenie parametru
negacja	parametr konstrukcji z centrum czasownikowym (jednostki zdaniowe, fraza finitywna, werbalna i zdaniowa) informujący o jego zaprzeczoności (wartość `nie` lub `ani`) lub niezaprzeczoności (wartość `tak`); jest ponadto używany w definicji frazy wymaganej, luźnej i wszystkich typach fraz szczegółowych w różnych celach, np. dla opisu związku przeczenia z rządem czasownika,

[7] Terminologii tej będę używać w dalszej części pracy, mówiąc o uzgodnieniach poziomych i pionowych.

[8] Opis szerszy wraz z dyskusją jego konsekwencji znalazł się w pracy [Woliński, 2004], s. 87–99, nie zamierzam go tu zatem powtarzać.

[9] Patrz s. 3.

Nazwa parametru	Opis i znaczenie parametru
inkorporacja	parametr informujący o wystąpieniu wewnątrz jednostki spójnika inkorporacyjnego (o wartości odpowiadającej nazwie spójnika lub `ni` dla wartości nieinkorporacyjnych), użyty m. in. do wykluczenia inkorporacyjności zdań względnych,
zależność	najważniejszy parametr GFJP, tworzący „ducha gramatyki" i określający własności składniowe jednostki z centrum finitywnym (pytajna, niepytajna, pospójnikowa, pytajnozależna lub względna — z uwzględnieniem typów spójników i zaimków lista wartości liczy 40 kategorii), obrazujący uzależnienie własności składniowych zdania od kontekstu,
typ frazy zdaniowej	parametr frazy zdaniowej (oznaczenie początkowego spójnika podrzędnego lub względnego albo wartość pytajnozależna); ogranicza kontekstowo użycie frazy danego typu (np. fraza typu `gdyby` jest składnikiem inicjalnym zdania prostego lub realizacją frazy luźnej),
korelatywność	parametr frazy zdaniowej związany z jej typem i określający dopuszczalność wystąpienia korelatu — elementu *to*, *tego*, *temu*, *tym* z ew. przyimkiem,
ograniczenie wewnętrzne	parametr zdania elementarnego i frazy finitywnej, uzależniający własności gramatyczne centrum od sąsiedztwa spójnika podrzędnego (np. spójnik *gdyby* wpływa ograniczająco na wartość parametru czasu centrum),
klasa	parametr frazy przyimkowej, nominalnej, przymiotnikowej i przysłówkowej określający jej rodzaj; wprowadza m. in. ograniczenie typu składnikowej frazy nominalnej w złożonej konstrukcji nominalnej (poprawna jest np. konstrukcja *książka którego autora*, ale nie *książka kogo*).

7.1.6 Hierarchia składników

GFJP zawiera wielopoziomową hierarchię składników, którą można podzielić na trzy zasadnicze grupy jednostek: zdania, frazy i jednostki funkcyjne.

Poniżej przytaczam jedynie zgrubną charakterystykę jednostek — bardziej szczegółowe opisy fragmentów gramatyki zawiera rozdział 11 (s. 119), w którym analizuję budowę poszczególnych konstrukcji pod kątem ich produktywności i pełności opisu.

Jednostki zdaniowe

Jednostkę najwyższego poziomu stanowi *wypowiedzenie*, realizowane w gramatyce jako zdanie złożone zakończone znakiem interpunkcyjnym typu kropki.

Hierarchia właściwych jednostek zdaniowych przedstawia się następująco:

- *zdanie równorzędne* — zdanie złożone z centrum spójnikowym równorzędnym (spójnikiem równorzędnym centralnym bądź nieciągłym wyrażeniem spójnikowym),
- *zdanie szeregowe* — zdanie równorzędne oparte na spójniku szeregowym,
- *zdanie jednorodne* — zdanie szeregowe oparte na tym samym spójniku szeregowym; jednostka pomocnicza,
- *zdanie proste* — redukowalne do pojedynczej frazy finitywnej; zawiera jedno zdanie elementarne poprzedzone spójnikiem podrzędnym i frazę zdaniową albo dwa zdania elementarne połączone spójnikiem podrzędnym,
- *zdanie elementarne* — nie posiadające centrum spójnikowego, zorganizowane wokół pojedynczej frazy finitywnej.

Rekursywność definicji zapewnia szczególna realizacja zdania elementarnego, które może być w szczególności właściwą jednostką zdaniową najwyższego poziomu — zdaniem równorzędnym.

Jednostki poziomu frazowego

Tradycyjny podział fraz — na finitywne, wymagane i luźne — jest dla potrzeb GFJP zbyt ogólny, autor nazywa zatem ich szczegółowe realizacje morfologiczne:

- *fraza werbalna* (finitywna i niefinitywna) — redukowalna do formy czasownika,
- *fraza przyimkowa* — złożona z przyimka i frazy nominalnej,
- *fraza nominalna* — równoważnik dystrybucyjny rzeczownika,
- *fraza przymiotnikowa* — redukowalna do formy przymiotnikowej,
- *fraza przysłówkowa* — realizowalna przez formę przysłówkową przymiotnika[10]),
- *fraza zdaniowa* — reprezentująca składowe zdanie podrzędne i realizowana jako zdanie lub zawierająca zdanie jako składnik bezpośredni.

Szczególną (choć w żaden sposób nie wyróżnioną) realizacją frazy wymaganej jest fraza nominalna w mianowniku, tradycyjnie klasyfikowana jako podmiot.

Jednostki frazowe posiadają też własne struktury podrzędne, wzorowane na strukturze jednostek zdaniowych.

Jednostki funkcyjne i elementarne

GFJP zawiera ponadto katalog jednostek pomocniczych, łączących jednostki poziomu zdaniowego lub frazowego. Są nimi:

[10]Por. też uwagi w rozdziale 9.2.1 (s. 93).

- spójniki,
- pytajniki partykułowe (*czy* i *czyżby*),
- zaimki pytajne,
- zaimki względne,
- aglutynanty (*-m*, *-ś*, *-śmy*, *-ście*),
- korelaty (*to*, *tego*, *temu*, *tym*),
- znaki interpunkcyjne.

Definicje powyższych jednostek funkcyjnych oraz jednostek elementarnych stanowiących właściwe części składowe fraz (form czasownikowych, przyimkowych, rzeczownikowych, przymiotnikowych i przysłówkowych) odwołują się do słownika jednostek terminalnych — *leksykonu*, którego istnieje Świdziński zakłada, ale z oczywistych powodów nie definiuje.

Oto przykład definicji jednostki reprezentującej zaimek rzeczowny:

```
zaimrzecz(F, P, R/L)
--> [F],
    { slow (F, zaimrzecz, P, R/L) }.
```

7.2 Analizator składniowy Świgra

7.2.1 Zadanie analizy składniowej

Zadanie *analizy składniowej* rozumiemy jako potwierdzenie lub zaprzeczenie zgodności napisu na wejściu z daną gramatyką formalną, a w razie stwierdzenia zgodności — także wygenerowanie wszystkich drzew możliwych rozbiorów gramatycznych napisu. W ogólnym przypadku takich rozbiorów może być wiele, nie tylko z racji znaczących różnic w strukturyzacji, ale także np. z powodu wielu możliwości grupowania struktur współrzędnych. Z tej przyczyny, mimo iż zadanie analizy składniowej może być wykonane ręcznie, stopień złożoności języka naturalnego znajdujący odzwierciedlenie w złożoności opisującej go gramatyki sprawia, że dopiero analiza automatyczna ujawnia większość potencjalnych niejednoznaczności.

7.2.2 Wcześniejsze próby wykorzystania GFJP do analizy automatycznej

Zadanie analizy składniowej nie jest trywialne — świadczą o tym kilkukrotne próby komputerowej realizacji GFJP podejmowane w zasadzie od momentu jej powstania.

Pierwszą z nich był analizator AMOS[11] realizowany w latach 1994–1996 pod kierunkiem Janusza S. Bienia w ramach grantu KBN nr 8 S503 032 07 *Analizator morfologiczno-syntaktyczny dla obszernego podzbioru języka polskiego*. Komponent morfologiczny stanowił w AMOS-ie program SAM[12] Krzysztofa Szafrana — pierwszy

[11] Patrz [Bień, 1996], [Bień, 1997a].
[12] Patrz [Szafran, 1996].

polski analizator morfologiczny udostępniany bezpłatnie w Internecie. W analizatorze składniowym zastosowano dostępną standardowo dla gramatyk metamorficznych zstepującą strategię analizy z wykorzystaniem oryginalnej postaci gramatyki Świdzińskiego. W tym etapie opracowano także sposób generowania i prezentacji drzew analizy, wykorzystywany w prawie nie zmienionym kształcie w dzisiejszym rozwiązaniu.

Próbę drugą stanowił analizator AS [Bień, 2000], jeden z wyników projektu KBN nr 8 T11C 002 13 *Zestaw testów do weryfikacji i oceny analizatorów języka polskiego*, realizowanego także pod kierunkiem Janusza S. Bienia w latach 1997–1999. Gramatykę Świdzińskiego przekształcano do postaci równoważnej, ale o lepszych parametrach obliczeniowych, dla której można było zastosować wstępującą strategię analizy. Po zakończeniu analizy składniowej jej wyniki normalizowano w procesie nazwanym „świdzińskizacją", polegającym na konwersji wyników do postaci zgodnej z GFJP, tak by użytkownik analizatora odniósł wrażenie, że korzysta z gramatyki oryginalnej. Dla polepszenia wydajności działania analizatora stosowano ponadto zapamiętywanie wyników pośrednich. Większość z przyjętych rozwiązań okazała się słuszna i jest wykorzystywana w najnowszej wersji analizatora.

Oba prototypy analizatorów powstały w zasadzie dla weryfikacji realizowalności gramatyki, toteż intencją ich twórców nie była optymalizacja efektywności ich działania. W konsekwencji analiza nawet krótkich zdań w rodzaju *„Jan umarł."* mogła trwać w analizatorze AMOS nawet kilka godzin — zbyt długo, by analizator ten można było uznać za więcej niż eksperyment informatyczny. Efektywność AS-a, przede wszystkim wskutek zmiany strategii analizy, okazała się już znacznie lepsza (na poziomie jednej sekundy dla jednego drzewa analizy), ale wciąż zbyt niska dla zastosowań praktycznych. Były to jednak eksperymenty nader pożyteczne, gdyż właśnie na ich bazie mógł powstać pierwszy analizator o w pełni zadowalających parametrach obliczeniowych — Świgra.

7.2.3 Świgra — komputerowa realizacja GFJP

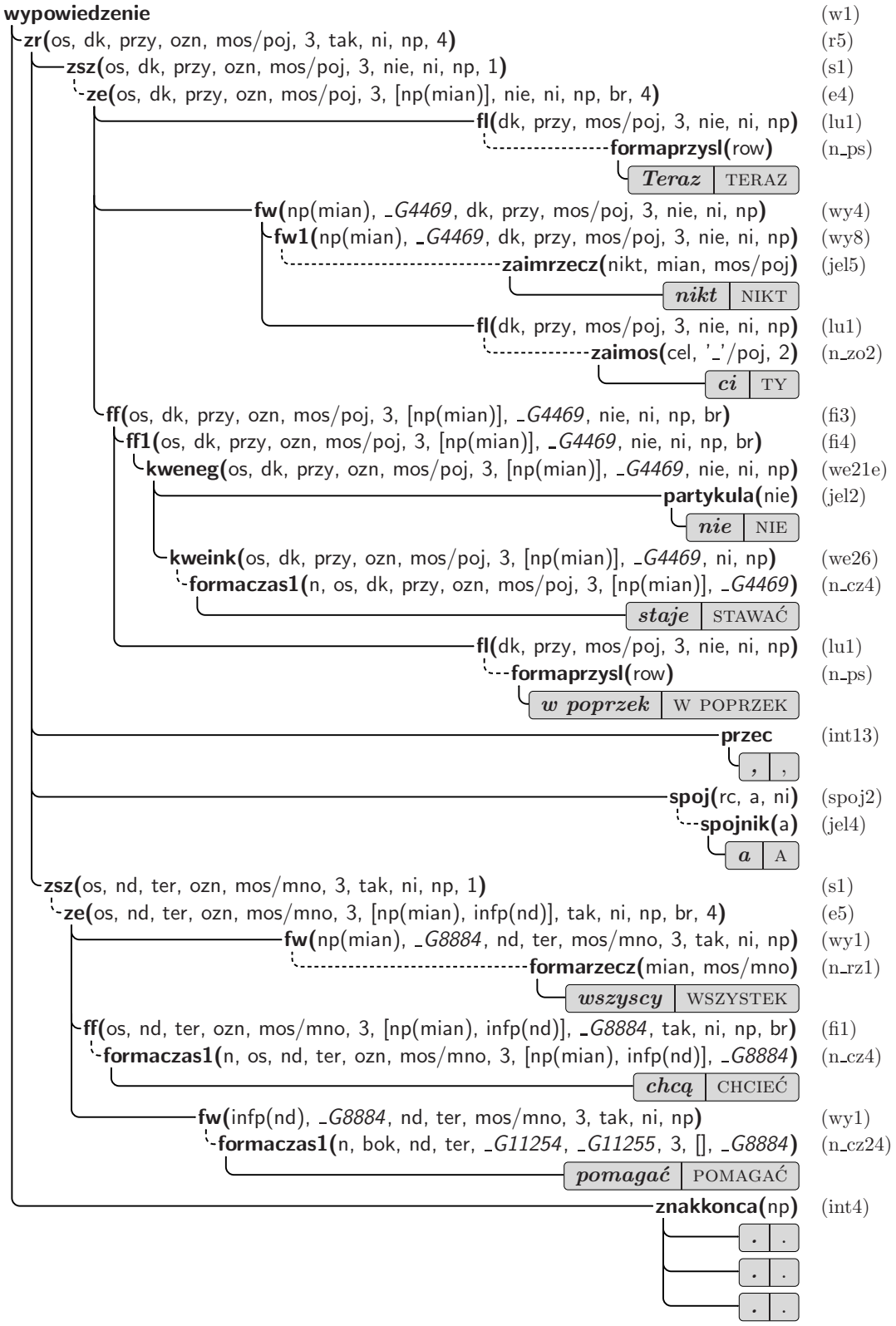
Na bazie wcześniejszych eksperymentów z gramatyką Świdzińskiego wykazujących, że własności języka polskiego skłaniają ku zastosowaniu wstępującego porządku analizy, powstał analizator składniowy Świgra o strategii wstępującej z zapamiętywaniem wyników pośrednich. Analizator ten zaimplementowany został w języku Prolog w sposób umożliwiający praktyczną komputerową realizację GFJP w postaci możliwie bliskiej oryginału.

Wobec lingwistycznego charakteru prac nad tworzeniem gramatyki, który spowodował, że nie była ona optymalizowana pod kątem zastosowań informatycznych[13] algorytm działania analizatora musiał uwzględniać jej odpowiednie przekształcenie. Efektywność czasową i pamięciową programu poprawiło użycie techniki przechowywania wyników analizy pośredniej w postaci „upakowanego lasu", dzięki czemu poddrzewa analizy przechowywane są w pojedynczych egzemplarzach i nie są generowane wielokrotnie.

[13] W często cytowanym zdaniu z opisu gramatyki Świdziński wyraża nawet przekonanie, że *bezpośrednia implementacja nawet fragmentów podanej gramatyki nie wydaje się możliwa*, patrz [Świdziński, 1992a], s. 58.

Komponentem morfologicznym Świgry jest znany nam już dobrze, dostępny w postaci zewnętrznej biblioteki analizator Morfeusz, którego wyniki są po stronie Świgry konwertowane do taksonomii wymaganej przez GFJP. Dodatkowym składnikiem analizatora Świgra jest interfejs do graficznej prezentacji drzew analizy składniowej z wykorzystaniem prologowej biblioteki XPCE. Przykładowy wynik działania analizatora dla wypowiednika `3040` *„Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać..."* przedstawia rys. 7.1.

Jednym z wyników pracy jest udostępnienie na dołączonej płycie pełnego środowiska analizy składniowej złożonego z przekształconej w opisany niżej sposób gramatyki Świdzińskiego, analizatora Świgra i interpretera Prologu, opisane szczegółowo w rozdziale D (s. 205).



Rysunek 7.1: Skrócone drzewo analizy zdania *„Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...”*

Rozdział 8

Wstępna weryfikacja składniowa wypowiedników

Poprzednie etapy weryfikacji dotyczyły „niskopoziomowej" analizy tekstów wypowiedników, w bieżącym rozdziale postaram się zatem zająć wyodrębnionymi przez autorów korpusu strukturami wykraczającymi poza granice segmentów czy słów. Po zarysowaniu ważnych różnic w opisie gramatycznym przyjętym dla korpusu i gramatyce formalnej Świdzińskiego porównuję uzyskane na bazie obu formalizmów reprezentacje wynikowe struktur frazowych, a następnie dokonuję wstępnej analizy składniowej pełnych tekstów wypowiedników.

8.1 Analiza składników frazowych

Inicjalnym etapem sprawdzającym opis elementów wskazanej struktury gramatycznej była ich analiza składniowa z wykorzystaniem programu Świgra. Ćwiczenie to jest w zasadzie niezależne od zadania pełnej weryfikacji polegającej na analizie składniowej kompletnych tekstów wypowiedników i z założenia może ono dać lepsze wyniki, gdyż oba poziomy opisu gramatycznego — frazowy i zdaniowy — mogą być testowane oddzielnie.

Weryfikacja tego rodzaju wydaje się zgodna z intencją Świdzińskiego, który przy opisie procesu wstępnej analizy danych z korpusu wypowiedników omawia problem ograniczeń GFJP na konkretnych przykładach, podając przy tym strukturalizację frazową także wypowiedzeń niepoprawnych w sensie GFJP[1].

Poniżej opisuję wyniki analizy przeprowadzonej z wykorzystaniem oferowanej przez analizator Świgra możliwości sprawdzenia interpretacji fragmentów wypowiedzeń poprzez podanie jako parametru wywołania procesu analizy nazwy jednostki gramatyki, względem której dokonujemy weryfikacji badanego napisu. W naszym przypadku jednostkę tę stanowi fraza typu odpowiadającego oznaczeniu fragmentu w korpusie (patrz rozdział 1.3.2, s. 19), zatem pierwszy krok przygotowawczy stanowi ustalenie odpowiedniości ogólnie zdefiniowanych struktur frazowych korpusu wypowiedników i jednostek składniowych gramatyki Świdzińskiego. Korpusowym

[1] Por. [Świdziński, 1993a], s. 21–22.

frazom finitywnym, wymaganym i luźnym w naturalny sposób odpowiadają frazy finitywne, wymagane i luźne GFJP, frazie podmiotowej przypisuję jako odpowiednik frazę nominalną. Odpowiadające sobie jednostki różnią się jednak nieco zakresem realizacji, każdy podrozdział wykorzystuję zatem również do porównania sposobów opisu odnośnej konstrukcji w GFJP i korpusie wypowiedników.

Warto w tym miejscu zaznaczyć, że jakość oznaczeń frazowych nie ma i w zasadzie nie może mieć żadnego wpływu na efekt analizy składniowej pełnych tekstów wypowiedników — właściwej analizie podlega jedynie ciąg słów, bez wykorzystania dodatkowych oznaczeń (choć można by się pokusić o uzupełnienie Świgry o możliwość wskazania „pożądanej" interpretacji fragmentów analizowanego tekstu, by np. ujednoznacznić warianty analizy). Proces ten zdecydowałem się jednak przeprowadzić jako dodatkowy etap weryfikacyjny, m. in. ułatwiający wskazanie składników oznaczonych niestandardowo, a więc potencjalnie nadmiarowych oraz dający ostrzejsze, bo „lokalne" spojrzenie na ewentualne problemy składniowe. Wiele z obserwacji okazało się bardzo przydatnych także w fazie analizy gramatyki; poniższy opis jest ponadto ilustracją ciekawych cech składniowych fraz poszczególnych typów, w większości nie ujętych w opracowaniu [Świdziński, 1996].

Eksperyment z analizą syntaktyczną fraz przeprowadziłem dla wszystkich składników frazowych w całym korpusie (co w praktyce oznaczało wypowiedniki zdaniowe, gdyż tylko dla nich dostępne było rozbicie struktury) po sczytaniu fraz wszystkich typów. W ten także sposób ujawniły się błędy w oznaczaniu fraz, uzupełniające listę z rozdziału 5.2 (s. 50)[2].

8.1.1 Frazy finitywne

Liczba fraz oznaczonych w korpusie wypowiedników jako finitywne wynosi 4807 (oba segmenty frazy nieciągłej — 140 przypadków — traktuję jako pojedynczą frazę), z czego różnych — 2851.

Oprócz „typowych" realizacji odpowiadających GFJP (forma leksemu czasownikowego lub quasi-czasownikowego, syntetyczna lub analityczna, być może nieciągła, być może z *się*, ewentualnie poprzedzona partykułą negacji lub mająca postać formy klitycznej leksemu *być*[3]) instrukcja dopuszcza ponadto reprezentację frazy finitywnej jako zdania elementarnego–reszty („orzeczenia szeregowego", ew. z pewnym składnikiem wspólnym). W korpusie oznaczono tak 145 wystąpień frazy finitywnej, np.

(43) *...=że= (coś) <tam u nich podpatrzę i przeniosę do naszej sekcji>.* *[3065]*

Świdziński opisuje tego rodzaju frazy jako składniki zdań elementarnych o postaci nietypowej[4], które następnie klasyfikuje jako *czasownikowe, quasi-czasownikowe oraz quasi-czasownikowe „składniowe"*. Konstrukcje te są na tyle ciekawe, że warto się nad nimi na chwilę zatrzymać, zwłaszcza, że w GFJP zdanie elementarne oparte

[2]Rozdział C.2 (s. 199) zawiera pełną listę usterek w oznaczeniach struktury frazowej, łącznie z wykrytymi na tym etapie błędami wynikającymi z błędnej klasyfikacji fraz.

[3]Opis na podstawie instrukcji dla edytorów: patrz [Świdziński, 1994a], s. 3.

[4]Patrz [Świdziński, 1996], s. 105–107.

jest na dokładnie jednej frazie finitywnej, a sama konstrukcja przytoczonego przykładu jest daleka od nietypowości.

Zdania z centrum czasownikowym i quasi-czasownikowym to konstrukcje rozróżniane w sposób zgodny z opisem schematów zdaniowych, czyli ze względu na kryterium obecności podmiotu-mianownika; zdania z centrum quasi-czasownikowym „składniowym" to z kolei konstrukcje zawierające wyrzuconą „przed nawias" pewną jednostkę funkcyjną (np. frazę zaimkową względną czy podrzędnik odnoszący się do wszystkich spojonych współrzędnie członów frazy finitywnej).

Opisując tego rodzaju konstrukcje Świdziński przyznaje otwarcie[5]:

> Interpretacja składniowa przykładu PU (71-1-4) »=że= [przekształcenie człowieka przez miłość] <będzie i może być> (tylko dziełem bóstwa),...« [5692] jest prowizoryczna i w gruncie rzeczy niewłaściwa. GFJP nie dostarcza, niestety, skutecznego mechanizmu opisu wyrażeń typu »będzie i może być« („orzeczeń modalnych").

nie wskazuje natomiast odpowiedniego sposobu postępowania z tego rodzaju przypadkami, podczas gdy taka interpretacja składniowa narzuca się w sposób dość oczywisty.

Zdania tego rodzaju zawierają — w moim przekonaniu — konstrukcję, którą na podobieństwo grupy nominalnej nazwałbym *grupą werbalną* (lub quasi-werbalną). Tak zdefiniowana grupa stanowiłaby łącznie centrum elementarnego zdania „nietypowego" — właśnie w postaci funkcjonującego w składni tradycyjnej pojęcia orzeczenia szeregowego.

Podobnie należałoby interpretować zdanie:

(44) *Zaczął jeść i mówić.* *[5212]*

które gramatyka Świdzińskiego błędnie akceptuje jako złożone z dwóch zdań elementarnych — *„Zaczął jeść."* i *„Mówić."*[6], podczas gdy jest to jedno zdanie elementarne z grupą werbalną *jeść i mówić*. Dla zdań zawierających tego rodzaju grupę przewaga centrum spójnikowego nad finitywnym, na którym to założeniu opiera się konstrukcja zdania złożonego w GFJP, realizuje się w specyficzny sposób, gdyż nie rozgranicza zdań elementarnych.

Idąc dalej w podjętych rozważaniach ośmielę się postawić tezę, że zastosowanie systematyki grup do realizacji czasownikowych może prowadzić także do uznania rozumianego ogólnie zdania złożonego właśnie za swego rodzaju grupę werbalną — choć ze względów dekompozycyjnych wygodnie jest uważać za jego centrum spójnik, „prawdziwym" centrum zdania złożonego byłaby jednak wciąż konstrukcja czasownikowa. Taka interpretacja mogłaby być nawet dobrym poparciem przyjętego przez

[5] [Świdziński, 1996], s. 106.

[6] Jest to wynikiem istnienia w gramatyce reguły `we22` definiującej konstrukcję werbalną z negacją jako konstrukcję werbalną z inkorporacją, po której następuje fraza luźna, posiadająca w szczególności realizację pustą (reguła `lu14`). W ramach zadania usuwania reguł epsilonowych w gramatyce przekształconej przez Wolińskiego na potrzeby Świgry odnośna reguła nosi numer `we22e`.

Świdzińskiego[7] założenia o rezygnacji z wprowadzanego przez tzw. składnię tradycyjną podziału na zdania pojedyncze i złożone. Jednostki nazywane przez Świdzińskiego zdaniami poszczególnych typów (równorzędne, szeregowe, jednorodne itd.) są zgodnie z tym duchem realizacjami grupy werbalnej tworzącej *zdanie* o dowolnym stopniu złożoności.

Świdziński również zdaje sobie sprawę z arbitralności tego rodzaju klasyfikacji, wyrażając wątpliwość przy opisie przykładu zdania elementarnego z centrum quasiczasownikowym:

(45) *\Chyba\ (mi) <się przywidziało albo przyśniło>.* *[1952]*

(46) *..., <wolno czy nie wolno> (wykonywać planów pozyskania zwierzyny) /na terenach nowo utworzonych obwodów, na które nie zostały zawarte umowy dzierżawcze/, ...* *[6527]*

Przykład drugi mógłby być interpretowany inaczej, jako zdanie złożone z centrum *czy*. Jego poprawność gramatyczna jest zresztą wątpliwa.[8]

Budowa trzech zaobserwowanych przykładów fraz finitywnych nietypowych wykracza jeszcze dalej poza opisany wyżej schemat „składników wyrzuconych przed nawias":

(47) *<Ki diabeł? - mówi do siebie...> [gospodarz] <...i woła przez drzwi: - A kto tam?>.* *[5194]*

(48) *..., =że= [ja] <nie chodziłem, nie buntowałem moich parafian, nie mówiłem im "wychodźcie, wychodźcie, bo to nie nasza msza", gdy oni mieli stanąć przed ołtarzem>.* *[5321]*

(49) *[Książę Bismarck] <nie wytrzymał tego długo i zawołał: - Cóż to ma znaczyć>?* *[5564]*

jednak przypadki zawierające mowę zależną wydają się w ogóle oderwane od omawianej klasyfikacji frazowej i powinny być raczej rozpisywane na kolejne wypowiedniki niż analizowane jako elementarne, wyłączam je zatem zupełnie z dalszego przetwarzania.

Frazy o strukturze nietypowej zdecydowałem się eksperymentalnie poddać analizie traktując je jak kompletne wypowiedzenia. Ze 142 tak wyodrębnionych wypowiedzeń (odpowiadających wszystkim frazom oprócz powyższych trzech) powiodła się analiza 103 (72,5%).

W zbiorze fraz finitywnych uznanych za realizacje typowe z 2694 różnych fraz analiza 2671 zakończyła się sukcesem (99%). W przeważającej liczbie przypadków (2650) rezultatem analizy są pojedyncze drzewa; pozostałe dają w wyniku po dwa drzewa,

[7] Por. [Świdziński, 1992a], s. 21.

[8] [Świdziński, 1996], s. 106.

co jest konsekwencją faktu, że Morfeusz pewne formy czasowników traktuje jako wieloznaczne morfologicznie[9].

Przypadki nieakceptowane obejmują konstrukcje nieciągłe (w sensie GFJP) z zaimkiem *się* rozdzielonym od czasownika partykułą *nie* (21 wystąpień), np.

(50) *\Chyba\ <się nie gniewasz>.* *[1306]*

(51) *...=,= [to] <się nie opłaca>.* *[2765]*

Ich analiza nie jest możliwa bez zmiany gramatyki, która byłaby w tym przypadku prosta technicznie, ale niosąca spore konsekwencje logiczne, wobec czego nie zdecydowałem się jej wprowadzić. Oto wyjaśnienie: parametr negacji pojawia się w GFJP w konstrukcjach finitywnych dopiero na poziomie konstrukcji werbalnej z negacją, natomiast dla zaimka zwrotnego zakłada się w regułach `n_cz2` i `n_cz3` (bardzo surowo!) jego wystąpienie w bezpośrednim sąsiedztwie czasownika (uda się zatem zanalizować zdanie *„Nie uda się."*, nie uda się natomiast zdania *„Się nie uda."*). Rozwiązaniem problemu mogłoby być wprowadzenie wariantów wymienionych reguł dopuszczających możliwość wystąpienia partykuły negatywnej, wymagałoby to jednak uwzględnienia parametru negacji na niższym poziomie niż do tej pory, a więc sporej zmiany koncepcyjnej. Poprawki tej zdecydowałem się nie wprowadzać także ze względu na jej doraźność, gdyż problem dotyka znacznie poważniejszej kwestii obsługi ruchomego zaimka zwrotnego (*się* i czasownik może rozdzielać więcej jednostek niż wyłącznie partykuła *nie*), która nie może już zostać rozwiązana za pomocą tak prostych metod, bez znacznego przeorganizowania gramatyki.

Ostatni przypadek, dla których analiza frazowa zakończyła się porażką pochodzi z wypowiednika:

(52) *(1 Gdzie 1) <by...> (2 tego 2) /znów/ <...wysłać>*[10] *[1048]*

i ilustruje nie opisywaną przez gramatykę konstrukcję z bezokolicznikiem i partykułą warunkową.

8.1.2 Frazy podmiotowe

Liczba fraz oznaczonych w korpusie wypowiedników jako podmiotowe wynosi 2292 (oba segmenty frazy nieciągłej traktuję łącznie), z czego różnych — 1693.

Budowa fraz podmiotowych w zasadzie nie wykazuje nietypowości, z wyjątkiem realizacji w postaci frazy zdaniowej (29 przypadków), np.

(53) *<Wymknęło...> (ci) <...się> [, że przesadą byłoby zginąć za komunizm].* *[4626]*

[9] Są to czasowniki *patrzeć/patrzyć* (9 wystąpień), *znajdować/znajdywać* (7 wystąpień), *śmiać/śmieć* (2 wystąpienia), *zbladnąć/zblednąć* (1 wystąpienie), *winien/winny* (1 wystąpienie) i *zwiedzić/zwiedzieć* dla pojedynczego wystąpienia formy *zwiedzą* (w tym akurat przypadku pierwsza interpretacja jest błędna, gdyż nie istnieje czasownik *zwiedzieć*, a jedynie *zwiedzieć się*).

[10] Zapis oryginalny — bez znaku końca wypowiedzenia.

i mianownikowej formy zaimka *się*[11] (60 wystąpień), np.

(54) *(Spis wyborców) <układa> [się] /1 oddzielnie 1/ /2 dla każdego obwodu głosowania 2/.* *[6492]*

Zarówno GFJP, jak i jej komputerowa realizacja są przygotowane do analizy zdań zawierających nietypowe frazy podmiotowe pierwszego rodzaju — poprzez uzupełnienie słownika wymagań o odpowiednie frazy zdaniowe. W drugim przypadku poprawna analiza zdania była możliwa dopiero po wprowadzeniu opisanej w rozdziale 11.2.12 (s. 134) modyfikacji, gdyż gramatyka nie przewiduje zaimkowej realizacji frazy wymaganej.

Na 1693 analizowane cząstkowo różne frazy podmiotowe sukcesem zakończyła się analiza 1290 fraz (76,17%).

8.1.3 Frazy wymagane

Liczba fraz oznaczonych w korpusie wypowiedników jako wymagane wynosi 4483 (oba segmenty frazy nieciągłej traktuję łącznie), z czego różnych — 3716.

Na potrzeby analizy wszystkie frazy wymagane traktuję w jednolity sposób, niezależnie od niejasnego opisu wykluczającego je z dalszego przetwarzania (zapewne z racji zachowania zgodności z przyjętym modelem schematów zdaniowych)[12]:

> Jako „człon inny" traktowane są też takie jednostki składniowe, jak fraza wymagana trzecia, czwarta itp., fraza luźna trzecia, czwarta itp. Wtedy odpowiednie składniki ujmowane są w nawiasy odpowiadające danemu typowi frazy (z indeksem 3 lub 4), a ich opis podawany w sposób specjalny. Jest to jednak tylko technika zapisu, nie zaś interpretacja składniowa.

Wykluczenie to jest zresztą tylko postulatem, gdyż w przypadku fraz wymaganych informacja o wystąpieniu frazy trzeciej i czwartej podawana jest co prawda w polu przechowującym typ członu innego i ogranicza się do kategorii frazy, lecz pole opisu zawiera pełną charakterystykę składniową fraz niższych rzędów.

Oto poglądowy rozkład rodzajów fraz wymaganych, przygotowany na bazie wartości pola *OPIS*[13] i posortowany względem częstości wystąpień. Realizacje frazeologizmów zliczam łącznie z realizacjami standardowymi:

[11] Por. np. [Saloni, 1982]. Podmiotowa konstrukcja tego rodzaju traktowana bywa także łącznie, jako nieosobowa forma czasownika — por. np. [Kopcińska, 1997], s. 29.

[12] [Świdziński, 1996], s. 42.

[13] Jedną z dopuszczalnych wartości jest też `m` oznaczające frazę nominalną; tego typu informacja przysługuje zdaniom bez fraz wymaganych w ścisłym rozumieniu korpusowym (oznaczonych nawiasami okrągłymi), za to zawierającym oznaczoną parą nawiasów kwadratowych frazę podmiotową; kategorię tę pomijam w podanym wykazie.

Rodzaj frazy wymaganej	Częstość wystąpień
`fraza nominalna`	42,96%
`fraza przyimkowo-nominalna`	16,58%
`fraza przysłówkowa`	11,92%
`fraza bezokolicznikowa`	11,88%
`fraza zdaniowa`	8,60%
`fraza przymiotnikowa`	7,87%
`fraza przyimkowo-przymiotnikowa`	0,19%

Sukcesem zakończyła się analiza 2428 fraz oznaczonych jako wymagane (65%).

8.1.4 Frazy luźne

Liczba fraz oznaczonych w korpusie wypowiedników jako luźne wynosi 2125 (oba segmenty frazy nieciągłej traktuję łącznie), z czego różnych — 1655. Podobnie jak w przypadku fraz wymaganych analizuję również trzecie (30 wystąpień) i czwarte (4 wystąpienia) frazy luźne.

W odróżnieniu od składników innych typów frazom luźnym nie jest przypisywana żadna charakterystyka jakościowa (wyłącznie parametry ilościowe — długość pierwszej i drugiej frazy luźnej w słowach), jedyne więc zadanie weryfikacyjne może dotyczyć sprawdzenia ich akceptowalności przez aparat GFJP.

Sukcesem zakończyła się analiza 992 fraz luźnych (59,94%).

8.1.5 Człony inne

Człony inne, zdefiniowane w sposób ogólny jako składniki różne od wyżej wymienionych, nie mają odpowiednika w GFJP. Ich charakterystyka jest trudna do określenia, w związku z czym nie mogą być one analizowane z użyciem arbitralnej jednostki nieterminalnej. Wykorzystując możliwości automatycznego analizatora można jednak pokusić się o sprawdzenie, którego nie dało się wykonać w momencie powstawania projektu korpusu wypowiedników — mianowicie jakiego rodzaju jednostki są reprezentowane w zbiorze fraz „innych” i jaki jest ich rozkład statystyczny. Praca ta została wykonana poprzez próbę rozbioru jednostek z użyciem każdej z dostępnych reguł składniowych GFJP. Oto jej wyniki:

Nieterminal GFJP	Liczba jednostek
`fl`	174
`fw`	151
`fno`	65
`fzd`	34
`spojnik`	29
`fpt`	24
`ff`	23
`fpm`	19
`zd`	10

Łączna liczba fraz oznaczonych w korpusie wypowiedników jako człony inne wynosi 1154, z czego różnych — 318; wyniki w tabeli nie sumują się do tej wartości, gdyż analizator mógł zwrócić więcej niż jedną interpretację danego członu.

Dane obrazują liczbę analiz dla danej jednostki gramatyki; jeśli człon został zinterpretowany na wiele sposobów, wybierałem jednostkę najwyższego poziomu, tak więc np. dla zbioru wyników:

```
, atanazy,: fl, fl1
```

wybraną jednostką jest fraza luźna, nie zaś fraza luźna właściwa. Taka interpretacja jest konsekwencją budowy gramatyki, w której aż 54 klauzule (z części zasadniczej, tj. bez uwzględniania definicji jednostek elementarnych) to proste reguły przepisywania pomiędzy pojedynczymi jednostkami nieterminalnymi, uzupełnione ewentualnie dodatkowymi warunkami ograniczającymi wartości parametrów. W naszym przypadku oznacza to w szczególności, że jeśli dany napis jest interpretowany jako realizacja jednostki podrzędnej i nadrzędnej (dla pary jednostek pozostających w takim związku), zostanie wybrana jedynie interpretacja z jednostką nadrzędną, jeśli natomiast interweniował warunek ograniczający i jednostka nadrzędna nie jest już właściwą realizacją, zostanie uwzględniona tylko jednostka podrzędna. Oto przykład takiej reguły właśnie dla frazy luźnej:

```
fl(A, C, Rl, O, Neg, I, Z)
--> s(lu1),
    fl1(A, C, Rl, O, Neg, I, Z).
```

Pełen graf związków między jednostkami nieterminalnymi (strzałka wskazuje jednostkę nadrzędną) przedstawia rys. 8.1.

W przypadku wielości interpretacji z różnych gałęzi drzewa obie analizy uznawane były za równorzędne, np. dla wyniku:

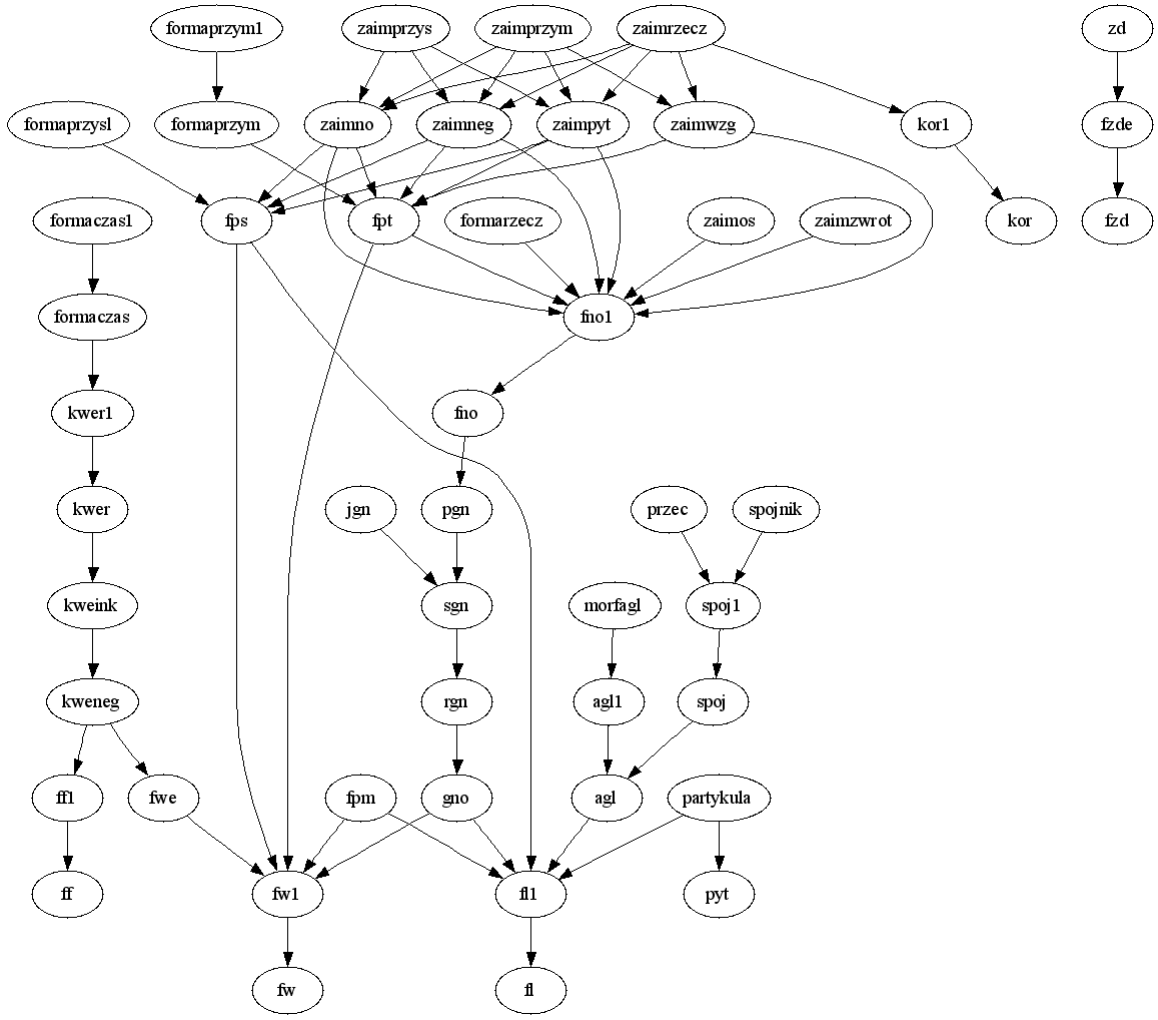
```
pewnie: fw, fw1, fl, fl1, fps, formaprzysl
```

w tabeli z poprzedniej strony znalazła się zarówno analiza z frazą wymaganą, jak i z frazą luźną.

Człony inne posiadające interpretację „wymaganą" nie są wspominanymi przez Świdzińskiego trzecimi i następnymi frazami wymaganymi[14], które miały tak zostać oznaczone ze względu na przyjęte ograniczenia — w posiadanej wersji korpusu w zdaniach zawierających nawet cztery frazy wymagane wszystkie one zostały oznaczone poprawnie.

Analiza 71 członów zakończyła się niepowodzeniem: nie została znaleziona konstrukcja gramatyki pozwalająca je opisać. Wynik ten jest, paradoksalnie, najbardziej interesujący, gdyż badane elementy należą prawie wyłącznie do klasy partykułoprzysłówków (*kublików*) — resztkowej klasy leksemów nieodmiennych nie będących

[14]Patrz [Świdziński, 1996], s. 42.



Rysunek 8.1: Graf zależności między jednostkami nieterminalnymi

przyimkami ani spójnikami, którym Morfeusz przypisuje kategorię `qub`. Oznacza to, że zdania zawierające konstrukcje kublikowe nie mogą zostać w ogóle zaakceptowane przez GFJP. Dzieje się tak dlatego, że jednostki o charakterystyce kublikowej w uzupełnionej przez Wolińskiego gramatyce wykorzystywane są w niezwykle ograniczonym zakresie (jedynie do definicji partykuł, praktycznie nie używanych w ogólnym sensie w regułach wyższego poziomu). Przykładowo, nie zagospodarowuje ich realizacja frazy luźnej, ograniczająca się w GFJP w uproszczeniu do frazy przymiotnikowej, nominalnej, przysłówkowej, werbalnej, „wtrącenia frazowego", frazy zdaniowej albo grupy wymienionych konstrukcji.

Na potrzeby analizy wypowiedników problem został rozwiązany w możliwie prosty sposób — albo poprzez zastąpienie partykuło-przysłówkowej charakterystyki jednostki charakterystyką przysłówkową, akceptowaną przez definicję frazy luźnej (jak pamiętamy z rozdziału 9.2.1, s. 93), albo (szczątkowo) poprzez opisaną w rozdziale C eliminację nietypowych fraz luźnych z analizy. Rozwiązanie pierwsze uzasadniam faktem, że niektórym członom tego rodzaju (np. *prawie*, *wreszcie*) słownik [SJP, 2002] przypisuje wprost charakterystykę przysłówkową, a wielu innym skłonni

bylibyśmy przypisać ją intuicyjnie (np. *całkiem*, *inaczej*).

Propozycją wariantową mogłoby być wprowadzenie nowej jednostki reprezentującej „człon inny", realizowanej przez wyrazy lub grupy wyrazów o charakterystyce partykuło-przysłówkowej i stanowiącej szczególną realizację frazy luźnej, do której wydają się najbardziej zbliżone charakterem i dystrybucją. Dzięki przyjęciu takiego rozwiązania kublikową interpretację mogłyby zyskać jednostki dostawione w rodzaju wtrąceń czy partykuł ekspresywnych, które w bieżącym kształcie GFJP nie mogą być w inny sposób zinterpretowane jako składniki zdania.

8.2 Zakres opisu składniowego GFJP

Przeprowadzenie automatycznej analizy korpusu wypowiedników z wykorzystaniem GFJP ma sens jedynie dla pewnego podzbioru danych korpusowych, który określimy odwołując się do podstawowych własności gramatyki Świdzińskiego. Właściwością nadrzędną jest dostarczenie opisu jednostek wyłącznie zdaniowych (a nie oznajmień), zatem z 6721 wszystkich wypowiedników analizie może zostać poddanych nie więcej niż 5262 wypowiedników zdaniowych. W związku z ograniczeniami GFJP (patrz rozdział 7.1.4, s. 65) bezcelowa jest także analiza wypowiedników zawierających:

- frazy nieciągłe i elipsy innego rodzaju niż zakładane przez projekt GFJP — 347 wypowiedników,
- mowę niezależną (w tym przypadku analizie zostanie poddana zarówno cytowana wypowiedź, jak też człon ją wprowadzający, o ile stanowią one wypowiedniki zdaniowe, zarejestrowane jako osobne rekordy) — 52 wypowiedniki,
- konstrukcje porównawcze z *niż* — 37 wypowiedników,
- konstrukcje ze spójnikami *czyli* i *mianowicie* — 22 wypowiedniki,
- jednostkę typu *ktokolwiek* — 14 wypowiedników,
- formy liczebników głównych — 462 wypowiedniki,
- frazę przyimkowo-przymiotnikową — 1 wypowiednik,
- znaki interpunkcyjne spoza akceptowanego zestawu — średniki, pauzy, nawiasy, wielokropki niekończowe, dwukropki, cudzysłowy[15] — 106 wypowiedników.

Osobny przypadek stanowi 45 przykładów zawierających wtrącenia w nawiasach. Mogłyby one zostać wyeliminowane z dalszego przetwarzania ze względu na obecność w treści znaku spoza dopuszczalnego zbioru. Zdecydowałem się jednak poddać je analizie po usunięciu wtrąceń, gdyż stanowią one we wszystkich przypadkach charakter objaśnienia/komentarza i nie interweniują poza obrębem nawiasu. Oto przykład:

[15]Oprócz przypadków, w których znak cudzysłowu służy jedynie wyróżnieniu fragmentu tekstu, jak w przypadku *„Ale to nie oni i nie doktor Borowik są odpowiedzialni za moje „opętanie" morzem."* czy *„Należy do nich bez wątpienia Daimler-Benz, który realizuje obecnie „najpotężniejszy program swej historii"."* — wypowiedniki oznaczone w ten sposób są poddawane analizie po usunięciu znaków cudzysłowu.

(55) *Badanie chemiczne wymaga bowiem posiadania przynajmniej jednego mikrograma (milionowa część grama) badanej substancji.* [3713]

(56) *Uważam, że światopogląd może być materialistyczny lub idealistyczny (w prymitywnych cywilizacjach nawet magiczny), ale nie morski, lotniczy, rolny czy leśny.* [5680]

W sześciu przypadkach[16] wtrącenia stanowią samodzielne wypowiedniki zdaniowe, opisane jako osobne rekordy bazy wypowiedników; zostały one oczywiście poddane analizie składniowej.

Po zastosowaniu powyższych zabiegów oraz usunięciu powtórzeń do analizy mogło zostać skierowanych 4242 wypowiedników.

8.3 Wyniki analizy

Po wprowadzeniu opisanych niżej usprawnień analizatora morfologicznego tekst wypowiedników mógł zostać poddany analizie składniowej. Nazywam ją *podstawową* w kontekście opisywanych w następnych rozdziałach rozszerzeń i modyfikacji, gdyż oprócz zapewnienia dostępności pełnej informacji morfologicznej dla analizowanych tekstów nie wprowadziłem jeszcze żadnych dodatkowych usprawnień ani w samej gramatyce, ani w mechanizmie analizy.

Etap ten ma na celu wykazanie stopnia korelacji GFJP z opisem wypowiedników i jest odpowiedzią na postulat Świdzińskiego[17]:

> Opis niniejszy wymaga z pewnością starannej weryfikacji empirycznej. Warto zbadać, i to na kilku bardzo dużych próbach wypowiedzeń empirycznych, stopień reprezentatywności opisu. Próbę mógłby stanowić zarówno zbiór wylosowanych z tekstów polszczyzny pisanej wypowiedzeń empirycznych, jak też zbiór wypowiedzeń już intuicyjnie zanalizowanych — na przykład spełniających warunki podane w punkcie 1.6[18]. Niezbędna jest też swojego rodzaju weryfikacja leksykograficzna, a więc — zbadanie, czy jednostki funkcyjne zdefiniowane w tej pracy wyczerpują zbiór typów jednostek funkcyjnych zawarty w danym słowniku (np. w SJP PWN); czy spis typów fraz wymaganych i spis typów wymaganych fraz zdaniowych są empirycznie wyczerpujące; czy własności składniowe jednostek funkcyjnych zostały opisane w sposób wystarczająco subtelny. Widać zatem potrzebę bardzo szeroko zakrojonych badań materiałowych, które mogą przy okazji dostarczyć nieocenionych danych do nowego słownika języka polskiego.

Jego pierwszą próbą realizacji był projekt korpusu wypowiedników, w niniejszej pracy chciałem zweryfikować niektóre jego założenia dokonując analizy automatycznej.

[16] Wypowiedniki o numerach 3834, 5839, 5863, 6037, 6082 i 6100.

[17] [Świdziński, 1992a], s. 307.

[18] Chodzi o konstrukcje nieakceptowane przez oryginalną gramatykę, por. rozdział 7.1.4, s. 65.

Próba analizy wypowiedników zdaniowych z wykorzystaniem oryginalnej gramatyki Świdzińskiego (w wersji dołączonej do Świgry, a więc już uzupełnionej i zmodyfikowanej przez Wolińskiego) zakończyła się umiarkowanym sukcesem:

Wypowiedniki	**Liczba**	**Udział %**
akceptowane	1312	30,93 %
nieakceptowane	2929	69,05 %
nie dające się zanalizować (analiza trwa dłużej niż 8 godzin)	1	0,02 %

Wypowiednikiem analizowanym najdłużej okazał się (nie bez powodu, co dyskutuję w rozdziałach 13.2.2 na s. 160 i 13.1 na s. 13.1) zawierający złożone konstrukcje współrzędne tekst:

(57) *Na to, aby młody Polak stał się patriotą poprzez trudy, walki, nadzieje, zwycięstwa i klęski przemysłu tak, jak stał się poprzez walki i klęski orężne, trzeba było, aby istniał po pierwsze polski kapitał, po drugie polska kadra techniczna.* *[5985]*

Uzyskany wynik świadczy o występowaniu w korpusie wypowiedników, a więc zapewne i w tekstach „z życia wziętych", dużej liczby konstrukcji celowo lub przypadkowo nie objętych opisem GFJP — ich szczegółowy przegląd zaprezentuję w kolejnych rozdziałach przy okazji opisu procesu przygotowania korpusu do analizy oraz dostosowania gramatyki do akceptacji większej liczby przykładów korpusowych.

Warto w tym miejscu zauważyć, że wynik ten odbiega znacznie od wartości podawanej przez Świdzińskiego w pracy [Świdziński, 1993a] (przeszło 90-procentowy udział zdań akceptowanych), na co wpłynęła zapewne metoda zatrzymywania analizy na granicy składników frazowych (ten ważny problem opisuję znacznie szerzej w rozdziale 12.2 (s.145) porównującym wyniki analizy ręcznej i automatycznej; w dalszym ciągu pracy przyjmuję uzyskany wynik trzydziestoprocentowy jako bazowy dla dalszych badań nad gramatyką i korpusem). Tym ważniejsza wydaje się więc weryfikacja z pełnym wykorzystaniem automatu, który uwypukla pewne aspekty mogące zostać zaniedbane przy nawet najuważniejszej analizie ręcznej.

Rozdział 9

Korekty wspomagające weryfikację składniową

W bieżącym rozdziale staram się zastanowić nad przyczynami słabego wyniku akceptowalności wypowiedników zdaniowych i opisuję wstępne działania podjęte w celu zapewnienia lepszych parametrów gramatycznych korpusu.

9.1 Problemy w zapisie postaci tekstowej

Trudności z automatyczną analizą korpusu wypowiedników wynikają w pierwszym rzędzie z niedostatków jakości ręcznego opracowania korpusu. Już Skibicki pisze[1]:

> Ilość zauważonych błędów, które pozostały (...) po ręcznym procesie wypełniania i weryfikacji bazy, nakazuje ograniczone zaufanie do wartości informacji syntaktycznej, która nie podlegała weryfikacji.

Oprócz opisywanych w rozdziale 5 (s. 49) problemów wynikających z tekstowej postaci korpusu, poprawionych jeszcze przed dokonaniem analizy inicjalnej, w korpusie dostrzegłem szereg innego rodzaju problemów mogących zaburzyć proces weryfikacji składniowej.

9.1.1 Brak oznaczenia członów nie należących do wypowiednika

Jednym z częstszych problemów w korpusie[2] jest nieprawidłowe oznaczenie spójnika lub partykuły wprowadzającej wypowiednik podrzędny jako członu „innego" lub nawet wymaganego, a zatem należącego do wypowiednika. W poniższym przykładzie (59) spójnik `i` został ujęty w nawiasy przysługujące oznaczeniu frazy wymaganej mimo że w tekście wypowiednika źródłowego (58) stanowi on centrum struktury złożonej, z założenia nie reprezentowanej na poziomie zdań składowych:

[1] [Skibicki, 2000], s. 31.

[2] Ok. 100 przypadków, szczegółową listę zawiera rozdział C.6 — patrz s. 203.

(58) *<Jest> (u nas we wsi) [taka, co jej było przez osiem roków odmówione, a dziewiątego zaszła i teraz ma Jadwisię.]* *[1179]*

(59) *...(i) /teraz/ <ma> (Jadwisię).* *[1184]*

Jest to oczywisty błąd edytorów korpusu, gdyż pozostała większość wypowiedników podrzędnych lub współrzędnych środkowych i końcowych zapisana jest poprawnie, z oznaczeniem członu spajającego jako nie należącego do wypowiednika, jak w poniższym przykładzie:

(60) *..., =a= [pańskie na nie odpowiedzi] <przyczynią się> /niewątpliwie/ (do rozwikłania pewnej zawiłej sprawy związanej, być może, z mającą nastąpić kradzieżą pewnych przedmiotów niewielkich, lecz jakże cennych)...* *[154]*

Duża częstość występowania i równomierność rozkładu tego błędu może być po części wyjaśniona prawdopodobnie niedokładną definicją charakteru członów uznawanych za wymagane lub „inne" — zgodnie z intuicją należałoby je ograniczyć odpowiednio do wymagania czasownikowego i członów o nieznanym charakterze, ale dających się z całą pewnością zaliczyć do wypowiednika podrzędnego. Przy takim rozumieniu klasyfikacji wyróżnianych fraz np. wymaganie frazy zdaniowej realizuje się na poziomie wypowiednika ją włączającego, a nie na poziomie zdania tworzącego tę frazę. W tego rodzaju wypadkach do poprawnej analizy składniowej konieczne było ręczne poprawienie opisu korpusowego poprzez wyłączenie nadmiarowej frazy z treści analizowanego przykładu.

Do tego i podobnych celów wprowadziłem do tekstów wypowiedników nowe oznaczenie — nawias rozpoczynający się i kończący znakiem gwiazdki, w który ujęte zostają fragmenty usunięte przeze mnie na potrzeby poprawnej analizy składniowej:

(61) **...(i)* /teraz/ <ma> (Jadwisię).* *[1184]*

9.1.2 Składniki niezdaniowe i nietypowe człony luźne

Osobno należałoby wyłączyć z procesu analizy wypowiedniki złożone, które zawierają jako składniki człony niezdaniowe. Niestety, ze względu na niekonsekwencję oznaczeń w korpusie, a mianowicie częste występowanie struktury frazowej już w wypowiedniku złożonym, nie jest wskazane automatyczne wykluczenie wypowiedników złożonych zawierających niezdaniowe, gdyż doprowadziłoby to do eliminacji zbyt wielu wypowiedników z poprawnie naniesioną strukturą frazową. W tym przypadku w treści wypowiedników złożonych oznaczałem człony niezdaniowe, które były następnie wykluczane podczas analizy automatycznej.

Oto przykład — w zestawie (dodatkowo oznaczono także typ wypowiednika):

(62) *Z: ...=,= \sam\ /1 mi 1/ /2 tu 2/ /3 przed chwilą 3/ <jęczałeś> (, po co, na co) =,=...* *[300]*

(63) *E: ..., po co =,=...* *[301]*

(64) *E: ...=,= na co,...* *[302]*

wypowiedniki eliptyczne (63) i (64) uniemożliwiają automatyczną analizę wypowiednika zdaniowego (62), którego są składnikami; wypowiednik (62) powinien jednak zostać poddany analizie, gdyż nie istnieje osobny wypowiednik dla zdania *„Sam mi tu przed chwilą jęczałeś.”*. Z tego względu człony eliptyczne wykluczam z treści wspomnianych wypowiedników, zachowując jednak podstawową część wypowiednika nadrzędnego.

W podobny sposób wyłączone zostały człony luźne wykraczające poza opis frazy luźnej w gramatyce Świdzińskiego[3]:

> Oto lista realizacji przewidzianych w niniejszym opisie:
>
> 1. fraza przyimkowa,
> 2. fraza nominalna niewołaczowa,
> 3. fraza przysłówkowa;
> 4. aglutynant,
> 5. fraza werbalna imiesłowowa,
> 6. fraza nominalna wołaczowa,
> 7. słowo puste,
> 8. fraza zdaniowa.
>
> Listę tę można by z powodzeniem rozszerzyć o jeszcze inne realizacje, jak np. wtrącenie, wyrażenie parentetyczne, fraza przymiotnikowa, fraza wykrzyknikowa itp. Wydaje się jednak, że ograniczenie się do podanych wyżej realizacji nie zubaża zbytnio przedstawionego tu opisu — zwłaszcza że fraza luźna jest tylko jednym z typów rozważanych w tej pracy konstrukcji, wcale nie najważniejszym.

Szczególnym przypadkiem takiego członu jest także początkowy lub inkorporacyjny spójnik w zdaniu samodzielnym, jak np. w poniższym wypowiedniku:

(65) *\I\ <trzeba> (to Jance koniecznie wytłumaczyć).* *[2978]*

9.1.3 Spójniki na granicy wypowiedników

Pokrewnym problemem jest też, nie będący błędem, fakt funkcjonowania w języku spójników jako partykuł wzmacniających, jak np. w treści następującego wypowiednika:

(66) *\A\ [co] <się stało>?* *[3721]*

[3] [Świdziński, 1992a], s. 203.

lub członów spajających dłuższe fragmenty wypowiedzi, jak spójnik *ale* w poniższej sekwencji:

(67) */Później/ <należałoby> (dać sylwetkę jakiego starego działacza, możliwie żyjącego jeszcze, którego nazwisko miałoby szansę patronować kiedyś naszej instytucji).* *[363]*

(68) *..., /którego/ [nazwisko] <miałoby> (szansę patronować kiedyś naszej instytucji).* *[364]*

(69) {*Ale okres ten ujęłoby się raczej krótko*} {,} {*bądź co bądź czasy to aż nadto zamierzchłe i współczesnemu czytelnikowi mało mówią*} *[365]*

Poprawna analiza tak reprezentowanego wypowiednika wymaga zastosowania jednego z trzech sposobów: dopuszczenia zdań złożonych z potencjalnie wielu wypowiedników, interpretacji konstrukcji spajających jako luźnych albo usunięcia członu łączącego z początku drugiego składnika. Rozwiązanie pierwsze wymagałoby zastosowania mechanizmów wykraczających poza gramatykę zdań, drugie — proste i jak najbardziej poprawne — wprowadzenia dodatkowego stopnia swobody nadmiernie zwiększającego liczbę analiz w zdaniach w zasadzie jednoznacznych. Najnaturalniejsze wydaje się rozwiązanie trzecie, gdyż stanowi konsekwencję przyjętego w korpusie wypowiedników założenia odcinania spójnika przy podziale wypowiednika złożonego na składowe. W oczywisty sposób możemy w tym przypadku założyć równoważność wypowiednika oryginalnego z jego wariantem nie zawierającym tak określonego członu luźnego.

Fragmenty treści wypowiedników będące tego rodzaju członami, często niezależne od struktury frazowej wypowiednika zostały również oznaczone poprzez ujęcie ich w parę znaków gwiazdki:

(70) */*I* teraz/ (małowiele) <czuję>.* *[239]*

9.1.4 Błędna interpunkcja wypowiedników składowych

Ze względu na niekonsekwencje zapisu interpunkcji wypowiedniki były ponadto dodatkowo przetwarzane w celu utworzenia poprawnych wypowiedzeń. Przypadek ten dotyczył zarówno wypowiedników samodzielnych, często pozbawionych interpunkcji końcowej (!) lub zakończonych błędnie (np. przecinkiem)[4], jak też wypowiedników składowych.

Przykładowo, dla zestawu:

(71) *A ja skończyłam weterynarię, ale...* *[2447]*

(72) *\A\ [ja] <skończyłam> (weterynarię),...* *[2448]*

[4]Przypadki te zostały uznane za usterki typograficzne tego samego rzędu co literówki i zostały poprawione; szczegółową listę wystąpień zawiera rozdział C.1 — patrz s. 197.

analiza oryginalnego kształtu wypowiednika (72), oczywiście po usunięciu struktury frazowej, nie ma szans zakończyć się powodzeniem, gdyż po odcięciu wielokropka (który pełni w przypadku wypowiedników podrzędnych funkcję oznaczenia podrzędności) otrzymujemy tekst zakończony przecinkiem. W podobnych przypadkach przed skierowaniem wypowiednika do analizy automatycznej końcowy przecinek lub średnik był zamieniany na kropkę.

Inny problem stanowi błędny zapis końcowych znaków przestankowych w wypowiednikach składowych. Spójrzmy:

(73) *A jutro co jest, niedziela?* *[2829]*

(74) *\A\ (jutro) [co] <jest>...* *[2830]*

Znak zapytania powinien znaleźć się także w zapisie wypowiednika składowego. Jest to o tyle istotne, że wypowiednik nadrzędny (jako zawierający składnik oznajmieniowy) nie znajdzie się w grupie zdań analizowanych automatycznie (patrz rozdział 8.2, s. 84).

Podobnie w przypadku przeciwnym, gdy elipsa pytająca tworzy tzw. *pytania rozłączne* (ang. i u Świdzińskiego *question tags*, tu: człon *„dobrze?”* w przykładzie 75):

(75) *<Niech...> (1 mu 1) [pan] <...powie> (2 , żeby do mnie zajrzał, jak się znowu pokaże 2) /, dobrze/?* *[2808]*

(76) *=żeby= (do mnie) <zajrzał> /, jak się znowu pokaże/,...* *[2809]*

Wypowiednik podrzędny (76) nie jest już (słusznie) zakończony znakiem zapytania, jednak usunięcie członu pytającego z wypowiednika nadrzędnego (75) przed skierowaniem go do analizy musi skutkować również zmianą znaku końca na kropkę. Jest to, być może, niekonsekwencja sposobu oznaczania tego rodzaju konstrukcji — znak zapytania wydaje się być nieodłączną częścią pytania rozłącznego, mógłby zatem znaleźć się wewnątrz oznaczenia frazowego. Podobne usterki zostały poprawione bezpośrednio w tekstach korpusu.

9.1.5 Względne i pytajnozależne wypowiedniki podrzędne

Jeszcze inny rodzaj trudności wynika z użycia w wypowiedniku podrzędnym dokładnej postaci tekstowej wypowiednika nadrzędnego, także wówczas, gdy wypowiednik podrzędny jest pytajnozależną lub względną realizacją frazy zdaniowej. Spójrzmy na fragment korpusu z opisem wypowiednika zawierającego składnik pytajnozależny, funkcjonujący następnie jako samodzielny wypowiednik:

(77) *<Można> (mówić, co się chce).* *[966]*

(78) *..., (co) [się] <chce>.* *[967]*

Wypowiednik podrzędny został zapisany bez znaku zapytania, co wskutek ortograficznego zapisu pytajności w GFJP uniemożliwia jego poprawną analizę automatyczną. W celu poprawienia tego zapisu zamieniam w tego rodzaju przypadkach końcowy znak kropki na znak zapytania. Wypowiedniki, w których taka zamiana jest konieczna, są łatwo identyfikowalne — parametr typu składnika podrzędnego ma w ich przypadku wartość `PZ: x`, gdzie `x` jest nazwą pytajnika.

W przypadku wypowiednika zawierającego zaimek względny:

(79) [*Trzy aminokwasy, mianowicie cysteina, cystyna, która powstaje z połączenia dwóch cząsteczek cysteiny, oraz metionina*] <*zawierają*> *(siarkę).* *[4061]*

(80) *..., [która]* <*powstaje*> *(z połączenia dwóch cząsteczek),...* *[4062]*

warunkiem wystarczającym do jego poprawnego przetworzenia jest usunięcie składnika względnego treści, co jest, być może, zgodne z zamysłem Świdzińskiego, gdyż w szczątkowym zakresie zostało dokonane przez edytorów poprzez oznaczenie członu względnego jako nie należącego do treści. Poprawki tej dokonałem w korpusie ręcznie, ujmując zaimki względne w gwiazdki.

W przypadku ogólnym decyzja o usunięciu członu względnego nie jest właściwa, gdyż może prowadzić do niepoprawności składniowej — gdy np. jego wystąpienie spełnia wymaganie czasownika. Rozwiązaniem poprawnym mogłoby być w tym przypadku traktowanie treści wypowiednika jako frazy zdaniowej, jednak zdecydowałem się na rozwiązanie prostsze, umożliwiające jednorodne traktowanie wypowiedników oznaczonych jako zdaniowe (w opisie korpusu wypowiednik zdaniowy oznacza zawsze zdanie).

9.1.6 Wypowiedniki z członem aglutynacyjnym

Podobny problem powstaje w przypadku wypowiedników zawierających człon aglutynacyjny (11 przypadków):

(81) <*Trzeba*> *(cię tak pokazać, żebyś wszedł na listę).* *[706]*

(82) *..., =żeby=* <*-ś... wszedł*> *(na listę).* *[707]*

Tego rodzaju wypowiedniki podrzędne nie są samodzielne i nie mogą zostać zanalizowane bez dokonania ingerencji w tekst. Najmniejszą możliwą w tym przypadku zmianą jest sklejenie aglutynantu z formą pseudoimiesłowu, a następnie dokonanie analizy tak przekształconego tekstu wypowiednika. Dla przykładów korpusowych cel ten został osiągnięty bez modyfikacji treści próbek, lecz z wykorzystaniem dodatkowej listy wypowiedników zawierających zmienioną postać tekstu.

9.2 Uzupełnianie i zamiana analiz morfologicznych

Problem z analizą części wypowiedników spowodowany był także specyfiką tworzących je form wyrazowych, klasyfikowanych przez Morfeusza „niewłaściwie" w kontekście analizy z wykorzystaniem GFJP. Poniżej poruszam także problem jednostek wielowyrazowych, obsłużony na granicy analizy morfologicznej i składniowej.

9.2.1 Zmiany i rozszerzenia kategoryzacji form

Już przy pierwszych próbach zastosowania analizy składniowej tekstów wypowiedników okazało się, że wymaga ona gdzieniegdzie ingerencji w charakterystykę morfologiczną słowoform — czy to poprzez uzupełnienie analiz Morfeusza, czy poprzez zamianę analizy (lub analiz) na właściwą, czyli wymaganą do akceptacji przez GFJP (niekoniecznie oznacza to, że analiza oryginalna była błędna w rozumieniu „intuicyjnym"). Specyfika problemu sprawiła, że większość przypadków tego rodzaju była możliwa do wykrycia wyłącznie wyrywkowo, podczas eksperymentów z analizą składniową przykładów korpusowych.

Uzupełnienie analiz polegało albo na rozszerzeniu listy wartości którejś z kategorii morfologicznych (podaję kolejno formę analizowaną, jej postać hasłową i analizę zwróconą przez Morfeusza oraz postać hasłową i analizę zmienioną):

kajdanki	kajdanki	`subst:pl:nom:[p1, p2|p3]`
→	kajdanki	`subst:pl:[nom,acc|voc]:[p1, p2|p3]`

albo na zarejestrowaniu całkiem nowej analizy:

krzywa	krzywy	`adj:sg:nom:f:pos`
→	krzywy	`adj:sg:nom:f:pos`
→	krzywa	`subst:sg:nom:f`
szczecina	szczecina	`subst:sg:nom:f`
→	szczecina	`subst:sg:nom:f`
→	Szczecin	`subst:sg:gen:m3`

Proces zastępowania ograniczył się do operacji na pojedynczych analizach słów:

całkiem	całkiem	`qub`
→	całkiem	`adv:pos`
koło	koło	`qub`
→	koło	`prep:gen`

z których największą grupę stanowią właśnie partykuło-przysłówki[5]. Wydaje się to zbytnim ograniczeniem dla form takich jak *całkiem*, *nadal* czy *nieco*, będących

[5]Definicję tej klasy leksemów zawiera rozdział 8.1.5, s. 81.

składniowymi równoważnikami przysłówków. W tym kontekście należy zwrócić uwagę na nieprecyzyjność zadomowionego już w językoznawstwie polskim określenia *partykuło-przysłówek*, grupującego leksemy o różnych funkcjach składniowych.

W moim przekonaniu jest to niekonsekwencja przyjętego systemu znaczników morfosyntaktycznych, gdyż z jednej strony Woliński pisze definiując ten system[6]:

> Kryteria czysto fleksyjne nie pozwalają na podzielenie leksemów nieodmiennych na takie klasy jak przyimki, spójniki i partykuło-przysłówki. Rozróżnienie miedzy klasami nieodmiennymi odbywa się więc na zasadzie różnic własności składniowych. Jako przyimki wyróżniono leksemy pełniące w wypowiedzeniu funkcję łączącą i wymagające określonego przypadka. Za spójniki uznano leksemy pełniące funkcję łączącą, ale nie wymagające określonego przypadka. Pozostałe leksemy nieodmienne należące do systemu leksykalnego polszczyzny zaliczono do resztkowej klasy kublików (odpowiadającej mniej więcej partykuło-przysłówkom).

z drugiej zaś równoważy taksonomicznie formy o różnej dystrybucji (dodatkowo zaliczając do klasy kublików również zaimek zwrotny *się* o funkcji jeszcze innej od przysłówkowej czy partykułowej).

Aby umożliwić akceptację częstych konstrukcji tego rodzaju zmieniłem klasyfikację niektórych jednostek tego typu na przysłówkową. Rozwiązanie to zostało wybrane arbitralnie, jako wygodniejsze od — wydaje się — właściwszego zwiększenia pojemności frazy przysłówkowej GFJP w celu akceptacji realizacji partykułowej (której odpowiada kategoria kublika). Za wprowadzeniem pierwszego rozwiązania przemawia potencjalnie duża liczba nadmiarowych analiz dla tradycyjnych partykuł w rodzaju *nie*, za drugim — opis frazy przysłówkowej GFJP[7]:

> Fraza przysłówkowa realizowana jest w postaci najprostszej przez formę przysłówkową, tzn. albo przez formę przysłówkową przymiotnika, albo przez formę partykuło-przysłówka; np. dawno, dość.

Co ciekawe, konstrukcja przysłówkowa w GFJP nie zawiera realizacji kublikowej, a jedynie przysłówkową. Wobec braku definicji jednostek elementarnych oznacza to, że albo intencją Świdzińskiego było potraktowanie partykuło-przysłówków jako form przysłówkowych (mniej prawdopodobne ze względu na obecność parametru stopnia w definicji elementu `formaprzysł`), albo że założenie z opisu nie znalazło właściwej realizacji. Do tej drugiej interpretacji zdaje się też nawiązywać Woliński: dodając reguły elementarne jednostce `formaprzysł` przypisuje realizację ściśle przysłówkową, o nadrzędnej kategorii `adv`[8], nie zapisuje natomiast realizacji kublikowej.

Inny ciekawy przypadek pokrewny ilustruje zdanie:

(83) *Od dzisiaj jesteś w tym pokoju zupełnie skończony.* *[991]*

[6] [Woliński, 2003], s. 52.
[7] [Świdziński, 1992a], s. 77.
[8] [Woliński, 2004], s. 61–63.

Określeniom czasu zwykle uznawanym za przysłówkowe (jak *wczoraj*, *dziś*, *jutro*) Morfeusz przypisuje wyłącznie klasyfikację kublikową. Jest to interpretacja niepełna, zasługująca na uzupełnienie o kategoryzację rzeczownikową (zgodną ze słownikiem języka polskiego, np. Szymczaka [SJP, 2002]), odpowiadającą dystrybucji rzeczownikowych określeń czasu:

(84) *Nie będzie mnie od poniedziałku.*

(85) *Od miesiąca go nie widziałem.*

Uzupełnienia lub zamiany analiz Morfeusza (także w przypadku kodów form nie rozpoznawanych przez Morfeusza) dokonałem, ze względu na niedostępność narzędzi do ingerencji w słownik analizatora morfologicznego, definiując włączoną następnie do plików Świgry funkcję w języku Prolog, której działanie polega na dodatkowym przetworzeniu listy analiz Morfeusza i korekcie kodów — w sposób maksymalnie nieinwazyjny dla procedury analizy morfologicznej:

```
zamien([i(A, B, 'encefalograf', [], []) | Tail],
       [i(A, B, 'encefalograf', 'encefalograf',
                 subst:sg:nom:m3) | NewTail])
      :- zamien(Tail, NewTail), !.

zamien([i(A, B, kajdanki, kajdanki,
                subst:pl:nom:[p1, p2|p3]) | Tail],
       [i(A, B, kajdanki, kajdanki,
                subst:pl:[nom,acc|voc]:[p1, p2|p3]) | NewTail])
:- zamien(Tail, NewTail), !.

zamien([i(A, B, szczecina, szczecina, subst:sg:nom:f) | Tail],
       [i(A, B, szczecina, szczecina, subst:sg:nom:f),
        i(A, B, szczecina, szczecin, subst:sg:gen:m3) | NewTail])
:- zamien(Tail, NewTail), !.
...
zamien([Head|Tail], [Head|NewTail]) :- zamien(Tail, NewTail).
zamien([], []).
```

Rozwiązaniem wariantowym mogłoby być użycie słownika wyjątków Świgry — umożliwiającego zamianę kodów morfologicznych mechanizmu wbudowanego w predykat `getinput`, za pomocą którego analizator pobiera kolejne elementy wejścia. Metoda ta w zaproponowanym kształcie nie nadaje się jednak do bezpośredniego wykorzystania, gdyż uniemożliwia dodawanie wariantów wyników o innym haśle, a ponadto zakłada, że analiza morfologiczna zwróciła co najmniej jeden wynik, co uniemożliwia jej użycie w przypadkach słów Morfeuszowi nie znanych (patrz rozdział 6.3, s. 59). Prostszy i być może mniej elegancki mechanizm zaproponowany przeze mnie wad tych nie posiada, zdecydowałem się zatem stosować go konsekwentnie do modyfikacji wyników analizy morfologicznej zawsze, gdy zajdzie taka konieczność (por. rozdziały 9.2.2 — s. 96 i 11.2.12 — s. 134).

9.2.2 Analiza jednostek wielowyrazowych

W inicjalnym kształcie Morfeusz nie dostarcza mechanizmów wspierających analizę wielowyrazowych jednostek leksykalnych o stałej łączliwości. Zawierać je może wykorzystywany przez GFJP *leksykon*, który w rozumieniu Świdzińskiego mieści także odniesienia do jednostek złożonych[9]:

> Zakładam roboczo, że dysponuję gotowym spisem realizacji leksykalnych tych jednostek składniowych, które występują jawnie w opisie. Spis ten nazwać można leksykonem; rozwiązanie takie zostało przyjęte i uzasadnione w dwu pracach Szpakowicza i Świdzińskiego. Leksykon zawiera wszystkie elementy terminalne, ale również ich dystrybucyjne ekwiwalenty o dowolnym stopniu złożoności. Zadaniem leksykonu jest symulowanie analizy jednostek składniowych, które nie są szczegółowo definiowane.

ale Woliński posługuje się jednostkami nie przekraczającymi granic słów:

> Odpowiednik tak [tj. jak w powyższym opisie Świdzińskiego] rozumianego leksykonu w programie Świgra praktycznie nie obejmuje konstrukcji złożonych (z wyjątkiem tzw. analitycznych form fleksyjnych). Uwzględnienie złożonych realizacji jednostek elementarnych GFJP w programie komputerowym wymagałoby rozbudowy reguł gramatyki i zapewne wprowadzenia jakiegoś nowego komponentu przetwarzającego frazeologizmy.

Tymczasem do zastosowań prostszych, jakim jest właśnie opisywane użycie analizatora dla zamkniętego zbioru tekstów, można także w Świgrze próbować skorzystać z metody Świdzińskiego. Nawiązując do koncepcji leksykonu, interpretacji morfologicznej stałych związków wielowyrazowych dokonuję więc na granicy analizy morfologicznej i składniowej. Z technicznego punktu widzenia wykorzystuję do tego celu znaną funkcję modyfikującą analizy morfologiczne: jedną z oferowanych przezeń możliwości jest też zdolność sklejania wielu analiz Morfeusza w pojedynczą analizę wynikową o ustalonym haśle i wartościach kategorii morfologicznych. Oto przykład reguły umożliwiającej tego rodzaju operację, przypisującej zestawowi pojedynczych form *na*, *łapu* i *capu* zupełnie odrębną kategorię przysłówka leksemu złożonego *na łapu capu*:

```
zamien([i(A, B, na, na, prep:[acc|loc]),
        i(B, C, łapu, łap, subst:sg:gen:m3),
        i(C, D, capu, capu, qub) | Tail],
       [i(A, D, 'na łapu capu', 'na łapu capu', adv:pos) | NewTail])
:- zamien(Tail, NewTail), !.
```

[9] [Świdziński, 1992a], s. 56.

Zdefiniowany w ten sposób „leksykon" liczy 29 reguł-rekordów i zawiera oprócz wielowyrazowych jednostek nieodmiennych w rodzaju przytoczonej także realizacje jednostek reprezentujących nazwy geograficzne czy właśnie frazeologizmy.

O akceptowalności takiego rozwiązania zdają się świadczyć podobne wnioski, do jakich dochodzi Rudolf[10] dla pewnej klasy omawianych jednostek — „burkinostek" od jednostki *Burkina Faso*[11]. W celu przypisania im charakterystyki fleksyjnej, często znacząco różnej od własności poszczególnych składników, autor również postuluje wprowadzenie dodatkowego etapu przetwarzania po fazie analizy morfologicznej.

Warto zauważyć, że podobna idea została zrealizowana w korpusie słownika frekwencyjnego, w którym wyrazy jednej słowoformy analitycznej oddzielone zostały specjalnym symbolem `[+]`[12], a kod fleksyjny przypisywany był całej formie poprzez umieszczenie go po ostatnim członie. Oto fragment jednej z próbek tego korpusu[13]:

```
Warto przypomnieć, że słynne[241] trzęsienie[141] ziemi[121],
które[211] w[66] roku[161] tysiąc dziewięćset szóstym[261]
zniszczyło San[+] Francisco[/][141] obliczono na[64] osiem[34]
i dwadzieścia[34] pięć[34] setnych[122] w[66] tej[261] skali[161].
```

Dodatkowe oznaczenia nie zostały natomiast użyte w korpusie IPI PAN, bazującym m. in. na korpusie słownika frekwencyjnego. Oto fragment zacytowanej wyżej próbki oznakowany zgodnie z zasadami stosowanymi w korpusie IPI PAN:

```
...
<tok>
  <orth>zniszczyło</orth>
  <lex disamb="1">
    <base>zniszczyć</base>
    <ctag>praet:sg:n:perf</ctag>
  </lex>
</tok>
<tok>
  <orth>San</orth>
  <lex disamb="1">
    <base>san</base>
    <ctag>subst:sg:acc:n</ctag>
  </lex>
</tok>
<tok>
  <orth>Francisco</orth>
  <lex disamb="1">
    <base>francisco</base>
```

[10] [Rudolf, 2004], s. 32–34.

[11] Termin wprowadzony w artykule [Derwojedowa i Rudolf, 2003].

[12] Zwięzły opis oznaczeń metatekstowych korpusu słownika frekwencyjnego zawiera praca [Nazarczuk, 1997].

[13] Styl B — drobnych wiadomości prasowych, próbka nr 1722 (Trybuna Opolska, 31.03.1964, str. 1, kol. 5).

```
    <ctag>subst:sg:acc:n</ctag>
  </lex>
</tok>
...
```

Koncepcja ta, zgodnie z zasadą „maksymalny segment to słowo", traktuje wyrażenia tego rodzaju jak opisywane dalej związki apozycyjne (por. rozdział 10.3.1, s. 114), dopuszczając osobne funkcjonowanie każdego z członów, co w tym przypadku może wydawać się błędem.

Rozdział 10

Rozszerzenie gramatyki

Analiza składniowa danych korpusowych musiałaby zostać silnie ograniczona, gdyby formalna weryfikacja akceptowalności zdań korpusowych z wykorzystaniem dostępnego parsera Wolińskiego nie była wstępem do wynikającego z jego wstępnych wyników koniecznego rozszerzenia gramatyki Świdzińskiego o mechanizmy umożliwiające akceptację większości zdań intuicyjnie poprawnych. Poniżej opisuję zastosowane w tym celu metody i przedstawiam wycinki nowego wariantu gramatyki. Tekst rozdziału stanowi znaczne rozszerzenie artykułu [Ogrodniczuk, 2005a].

W opisie reguł zmienionych cytuję także w większości przypadków oryginalne brzmienie; reguł usuniętych nie podaję zakładając dostępność gramatyki. Reguły nowe dodaję z zachowaniem ciągłości numeracji — po klauzulach danej jednostki lub we wnętrzu sekcji, wówczas z użyciem oznaczeń literowych.

10.1 Konstrukcja liczebnikowa

GFJP nie zawiera definicji konstrukcji liczebnikowej. Świdziński pisze[1]:

> Fraza nominalna realizowana jest w postaci najprostszej przez formę rzeczownika, ale także liczebnika w rozumieniu Saloniego, a nawet przymiotnika; np. *to*, *Jaś*, *okazję*. [...] W tej pracy decyduję się pominąć frazy nominalne z centrum liczebnikowym, których budowa jest, jak wiadomo, wysoce specyficzna.

W istocie, frazy liczebnikowe (rozumiane syntaktycznie jako równoważniki dystrybucyjne formy liczebnikowej) cechują osobliwe własności składniowe — narzucanie mnogości, występowanie lub brak uzgodnienia zależne od klasy liczebnika, ustalony porządek i jednorodność składników konstrukcji złożonych przy reprezentatywności ostatniego składnika. Wszystkie te problemy zostały zaadresowane w opisie formalnym, który dodałem do GFJP posługując się materiałami z artykułów

[1] [Świdziński, 1992a], s. 77 i 240.

[Saloni i Gruszczyński, 1978], [Derwojedowa i in., 2003] oraz z poświęconej liczebnikom sesji wykładu monograficznego Świdzińskiego *Wprowadzenie do gramatyki formalnej*. Opis ten jest z założenia utylitarny w rozumieniu ograniczenia do przypadków występujących w korpusie wypowiedników, nie zajmuję się zatem konstrukcjami wykraczającymi poza tradycyjną listą leksemów liczebnikowych (nie uwzględniam np. możliwego liczebnikowego opisu leksemów innych rodzajów[2]) czy wchodzących w zależności nie ograniczające się do grupy nominalnej, czasownika i ewentualnego określenia atrybutywnego[3].

10.1.1 Formy liczebnikowe i rozszerzona kategoria akomodacyjności

Problemem pierwotnym jest opis jednostek elementarnych — form liczebnikowych. Liczebnik jako jednostka wchodząca w skład konstrukcji nominalnej posiada kategorię fleksyjną rodzaju, a na mocy uzależnienia od słowa czasownikowego — kategorię przypadka (przyjmuję, że formy liczebników posiadają też ustaloną wartość kategorii liczby — mnogiej).

Analiza związków konstrukcji liczebnikowej z nominalną wykazuje pewne niestandardowe zachowanie leksemów liczebnikowych: dla niektórych słów liczebnikowych wartość kategorii fleksyjnej przypadka nie jest rządzona formą rzeczownika nadrzędnej grupy nominalnej. Zwyczajowo podawany przykład ilustruje to zjawisko za pomocą zdań[4]:

(86) *Przyszło pięć kobiet.*

(87) *Przyszło pięciu mężczyzn.*

Przypadek formy rzeczownika jest wówczas równy dopełniaczowi, zaś zgoda frazy liczebnikowo-nominalnej z mianownikową formą czasownika zapewniona zostaje przez nadrzędną dystrybucyjnie formę liczebnikową[5].

Zachowanie to można w dużym, aczkolwiek wystarczającym na potrzeby przyjętego rozwiązania uproszczeniu opisać dzieląc słowa liczebnikowe — egzemplarze leksemów liczebnikowych — na dwie klasy, tj. wprowadzając dodatkową kategorię gramatyczną dla liczebników, odpowiadającą występowaniu lub brakowi uzgodnienia przypadka zawierającej dane słowo liczebnikowe konstrukcji liczebnikowo-nominalnej z nadrzędną konstrukcją czasownikową. Wartościami tej klasy będą odpowiednio `congr` dla występowania uzgodnienia i `rec` dla jego braku, przejęte z artykułu [Bień i Saloni, 1982][6], gdzie zbliżona kategoria (przysługująca jednak wyłącznie formom liczebnikowym rodzaju męskoosobowego w mianowniku) nosi nazwę

[2]Ich interesujący wykaz znalazł się w pracy [Saloni, 1977], s. 156–157.

[3]W artykule [Saloni i Gruszczyński, 1978] znaleźć można wykaz tworów o jeszcze szerszej realizacji — m. in. konstrukcje zawierające przysłówkowe podrzędniki dystrybucyjne w rodzaju *dokładnie dwadzieścia pięć* czy *w osiągalnym przybliżeniu pięćset lub sześćset*.

[4]Tu cytuję za [Saloni, 1977], s. 151.

[5]Saloni dyskutuje szczegółowo kwestię kwalifikacji przypadka dla formy liczebnika — tamże, s. 154.

[6]Patrz s. 37.

akomodacyjności; proponuję zachowanie tej nazwy na oznaczenie wprowadzanego rozszerzonego rozumienia kategorii opisującej wspomniane uzgodnienie.

Oto wartości kategorii akomodacyjności dla przykładowych zdań (pierwsze trzy ilustrują pokrywanie się kategorii z rozumieniem Bienia i Saloniego):

(88) *Przyszli dwaj chłopcy. (`congr`)*

(89) *Przyszło dwóch chłopców. (`rec`)*

(90) *Przyszło dwu chłopców. (`rec`)*

(91) *Przyszły dwie dziewczyny. (`congr`)*

(92) *Przyszło siedem dziewczyn. (`rec`)*

Zgodnie z artykułem [Bień i Saloni, 1982] (na podstawie którego powstał zaproponowany w artykule [Woliński i Przepiórkowski, 2001] system znaczników morfosyntaktycznych wykorzystany w korpusie IPI PAN[7]) działa też analizator Morfeusz. W późniejszym artykule [Woliński, 2003] postuluje się przypisywanie wartości akomodacyjności wszystkim formom liczebnikowym, jednak raczej ze względu na łatwiejsze przetwarzanie wyników analizy morfologicznej niż wskutek zmiany znaczenia kategorii; założenie to nie zostało w każdym razie zrealizowane.

Wartość kategorii akomodacyjności w bieżącym rozumieniu zależy od wartości kategorii przypadka i rodzaju danej formy liczebnikowej, konieczne jest zatem podanie pełnego paradygmatu leksemów zaliczanych do klasy liczebników. Podaję go w skrótowej formie, łącząc wartości kategorii gramatycznych, dla których występuje pełna neutralizacja. Wartość parametru akomodacyjności („+” dla występowania uzgodnienia — wartość `congr` i „−” dla jego braku — wartość `rec`) dodaję do tabeli odmiany zbudowanej na podstawie danych z artykułu [Saloni, 1977][8]; z niego pochodzą też oznaczenia klas poszczególnych form:

[7] Por. [Przepiórkowski, 2004], s. 22–25.
[8] Por. s. 154–155.

I	nom acc	m2 m3 n2	*dwa* +	*trzy* +	*cztery* +	*pięć* −
	nom acc	f	*dwie* +			
II	nom	m1	*dwaj* + *dwóch/dwu* −	*trzej* + *trzech* −	*czterej* + *czterech* −	*pięciu* −
	acc	m1	*dwóch/dwu* +	*trzech* +	*czterech* +	*pięciu* +
	gen loc	m1 m2 m3 n2 f				
	dat	m1 m2 m3 n2 f	*dwu/dwom* +	*trzem* +	*czterem* +	
IIa	inst	m1 m2 m3 n2	*dwoma* +	*trzema* +	*czterema* +	*pięcioma* +
	inst	f	*dwiema* +			
III	nom acc	n1 p1 p2	*dwoje* −	*troje* −	*czworo* −	*pięcioro* −
IIIa	gen	n1 p1 p2	*dwojga* −	*trojga* −	*czworga* −	*pięciorga* −
IIIb	dat loc	n1 p1 p2	*dwojgu* −/+	*trojgu* −/+	*czworgu* −/+	*pięciorgu* −/+
IIIc	inst	n1 p1 p2	*dwojgiem* −	*trojgiem* −	*czworgiem* −	*pięciorgiem* −

Wartość parametru akomodacyjności dla liczebników o wartości liczbowej wyższej niż 5 jest identyczna z podanym wzorcem dla liczebnika *pięć*; jak *dwa* odmieniają się ponadto liczebniki *oba* i *obydwa*.

Na dodatkową dyskusję zasługuje też wartość tego parametru dla biernika rodzaju męskoosobowego (`m1`) i dopełniacza wszystkich rodzajów męskich, nijakiego zwykłego (`n2`) i żeńskiego. Jej ustalenie jest w zasadzie arbitralne, gdyż przyjęcie wartości oznaczającej niezgodność przypadka implikowałoby podczas analizy wybór ścieżki ustalającej wartość kategorii przypadka na dopełniacz i analiza dokonałaby się poprawnie tak dla dopełniacza, jak i dla biernika (równemu w tym wypadku dopełniaczowi). Rozwiązaniem kanonicznym wydaje się jednak ustalenie zgodnej wartości akomodacyjności, stąd też wspomniane rozróżnienie.

Realizując technicznie wartość akomodacyjności zakładam jej słownikowość, nie dodaję więc dodatkowego warunku w definicjach reguł gramatyki, lecz przedefiniowuję morfeuszowe wyniki analizy morfologicznej liczebników zgodnie z powyższą tabelką, wzbogacając je, gdzie to konieczne, o dodatkowy parametr. Ponadto uzupełniam listę wyników Morfeusza o formy liczebników zbiorowych.

Kategoria akomodacyjności rozciąga się w oczywisty sposób na liczebniki tradycyjnie nazywane złożonymi (składnik końcowy determinuje akomodacyjność złożenia). W ich definicji przejmuję z niewielkimi zmianami klasyfikację z artykułu [Derwojedowa i in., 2003][9]. W oryginalnym kształcie używam definicji jednostek nazywanych w artykule *ten-numerals*, czyli reprezentujących liczby z zakresu 1–99. Zgodnie z koncepcją *trójek podstawowych* Saloniego[10] modyfikuję natomiast znaczenie jednostek z klasy *hundred-numerals*, włączając do niej reprezentacje wszystkich liczb z przedziału 1–999. Zabieg ten ma na celu uproszczenie opisu jednostki frazowej, która w ogólnej postaci składa się z trójek podstawowych przedzielonych formami leksemów odpowiadających kolejnym potęgom liczby 1000 (*tysiące*, *miliony* itd.) W zastosowanym rozwiązaniu ograniczam się do reprezentacji liczb naturalnych do wartości poniżej miliona, co pozwala znacznie (choć nie ponad miarę) uprościć opis składniowy.

Oto lista reguł elementarnych dla form liczebnikowych:

```
formalicz10(P, R/L, A)
--> s(n_licz1),
    [morf(_, H, num:Num:Cases:Gend:A)],
    { rowne(H, ['jeden', 'dwa', 'trzy', 'cztery', 'pięć', 'sześć',
                'siedem', 'osiem', 'dziewięć', 'dziesięć',
                'jedenaście', 'dwanaście', 'trzynaście',
                'czternaście', 'piętnaście', 'szesnaście',
                'siedemnaście', 'osiemnaście', 'dziewiętnaście',
                'dwadzieścia', 'trzydzieści', 'czterdzieści',
                'pięćdziesiąt', 'sześćdziesiąt', 'siedemdziesiąt',
                'osiemdziesiąt', 'dziewięćdziesiąt']),
      przypadki(Cases, P),
```

[9] Patrz pkt 2.3–2.5, s. 95–96.

[10] Patrz [Saloni i Gruszczyński, 1978], pkt 1.2, s. 24.

```
      liczba(Num, L),
      rodzaje(Gend, R) }.

formalicz10(P, R/L, A)
--> s(n_licz2),
    [morf(_, H1, num:_:_:_:_)],
    { rowne(H1, ['dwadzieścia', 'trzydzieści', 'czterdzieści',
                 'pięćdziesiąt', 'sześćdziesiąt', 'siedemdziesiąt',
                 'osiemdziesiąt', 'dziewięćdziesiąt']) },
    [morf(F, H2, num:Num:Cases:Gend:A)],
    { rowne(H2, ['jeden', 'dwa', 'trzy', 'cztery', 'pięć', 'sześć',
                 'siedem', 'osiem', 'dziewięć']),
      przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) }.

formalicz100(P, R/L, A)
--> s(n_licz3),
    formalicz10(P, R/L, A).

formalicz100(P, R/L, A)
--> s(n_licz4),
    [morf(_, H, num:Num:Cases:Gend:A)],
    { rowne(H, ['sto', 'dwieście', 'trzysta', 'czterysta', 'pięćset',
                'sześćset', 'siedemset', 'osiemset', 'dziewięćset',
                'kilkaset']),
      przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) }.

formalicz100(P, R/L, A)
--> s(n_licz5),
    [morf(_, H, num:Num:Cases:Gend:_)],
    { rowne(H, ['sto', 'dwieście', 'trzysta', 'czterysta', 'pięćset',
                'sześćset', 'siedemset', 'osiemset', 'dziewięćset']),
      przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) },
    formalicz10(P, R/L, A).

formalicz(P, R/L, A)
--> s(n_licz6),
    formalicz100(P, R/L, A).

formalicz(P, R/L, A)
--> s(n_licz7),
    [morf(_, 'tysiąc', subst:Num:Cases:Gend:A)],
```

```
    { przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) }.

formalicz(P, R/L, A)
--> s(n_licz8),
    [morf('tysiąc', _, num:Num:Cases:Gend:_)],
    { przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) },
    formalicz100(P, R/L, A).

formalicz(P, R/L, A)
--> s(n_licz9),
    formalicz100(P, R/L, A),
    [morf('tysiąc', _, num:Num:Cases:Gend:_)],
    { przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) }.

formalicz(P, R/L, A)
--> s(n_licz10),
    formalicz100(P, R/L, _),
    [morf('tysiąc', _, num:Num:Cases:Gend:_)],
    { przypadki(Cases, P),
      liczba(Num, L),
      rodzaje(Gend, R) },
    formalicz100(P, R/L, A).
```

10.1.2 Fraza liczebnikowo-nominalna

Tak zdefiniowane jednostki liczebnikowe wchodzą w uzgodnienia przede wszystkim z frazą nominalną (*„Dałem mu dwa dolary.”*) i przymiotnikową (*„Dałem mu dwa przedwojenne dolary.”*), uzupełniłem zatem gramatykę o odpowiednie realizacje frazy nominalnej. Możliwe związki z innymi konstrukcjami nietypowymi (przysłówkowymi, przyimkowymi) zaniedbuję jako mniej ważne wobec braku ich reprezentacji w korpusie. Ze względu na różnice dystrybucji fraz z konstrukcją liczebnikową i fraz bez liczebników (polegającą w skrócie na jednokrotności występowania liczebnika w konstrukcji frazy nominalnej) jednostka `fno` reprezentuje teraz frazę nominalną z możliwością wystąpienia liczebnika, czyli *liczebnikowo-nominalną*; dotychczasowa jednostka `fno` zostaje przemianowana na `fno1` — frazę nominalną właściwą.

Realizacja nominalna

Najprostszą realizację frazy liczebnikowo-nominalnej stanowi fraza nominalna właściwa:

```
fno(P, Rl, O, Neg, I, Z, Kl)
--> s(liczn1),
    fno1(P, Rl, O, Neg, I, Z, Kl).
```

Realizacje uzgadniające

Realizacje dla liczebników z klasy uzgadniającej mają postać formy liczebnika, ewentualnie uzupełnionej frazą nominalną zgodną w zakresie rodzaju i liczby:

```
fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn2),
    formalicz(P, Rl, congr).

fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn3),
    formalicz(P, Rl, congr),
    fno1(P, Rl, O, Neg, I1, Z1, Kl1).

fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn4),
    fno1(P, Rl, O, Neg, I1, Z, Kl1),
    formalicz(P, Rl, congr).
```

Realizacje nieuzgadniające niemianownikowe

Realizacje niemianownikowe dla liczebników z klasy nieuzgadniającej mają postać liczebnika w przypadku innym niż mianownik, ewentualnie uzupełnionego frazą nominalną właściwą w dopełniaczu:

```
fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn5),
    formalicz(P, Rl, rec),
    { rozne(P, mian) }.

fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn6),
    formalicz(P, Rl, rec),
    fno1(dop, Rl, O, Neg, I1, Z1, Kl1),
    { rozne(P, mian) }.

fno(P, Rl, O, Neg, I, z(_,[np]), rzecz)
--> s(liczn7),
    fno1(dop, Rl, O, Neg, I, Z, Kl1),
    formalicz(P, Rl, rec),
    { rozne(P, mian) }.
```

Realizacje nieuzgadniające mianownikowe

Realizacje dla liczebników z klasy nieuzgadniającej mają postać formy liczebnika w mianowniku, ewentualnie uzupełnionej frazą nominalną właściwą w dopełniaczu i/lub frazą przymiotnikową w mianowniku lub dopełniaczu[11]:

```
fno(mian, nij/poj, 3, Neg, I, z(_,[np]), rzecz)
--> s(liczn8),
    formalicz(mian, Rl, rec).

fno(mian, nij/poj, 3, Neg, I, z(_,[np]), rzecz)
--> s(liczn9),
    formalicz(mian, Rl, rec),
    fno1(dop, Rl, O, Neg, I1, Z1, Kl1).

fno(mian, nij/poj, 3, Neg, I, z(_,[np]), rzecz)
--> s(liczn10),
    fno1(dop, Rl, O, Neg, I, Z1, Kl1),
    formalicz(mian, Rl, rec).

fno(mian, nij/poj, 3, Neg, I, Z, rzecz)
--> s(liczn11),
    fpt(dop, Rl, St, Neg, I, Z, Kl),
    formalicz(mian, Rl, rec).

fno(mian, nij/poj, 3, Neg, I, Z, rzecz)
--> s(liczn12),
    fpt(dop, Rl, St, Neg, I, Z, Kl),
    formalicz(mian, Rl, rec),
    fno1(dop, Rl, O, Neg, I1, Z1, Kl1).

fno(mian, nij/poj, 3, Neg, I, Z, rzecz)
--> s(liczn13),
    fpt(mian, Rl, St, Neg, I, Z, Kl),
    formalicz(mian, R/L, rec),
    { rozne(R, mos), rozne(L, mno) }.

fno(mian, nij/poj, 3, Neg, I, Z, rzecz)
--> s(liczn14),
    fpt(mian, R/L, St, Neg, I, Z, Kl),
    formalicz(mian, R/L, rec),
    fno1(dop, Rl, O, Neg, I1, Z1, Kl1),
    { rozne(R, mos), rozne(L, mno) }.
```

[11] W artykule [Derwojedowa i in., 2003] takie połączenie (np. *tych pięć dziewczyn*) określane jest mianem frazy przymiotnikowo-liczebnikowej, co jest uzasadnione jej budową, ale już nie dystrybucją — jest to bowiem szczególna realizacja frazy nominalnej. Ze względu na dystrybucyjny charakter nazewnictwa w GFJP pomijam to określenie w tworzonym opisie.

10.1.3 Testy weryfikacyjne i analiza nowych wypowiedników

Materiału do wstępnych testów powstałego opisu gramatycznego dostarczył artykuł [Saloni, 1977], zawierający 89 numerowanych przykładów poprawnych zdań z konstrukcjami liczebnikowymi[12]. Analiza wszystkich przykładów zakończyła się powodzeniem; w następnym kroku przystąpiłem do analizy wypowiedników zawierających formy liczebników.

Frazy liczebnikowe występują w korpusie wypowiedników stosunkowo często — w 542 wypowiednikach, z czego aż w 492 zdaniowych, co stanowi 11,6% ogólnej liczby wypowiedników zdaniowych. Ze względu na wykluczenie z analizy wypowiedników zawierających niektóre konstrukcje nie przewidziane przez gramatykę (patrz rozdział 8.2, s. 84) analizie mogło zostać poddanych 298 nowych (w stosunku do analizy inicjalnej[13]) zdań. Sposób postępowania z liczebnikami w źródłowym korpusie słownika frekwencyjnego (liczby były zastępowane ich zapisem słownym) sprawił, że tekst wypowiedników nadaje się do analizy bez dodatkowego przetwarzania.

Na potrzeby analizy przykładów korpusowych usunąłem z gramatyki realizacje frazy nominalno-liczebnikowej w postaci pojedynczej formy liczebnikowej (*„Siedem szczekało."* czy *„Znam pięciu."* — reguły `liczn2`, `liczn5` i `liczn8`). Zawierające je zdania mają charakter zdecydowanie eliptyczny, podczas gdy kompletny opis stworzony wg założeń Świdzińskiego generuje bardzo dużo analiz nadmiarowych dla zdań zawierających nominalno-liczebnikową realizację frazy. Przykładowo, dla jednoznacznego zdania *„Szczekało siedem psów."* oprócz prawidłowej realizacji z wymaganą frazą podmiotową występuje aż 9 innych analiz z frazą luźną (wynikające z potraktowania formy liczebnikowej *siedem* jako podmiotu oraz uwzględnienia wieloznaczności rodzajowych).

10.2 Grupy składniowe

Struktura frazowa w gramatyce Świdzińskiego pozostawia wiele do życzenia jeśli chodzi o szczegółowość opisu. Zauważa to zresztą sam jej autor[14]:

> Lokalne hierarchie fraz nie są (...) tak istotne z punktu widzenia celów tej pracy, jak hierarchie jednostek zdaniowych: tylko jeden typ frazy, mianowicie fraza zdaniowa, zostanie tu opisany w sposób maksymalnie szczegółowy. (...) Dla uproszczenia wszystkie frazy z wyjątkiem frazy zdaniowej będą uważane za frazy elementarne, a więc niewspółrzędne.

Zgodnie z tym założeniem w opisie nie znalazły się np. złożone współrzędnie frazy nominalne takie jak np. *kot i pies*, wskutek czego zawierające je zdania nie są akceptowane przez bazujący na GFJP analizator. Formalny opis współrzędnych konstrukcji nominalnych powstał jednak na potrzeby artykułu [Szpakowicz i Świdziński, 1990],

[12] Zawierający je plik powstały po ekstrakcji przykładów z tekstu znajduje się m. in. na dołączonej do pracy płycie CD, patrz rozdział D.6, s. 211.

[13] Patrz rozdział 8.3, s. 85.

[14] [Świdziński, 1992a], rozdział 4.3.7, s. 77–78.

a jego uproszczona wersja została przeze mnie włączona do GFJP. Na bazie definicji dla grupy nominalnej rozbudowałem też opis frazy przysłówkowej, przymiotnikowej i przyimkowej (przyimkowo-nominalnej) umożliwiający akceptację przez gramatykę Świdzińskiego szerokiego zakresu konstrukcji współrzędnych.

Wszystkie definicje tworzyłem od razu w hierarchii uproszczonej, zgodnie z założeniem opisanym w poprzednim rozdziale. Jednostki reprezentujące grupy wykorzystują zatem funktory jednostek frazowych.

10.2.1 Definicja grupy nominalnej

Wspomniany artykuł Szpakowicza i Świdzińskiego wprowadza złożone jednostki nominalne na podobieństwo zdaniowych z GFJP: równorzędną grupę nominalną (`rgn`), grupę szeregową (`sgn`), jednorodną (`jgn`) i pojedynczą (`pgn`). Zbieżność konstrukcji jest celowa i ma według Świdzińskiego dobrze odpowiadać rzeczywistości językowej[15]. Także sam opis grupy nominalnej został stworzony z użyciem formalizmu podobnego do wykorzystywanego w GFJP, wobec czego może on zostać łatwo dołączony do definicji gramatyki wykorzystywanej przez analizator Świgra.

W artykule [Ogrodniczuk, 2005b] opisuję doświadczenia z pełnej implementacji oryginalnego opisu; na potrzeby korpusu wypowiedników zdecydowałem się mocno go uprościć, zachowując jednak wystarczającą siłę wyrazu. W porównaniu z pełnym opisem przyjmuję, że grupa zawsze uzgadnia się do wartości liczby mnogiej, nie jest zatem możliwa akceptacja zdań w rodzaju *„Nie tylko on, ale i ona przyszła”*, stanowiących notabene bardzo ciekawą konstrukcję uzgadniającą się w liczbie pojedynczej do członu nominalnego znajdującego się „bliżej czasownika”[16]. Nie biorę także pod uwagę szczątkowo wyróżnionego w opisie, lecz pozostawionego do późniejszego rozwinięcia pojęcia *klasy grupy nominalnej*, eliminuję zatem reguły opisujące konstrukcje z zaimkiem negatywnym i liczebnikiem dopełniaczowym. W opisie stosuję natomiast oczywiście niezbędne do poprawnej definicji uzgodnienia rodzaju, osoby i negacji selektywnej[17], obliczane na podstawie wartości odpowiednich parametrów członów składowych. Takie ograniczenie pozwoliło sprowadzić skomplikowaną definicję złożoną z 47 reguł do dziesięciu (dla zapewnienia choćby szczątkowej odpowiedniości zachowuję oryginalne oznaczenia reguł). Oto przykład reguły dla grupy równorzędnej:

```
fno(P, R/mno, 0, Neg, I, Z, Kl, _, rgn)
--> s(gno-r2),
```

[15]Por. [Świdziński, 1993a], s. 22. Świdziński uznaje to podobieństwo za pogłos idei „gramatyki jednej reguły” [Gazdar i in., 1985].

[16]Szczegółowy opis tej konstrukcji znalazł się we wspomnianym artykule [Ogrodniczuk, 2005b], patrz s. 180. Wydaje się, że ten rodzaj uzgodnienia wymaga jeszcze szczegółowych badań, gdyż niekiedy zdania utworzone na mocy powyższej reguły wydają się niezgodne z intuicją lingwistyczną (por. np. *„Powiedziała mu ona i on.”*). O komplikacji konstrukcji świadczy fakt, że dotyczy jej większość reguł opisu oryginalnego.

[17]Za [Szpakowicz i Świdziński, 1990], s. 8: „Odpowiednia wartość negacji selektywnej przysługuje grupie nominalnej ze względu na związki z grupą czasownikową (w szczególności finitywną). Negacja partykułowa jest zjawiskiem czysto ortograficznym: chodzi o obecność lub brak partykuły „nie” na początku danej grupy nominalnej.”.

```
spojr(l, Nr),
fno(P/R1/_/O1, Neg1, I, Z, Kl, _, Sgn1),
{ rowne(Sgn1, [sgn, jgn, fno, fno1, knodop, knopm, knoatr,
               knoink, knom]) },
przec,
spojr(p, Nr),
fno(P/R2/_/O2, Neg2, _, _, _, _, Sgn2),
{ rowne(Nr, [1, 2]),
  uzgr(R1, R2, R),
  uzgo(O1, O2, O),
  uzgsneg(Neg1, Neg2, Neg),
  rowne(Sgn2, [sgn, jgn, fno, fno1, knodop, knopm, knoatr,
               knoink, knom]) }.
```

Na przykładzie tej reguły warto zauważyć, że w definicji grupy nominalnej ujawnia się błędna decyzja połączenia kategorii rodzaju i liczby w jeden parametr, na co w innych kontekstach zwracają też uwagę Bień[18] i Woliński[19] — wynikowa kategoria rodzaju grupy musi zostać obliczona na podstawie kategorii rodzaju konstrukcji składowych.

Definicja wykorzystująca jednostkę frazową umożliwia włączenie grupy do GFJP bez dokonywania żadnych dodatkowych zabiegów. Jedynym wartym uściślenia szczegółem jest sposób zapisu pojedynczej grupy nominalnej, której definicja nie została w artykule podana: jest ona realizowana jako fraza nominalna w rozumieniu GFJP (jakkolwiek warto nadmienić, że taki opis frazy nominalnej jest znacznie prostszy od zakładanego przez autorów — pozostawionego do określenia w przyszłości).

Weryfikacja definicji

Weryfikacja definicji została przeprowadzona na zestawie zdań testowych. W związku z tym, że współrzędna grupa nominalna nie była przedmiotem opisu GFJP, nie istniał również dla niej gotowy zestaw testów; w rozdziale czwartym artykułu [Szpakowicz i Świdziński, 1990] podano jednak zdania przykładowe w formie zbliżonej do opisu z aneksu do pracy Świdzińskiego, mogły one zatem zostać wyekstrahowane i użyte na potrzeby testów tego i przyszłych opisów.

Zgodnie z tym założeniem na bazie tekstu artykułu Szpakowicza i Świdzińskiego utworzyłem plik zawierający 216 zdań testowych (213 różnych; 135 oznaczonych jako poprawne, 81 jako niepoprawne) zapisanych w formacie zgodnym z konwencją stosowaną w projekcie *Zestaw testów do weryfikacji i oceny analizatorów języka polskiego*[20] (każda próbka opatrzona numerem rozdziału tekstu, z której pochodzi, symbolem klauzuli, którą ilustruje numerem porządkowym oraz oznaczeniem poprawności/niepoprawności/wątpliwej poprawności) — oto przykładowe zdanie z pliku testowego, zamieszczonego także na dołączonej do pracy płycie[21]:

[18] [Bień, 1997b], s. 82.

[19] [Woliński, 2004], s. 93.

[20] Patrz [Bień, 2000]; format był już wykorzystywany na potrzeby testów zmian gramatyki, por. np. [Ogrodniczuk, 2005b].

[21] Patrz rozdział D.6, s. 211.

```
[R2; GNO; 1]
Zarówno chłopiec, jak i dziewczyna przyszli.

[*R2; GNO; 6]
Tak ojciec i matka, jak i córka przyszły.
```

Dla potrzeb powstałego podzbioru definicji równorzędnej grupy nominalnej stworzyłem również odnośny podzbiór przykładów testowych, zawierający 59 różnych zdań testowych (33 poprawne, 26 niepoprawnych; zbiór został również umieszczony na płycie).

Kolejnym testem była oczywiście wstępna analiza ciągów oznaczonych w korpusie wypowiedników jako frazy nominalne; w jej trakcie okazało się, że włączony opis wymaga jeszcze drobnych modyfikacji, np. na zamieszczonej w artykule liście spójników szeregowych nie znalazł się spójnik *czy*. Bez jego dodania nie była możliwa analiza zdań w rodzaju:

(93) *Dziewczęta z ludu, panny służące czy szwaczki były po innej cenie.* *[547]*

Spójnik *czy* został dodany do klasy, w której znajdował się już spójnik *bądź* o podobnej dystrybucji.

Nadmiarowość opisu

Krótkiego komentarza wymaga też na koniec kwestia nadmiarowości definicji Świdzińskiego i Szpakowicza, ujawniająca się w **każdym** przypadku analizy grupy nominalnej i prowadząca do wykładniczego wzrostu liczby drzew rozbioru w zależności od liczby grup nominalnych w analizowanym tekście.

Opis zawiera m. in. następujące definicje jednostek (notacja oryginalna):

```
 (R1) RGN(prlo, sneg, neg, e, poz, kl)
      = SGN (prlo, sneg, neg, e, poz, kl, typ).

 (S1) SGN(prlo, sneg, neg, e, poz, kl, NALTR)
      = PGN (prlo, sneg, neg, e, kl).

 (S2) SGN(prlo, sneg, neg, e, poz, kl, NALTR)
      = JGN (prlo, sneg, neg, e, poz, kl, nr)
        -ALT(nr, 1.2.6).

(J11) JGN(prlo, sneg, neg, e, poz, kl, 2)
      = PGN (prlo, sneg, neg, e, kl).
```

co niezależnie od znaczenia parametrów[22] oznacza, że każda analiza przebiega dwoma alternatywnymi ścieżkami:

[22]Warto tylko wyjaśnić, że zapis `-ALT(nr, 1.2.6)` wymusza, by wartość parametru `nr` była równa 1, 2 lub 6.

$$rgn(r1) \rightarrow sgn(s2) \rightarrow jgn(j11) \rightarrow pgn$$

$$rgn(r1) \rightarrow sgn(s1) \rightarrow pgn$$

Niedogodność ta jest wynikiem przejętej z opisu hierarchii zdaniowej obsługi konstrukcji dwuczłonowej, zapewnianej przez regułę `j11`. Świdziński pisze[23]:

> Reguła (J11) pozwala opisać grupy szeregowe o dokładnie dwóch członach nominalnych. Grupy takie nie zostały zdefiniowane przez żadną z reguł opisujących szeregową grupę nominalną nieprzecinkową, gdyż szeregowa grupa nieprzecinkowa zawiera co najmniej trzy człony nominalne. Szeregowa grupa przecinkowa natomiast jest, jak pamiętamy, ciągiem składającym się z jednorodnej przecinkowej grupy nominalnej, spójnika szeregowego końcowego oraz pojedynczej grupy nominalnej. Aby zatem możliwa była interpretacja grup szeregowych dwuczłonowych, opartych — paradoksalnie! — na spójniku nieprzecinkowym, konieczne jest uznanie ich za realizacje szeregowej grupy przecinkowej, której pierwszy składnik nominalny, jednorodna grupa nominalna przecinkowa, ma postać pojedynczej grupy nominalnej. To ostatnie ustanawia właśnie reguła (J11).

Jednym z prostszych sposobów rozwiązania problemu jest eliminacja nadmiarowej reguły `j11` i zapewnienie obsługi dwuczłonowych grup nominalnych w regule nadrzędnej. Nie przytaczam jednak tego rozwiązania, gdyż problem został ostatecznie wyeliminowany poprzez przebudowę definicji w ramach upraszczania hierarchii jednostek.

10.2.2 Grupa przymiotnikowa, przysłówkowa i przyimkowa

W sposób zbliżony do koncepcji grupy nominalnej zdefiniowałem również grupy złożone z jednostek równoważnych innym częściom mowy, umożliwiające odpowiednio analizę przykładów w rodzaju:

(94) *Jestem kobietą czułą i łagodną.* *[1367]*

(95) *Tu i ówdzie legły, wiły się ciała.* *[5617]*

(96) *Guanajuato powstało ze srebra i dla srebra.* *[3595]*

Pojęcie grupy rozszerza definicję frazy danego typu. Zachowuję ograniczenia łączliwości związane z różnicami dystrybucji konstrukcji równorzędnych, szeregowych i jednorodnych; w przeciwieństwie do definicji grupy nominalnej w grupie przymiotnikowej zakładam pełne uzgodnienie kategorii przypadka/rodzaju-liczby/stopnia, a w przysłówkowej — stopnia. W przypadku grupy przyimkowej w uproszczeniu zakładam, że reprezentuje ją pierwszy składnik; w szczególności pobierany jest z niego „przyimek grupy", co może mieć wpływ na spełnienie wymagania czasownika.

Oto przyjęty zestaw reguł, na przykładzie grupy przyimkowej:

[23] [Szpakowicz i Świdziński, 1990], s. 35.

```
fpm(Pm, P, Neg, I, Z, Kl, _, rgpm)
--> s(gpm1),
    spojr(l, Nr),
    fpm(Pm, P, _, I, Z, Kl, _, Sgpm1),
    { rowne(Sgpm1, [sgpm, jgpm, fpm]) },
    przec,
    spojr(p, Nr),
    fpm(_, _, _, _, _, _, _, Sgpm2),
    { rowne(Nr, [1, 2]),
      rowne(Sgpm2, [sgpm, jgpm, fpm]) }.

fpm(Pm, P, Neg, I, Z, Kl, _, sgpm)
--> s(gpm2),
    fpm(Pm, P, Neg, I, Z, Kl, _, fpm),
    fpm(_, _, _, _, _, _, Nr, Jgpm),
    { rowne(Nr, [1, 6]),
      rowne(Jgpm, [jgpm, fpm]) }.

fpm(Pm, P, Neg, I, Z, Kl, _, sgpm)
--> s(gpm3),
    fpm(Pm, P, Neg, I, Z, Kl, 2, Jgpm),
    { rowne(Jgpm, [jgpm, fpm]) },
    spojszk(Nr),
    fpm(_, _, _, _, _, _, _, fpm),
    { rowne(Nr, [1, 6, 7]) }.

fpm(Pm, P, Neg, I, Z, Kl, Nr, jgpm)
--> s(gpm4),
    spojsz(Nr),
    fpm(Pm, P, Neg, I, Z, Kl, _, fpm),
    przec,
    spojszk(Nr),
    fpm(_, _, _, _, _, _, _, fpm),
    { rowne(Nr, [1, 6]) }.

fpm(Pm, P, Neg, I, Z, Kl, Nr, jgpm)
--> s(gpm5),
    spojsz(Nr),
    fpm(Pm, P, Neg, I, Z, Kl, _, fpm),
    przec,
    fpm(_, _, _, _, _, _, Nr, Jgpm),
    { rowne(Nr, [1, 6]),
      rowne(Jgpm, [jgpm, fpm]) }.

fpm(Pm, P, Neg, I, Z, Kl, 2, jgpm)
--> s(gpm6),
    fpm(Pm, P, Neg, I, Z, Kl, _, fpm1),
```

```
    spojsz(2),
    fpm(_, _, _, _, _, _,  _, fpm2).

fpm(Pm, P, Neg, I, Z, Kl, 2, jgpm)
--> s(gpm7),
    fpm(Pm, P, Neg, I, Z, Kl, _, fpm),
    spojsz(2),
    fpm(_, _, _, _, _, _, _, 2, Jgpm),
    { rowne(Jgpm, [jgpm, fpm]) }.
```

10.3 Grupy jednostek równorzędnych

W rozdziale 9.2.2 (s. 96) przedstawiłem użyty przeze mnie sposób dodatkowego przetwarzania pewnej klasy jednostek wielowyrazowych po zakończeniu procesu analizy morfologicznej (w rozumieniu Wolińskiego) w celu korekty ich opisów — kiedy ciąg jednostek okazuje się ukonkretnieniem wyrażenia o stałej łączliwości. W przypadku jednostek nieodmiennych rozwiązanie to wydaje się wystarczające, jest jednak za mało ogólne do użycia w przypadkach jednostek wchodzących w oczywiste związki składniowe.

Poniżej przedstawiam próbę rozszerzenia gramatyki o dwie klasy jednostek o podobnym charakterze, która okazała się wystarczająca dla wypowiedników korpusowych.

10.3.1 Grupy apozycyjne

Szczególnym przypadkiem nie akceptowanych przez Świgrę konstrukcji wielowyrazowych są *grupy apozycyjne*. Za [Kallas, 1980][24]:

> Grupa apozycyjna to bezspójnikowe połączenie dwóch koreferencyjnych rzeczowników, z których drugi stoi w tym samym przypadku, co i pierwszy, bądź stoi w mianowniku, bądź też nie jest rządzony pod względem przypadka (nie ma natomiast reguł gramatycznych na wartość liczbową i, tym bardziej, rodzajową tychże rzeczowników). (...) Członem grupy apozycyjnej są też rozwinięte ekwiwalenty dystrybucyjne rzeczownika.

Oto przykłady tego rodzaju związków ilustrujące różne rodzaje braku uzgodnienia[25]: *artysta filozof*, *dusiciel widmo*, *o mieście Toruń*, *panowie szlachta*.

Pewien podzbiór konstrukcji tego rodzaju (ograniczony do pojedynczych rzeczowników, jak w konstrukcji *Pan Tadeusz*) Świdziński nazywa nieformalnie *zawiadomieniami*; w korpusie jako człony grupy apozycyjnej występują oczywiście pełne struktury frazowe w rodzaju *Przewodniczący Rady Ministrów ZSRR Chruszczow*. Ich częstość w języku naturalnym nie jest duża, stanowią one jednak zjawisko nie

[24] Patrz s. 12–13.

[25] Również na podstawie [Kallas, 1980].

rzadsze od wielu już reprezentowanych w gramatyce osobnymi regułami przepisywania, stąd decyzja o dołączeniu do GFJP osobnej reguły umożliwiającej akceptację wypowiedzeń zawierających podobne konstrukcje.

Włączona do GFJP definicja grupy jest następująca:

```
fno(P, R/L, O, Neg, I, Z, Kl)
--> s(no48),
    fno(P, R/L, O, Neg, I, Z, Kl),
    fno(P2, R2/L, O, Neg, I, Z, Kl),
    { P2 == mian; P2 = P }.
```

W zastosowanym opisie rezygnuję z dopuszczenia niezgodności liczby członów składowych; rozwiązanie to ogranicza nieco liczbę możliwych interpretacji. Oczywiście stopień złożoności analizowanych w pracy Kallas konstrukcji ogólnych jest również znacznie wyższy i obejmuje np. konstrukcje w rodzaju zapisanej w następującym wypowiedniku:

(97) *Puszczając na płynny metal silny strumień powietrza spala się metale: arsen, antymon, cynę.* *[3781]*

wykluczonym z przetwarzanego zbioru ze względów interpunkcyjnych.

10.3.2 Złożona konstrukcja przymiotnikowa

Aby być w stanie analizować zdania zawierające złożone konstrukcje przymiotnikowe w rodzaju *pierwszy polski (balon)* dodajemy regułę:

```
fpt(P, Rl, St, Neg, I, z(_,[np]), przym)
--> s(pt30a),
    fpt(P, Rl, St, Neg, I, z(_,[np]), Kl),
    formaprzym(P, Rl, St),
    { rowne(Kl, ['przym', 'tk']
```

Rozszerzenie kategorii klasy ma na celu dopuszczenie konstrukcji w rodzaju *taka szczęśliwa*:

(98) *Jestem taka szczęśliwa.* *[2952]*

10.4 Negacja a wymaganie bezokolicznika

Problem z analizą zdań typu

(99) *Nie mogę spać.* *[2000]*

(100) *Czy nie powinieneś się poradzić lekarza?* *[2887]*

(101) *Dlaczego nie pozwala mi spokojnie odpocząć?* *[4703]*

wymaga analizy sposobu uzgodnienia negacji między frazą wymaganą a finitywną. Świdziński pisze[26]:

> Spośród (...) trzynastu parametrów zdania elementarnego pięć uzgodnionych jest ze wszystkimi frazami składnikowymi, mianowicie aspekt, czas, rodzaj-liczba, osoba oraz negacja.

Założenie to jest spełnione dla wszystkich rodzajów wymagań poza wymaganiem bezokolicznika i frazy zdaniowej, przy czym dla frazy zdaniowej już w regułach GFJP zapisano brak uzgodnienia pionowego między jednostkami `fzd` a `fw1` (reguła `wy19`). Proste rozwiązanie polegające na rezygnacji z uzgodnienia negacji w odnośnej regule `wy6` nie jest poprawne, co wykazuje szerszy artykuł [Przepiórkowski i Świdziński, 1997] dyskutujący wpływ negacji na konstrukcje zdaniowe; oto przykłady błędnych zdań uzasadniające tę tezę[27]:

(102) **Piotrek nie chciał widzieć Marię.*

(103) **Piotrek nie chciał widzieć nic.*

Analiza problemu wykazuje dominację negacji frazy werbalnej najwyższego poziomu[28]:

> If we consider a verbal phrase with another (embedded) verbal phrase, we will see that the highest negativity affects everything, the value `NOT` being transported down the hierarchy. If the highest value of negation if `YES`, then nothing can be predicted concerning lower values: they depend on lower verbals.

Opisane rozwiązanie polega na uzupełnieniu frazy werbalnej o warunek „obliczenia negacji", spełniony albo gdy zarówno konstrukcja werbalna z negacją, jak i fraza wymagana mają wartość zanegowaną (*„nie spotkać Marii"*) — wówczas wartość parametru negacji frazy wynikowej jest dowolna (*„mogę nie spotkać Marii"*, *„nie mogę nie spotkać Marii"*, albo gdy konstrukcja werbalna nie jest zanegowana (*„spotkać Marię"*, *„spotkać Marii"*) — i w tym wypadku wynikowa wartość negacji jest równa wartości parametru negacji frazy wymaganej (*„mogę spotkać Marię"*, *„nie mogę spotkać Marii"*).

Oto przykład reguły przed dodaniem warunku:

[26]Patrz [Świdziński, 1992a], s. 164. Co ciekawe, w przypadku opisu realizacji bezokolicznikowej frazy wymaganej właściwej w dalszej części pracy (s. 196) Świdziński ponownie potwierdza konieczność pionowego uzgodnienia negacji!

[27]Patrz [Przepiórkowski i Świdziński, 1997], s. 12.

[28]Tamże, s. 13.

```
fwe(Wf, A, C, T, Rl, O, wym(_,ResztaWym), K, Neg, I, z(SwZ,NZ))
--> s(we2/4/6),
    kweneg(Wf, A, C, T, Rl, O, Wym, K, Neg, I, z(SwZ,Z)),
    { rowne(Wf, [bok,psu,psw]),
      zrowne(Z, ['p','px','pxx'], NZ) },
    wymagane(Wym, ResztaWym,
     [W1/fw(W1, K, A, C, Rl, O, Neg, ni, z(SwZ1,Z1)),
      W2/fw(W2, K, A, C, Rl, O, Neg, ni, z(SwZ2,Z2)),
      W3/fw(W3, K, A, C, Rl, O, Neg, ni, z(SwZ3,Z3))] ),
    { resztawym(ResztaWym),
      zrowne(Z1, ['p','np'], SwZ1),
      zrowne(Z2, ['p','np'], SwZ2),
      zrowne(Z3, ['p','np'], SwZ3) }.
```

i po modyfikacji:

```
fwe(Wf, A, C, T, Rl, O, wym(_,ResztaWym), K, Neg, I, z(SwZ,NZ))
--> s(we2/4/6),
    kweneg(Wf, A, C, T, Rl, O, Wym, K, Neg1, I, z(SwZ,Z)),
    { rowne(Wf, [bok,psu,psw]),
      zrowne(Z, ['p','px','pxx'], NZ) },
    wymagane(Wym, ResztaWym,
     [W1/fw(W1, K, A, C, Rl, O, Neg2, ni, z(SwZ1,Z1)),
      W2/fw(W2, K, A, C, Rl, O, Neg2, ni, z(SwZ2,Z2)),
      W3/fw(W3, K, A, C, Rl, O, Neg2, ni, z(SwZ3,Z3))] ),
    { resztawym(ResztaWym),
      zrowne(Z1, ['p','np'], SwZ1),
      zrowne(Z2, ['p','np'], SwZ2),
      zrowne(Z3, ['p','np'], SwZ3),
      (Neg1 == nie(_), Neg2 == nie(_)) -> true;
       Neg1 == tak -> Neg = Neg2 }.
```

Zapisanie reguły umożliwia m. in. poprawną interpretację wypowiednika

(104) *<Nie zdążył...> [Serab] <...dać odpowiedzi>.* *[5447]*

który przytaczam dla ilustracji innego ciekawego faktu z dziedziny opisu wypowiedników: wymagana fraza bezokolicznikowa została tu zinterpretowana jako fragment frazy finitywnej, co zdaje się wynikać właśnie z niejasności interpretacyjnej spowodowanej zagnieżdżeniem wymagań: poprawna interpretacja wymagałaby bowiem rozbicia wypowiednika na dwa rekordy, a to z kolei nie mogłoby zostać dokonane bez zmiany postaci tekstowej wypowiednika podrzędnego:

(105) *<Nie zdążył> [Serab] (dać odpowiedzi).* *[5447]*

(106) *<Dać> (odpowiedź).* *[5447a]*

(107) *<Nie dać> (odpowiedzi).* *[5447b]*

Wypowiednik podrzędny nie został w tym przypadku w ogóle zarejestrowany w korpusie.

Rozdział 11

Usprawnienie gramatyki i mechanizmu analizy

Kolejnym etapem weryfikacyjnym jest możliwość sprawdzenia, jakiego rodzaju usprawnień trzeba lub można dokonać w gramatyce i zasobach Świgry, by poprawić wynik akceptowalności składniowej zdań korpusowych.

11.1 Uproszczenie hierarchii jednostek

Praca Wolińskiego zawiera następującą hipotezę o równoważności dystrybucyjnej jednostek zdaniowych w GFJP, którą zamierzam przedyskutować[1]:

> Wymienione jednostki [reprezentujące zdania równorzędne, szeregowe, jednorodne, proste i elementarne] można swobodnie przepisywać jedna na drugą z zachowaniem wymienionych wartości parametrów. (...) To zaś oznacza, że wszystkie te jednostki są sobie dystrybucyjnie równoważne. Trudno twierdzić, że w GFJP zdanie równorzędne składa się z dwóch zdań szeregowych połączonych spójnikiem, skoro każde z tych zdań szeregowych ze względu na swoją budowę może być zdaniem równorzędnym lub elementarnym. Tak więc można by w uniformizacji pójść jeszcze dalej i zastąpić wszystkie jednostki występujące w cyklu jedną (na przykład o nazwie *zdanie*):
>
> *zdanie* → *zdanie, [i], zdanie.*
> *zdanie* → *zdanie, [gdy], zdanie.*
> *zdanie* → *['Jan', umarł].*
> *zdanie* → *['Maria', spała].*
> *zdanie* → *['Piotr', czytał].*
>
> Otrzymana w ten sposób gramatyka opisuje ten sam zbiór zdań i przypisuje im drzewa o tych samych kształtach (z dokładnością do długości nierozgałęziających się gałęzi). Jeżeli chcemy oddać jakoś typ budowy poszczególnych realizacji zdania, można wprowadzić parametr

[1] Patrz [Woliński, 2004], s. 76.

> *Typ_budowy*, ze świadomością wszakże, że nie bierze on udziału w uzgodnieniach i spełnia funkcję czysto dekoracyjną.

Przed rozpoczęciem rozważań warto zauważyć, że byłaby to koncepcja odpowiadająca w prostej linii „gramatyce korpusu wypowiedników" o ograniczonej do minimum hierarchii jednostek (pojęcie wypowiednika współrzędnego łączy koncepcje zdania równorzędnego, szeregowego i jednorodnego, z których dwa ostatnie stanowią w oczywisty sposób jednostki pomocnicze). Oznaczenie wypowiedników składowych w treści wypowiednika nadrzędnego za pomocą nawiasów klamrowych jest dokładnym odpowiednikiem wystąpienia wspólnej jednostki zdaniowej w gramatyce przekształconej w sugerowany sposób.

Niestety, stwierdzenie o równoważności dystrybucyjnej jest nieprawdziwe — nawet na podanym przykładzie widać, że do zbioru zdań generowanego przez tak sformułowaną gramatykę należy zdanie *„Jan umarł i Maria spała i Piotr czytał"*, niepoprawne w języku polskim i wykluczone przez Świdzińskiego. Nierównoważność dystrybucyjna konstrukcji zrealizowana jest w GFJP właśnie za pośrednictwem rozbudowanej hierarchii jednostek oraz dodatkowych warunków sterujących przetwarzaniem i ograniczających uzgodnienia.

Sposób powstania tej hipotezy i możliwość jej wykorzystania łatwo jednak wyjaśnić przyglądając się konstrukcji gramatyki wprowadzającej hierarchie jednostek.

11.1.1 Rekurencja w GFJP i jej konsekwencje implementacyjne

Możliwość reprezentacji potencjalnie nieskończonej liczby wyrażeń języka naturalnego jest w GFJP modelowana w specyficzny sposób, który wydaje się wypływać z następującego przekonania Świdzińskiego[2]:

> (...) dla zapewnienia rekurencji dopuścić trzeba realizację jednostki składniowej poziomu najniższego (najprostszej) przez jednostkę poziomu najwyższego.

Wynikiem tego stwierdzenia jest obecność w gramatyce pięciu cykli jednostek nieterminalnych w postaci zapętlonych klauzul zawierających pojedynczy nieterminal w nagłówku i treści. Dla przykładu, reguły

```
zr(Wf, A, C, T, Rl, O, Neg, I, Z)
--> s(r1),
    zsz(Wf, A, C, T, Rl, O, Neg, I, Z).

zsz(Wf, A, C, T, Rl, O, Neg, I, Z)
--> s(s1),
    zj(Wf, A, C, T, Rl, O, Neg, I, Z, Oz),
```

[2]Patrz [Świdziński, 1992a], s. 59.

```
    { rozne(Oz, lub) }.

zj(Wf, A, C, T, Rl, O, Neg, I, Z, przec)
--> s(j1),
    zp(Wf, A, C, T, Rl, O, Neg, I, Z).

zp(Wf, A, C, T, Rl, O, Neg, I, Z)
--> s(p1),
    ze(Wf, A, C, T, Rl, O, Wym, Neg, I, Z, br).

ze(Wf, A, C, T, Rl, O, Wym, Neg, I, z(SwZ,Z), Ow)
--> s(e19),
    zr(Wf, A, C, T, Rl, O, Neg, I, z(SwZ,Z)).
```

są elementami cyklu

$$zr \rightarrow zsz \rightarrow zj \rightarrow zp \rightarrow ze \rightarrow zr$$

(klauzula oznaczona jako `s1` zawiera także przykład wspomnianego warunku ograniczającego, który oznacza tu, że spójnik `lub` nie może być wartością parametru `Oz`).

Oto pozostałe cykle, zidentyfikowane w pracy Wolińskiego[3]:

$$fno \rightarrow knodop \rightarrow knopm \rightarrow knoatr \rightarrow knoink \rightarrow knom \rightarrow fno$$

$$fps \rightarrow kpspm \rightarrow kpsps \rightarrow kpsink \rightarrow kprzysl \rightarrow fps$$

$$fpt \rightarrow kptno \rightarrow kptpm \rightarrow kptps \rightarrow kptink \rightarrow kprzym \rightarrow fpt$$

$$fzd \rightarrow fzdsz \rightarrow fzdj \rightarrow fzdkor \rightarrow fzd$$

Obecność cyklu z pojedynczym nieterminalem po obu stronach wchodzących w jego skład reguł, przejęta z pracy Szpakowicza[4], sprawia poważny problem dla implementacji parsera gramatyki, gdyż prowadzi do zapętlenia obliczeń. Woliński eliminuje cykle w sposób najmniej inwazyjny dla gramatyki, uniemożliwiając wystąpienie zapętlenia poprzez wprowadzenie dla każdej z tworzących cykl klauzul dodatkowego parametru kontrolującego długość cyklu (wartość „kredytu zaufania" jest ustalana dla nowo rozpoznanych, potencjalnie „niebezpiecznych" nieterminali i zmniejszana o jeden przy przejściu przez jałową regułę). Operacja ta zapewnia w większości przypadków obliczenie zgodne z intencją autora gramatyki, blokuje jednak zdania w rodzaju

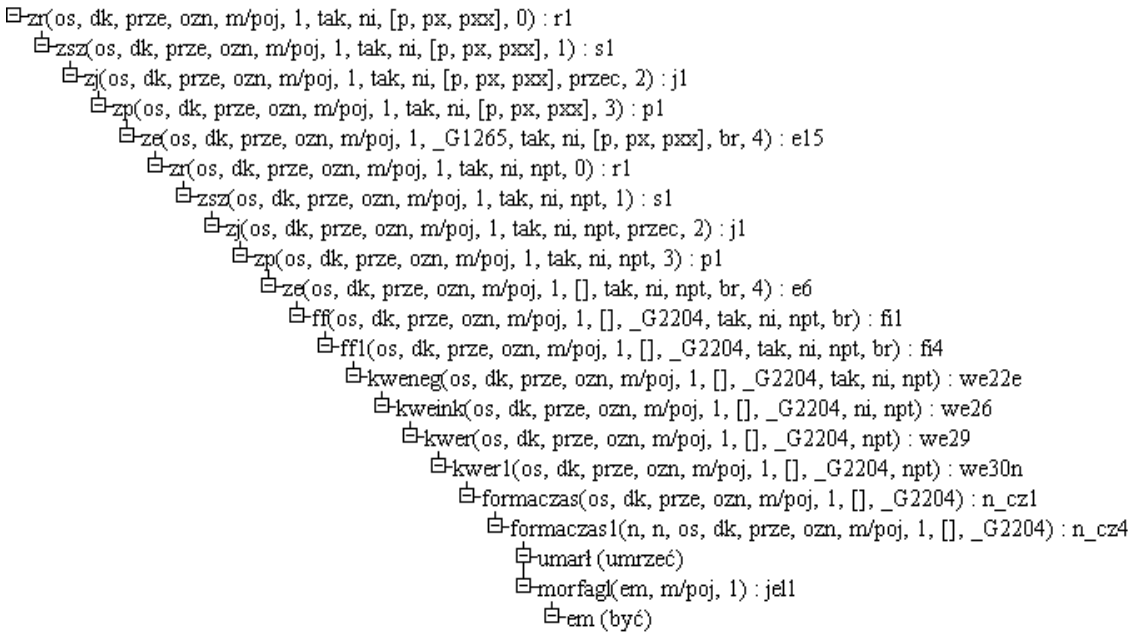
(108) *Ostatnia wreszcie zasada, na której opiera się nasz system szkolny, to zasada nauczania w języku ojczystym.* *[4269]*

[3] [Woliński, 2004], s. 74.

[4] Por. [Szpakowicz, 1986], np. s. 48–54 z regułą `rzskl4` zamykającą cykl $szfrz \rightarrow frz \rightarrow frz1 \rightarrow frz1w \rightarrow krzpodrz \rightarrow krzdop \rightarrow krzatr \rightarrow krz \rightarrow rzeczsk \rightarrow szfrz$. Reguły tego typu miały być jednak usunięte przy implementacji gramatyki.

zawierające atrybuty po obu stronach konstrukcji nominalnej z inkorporacją (*„nasz system szkolny”*). Naprawienie tego problemu w wersji GFJP z mechanizmem Wolińskiego wymaga modyfikacji reguł `no20`–`no37` opisujących różne warianty konstrukcji nominalnej z atrybutem. Najprostsze rozwiązanie polega na zamianie odwołań do jednostki `knoink` na odwołania do jednostki `knoatr`.

Na marginesie warto też wspomnieć, że w związku z „rozwidleniem” cyklu w definicji jednostki zdania elementarnego (znajduje się on w cytowanej regule `e19`, ale i w `e15`) brak blokady w tym drugim przypadku prowadzi do wystąpienia jałowego przebiegu pętli przy próbie analizy napisu jako jednostki niższego rzędu niż wypowiedzenie; oto drzewo analizy napisu *„umarłem”* jako zdania równorzędnego:



Rysunek 11.1: Jałowy przebieg analizy zdania równorzędnego

11.1.2 Uniformizacja jednostek

Pamiętając o nierównoważności jednostek pomysł Wolińskiego można jednak wykorzystać do spłaszczenia drzew rozbioru, co jednocześnie w naturalny sposób — bez wprowadzania sztucznych parametrów — zablokuje niepożądane cykle przy jednoczesnej eliminacji opisanych bezpośrednio powyżej niepożądanych efektów dla zdań z wielokrotnym atrybutem. Jego realizacja polega na wprowadzeniu wspólnego nieterminala dla każdego cyklu i jednoczesnym rozszerzeniu listy jego argumentów o dodatkowy parametr przechowujący nazwę jednostki z hierarchii — w przeciwieństwie do intencji Wolińskiego wchodzący jednak w uzgodnienia i służący do ograniczenia użycia jednostki w nieodpowiednich kontekstach. Koncepcja oczywiście nie jest nowa ani nie wydaje się rozspójniać gramatyki, przeciwnie, jest już stosowana przez Świdzińskiego w postaci klasy np. jednostek frazowych.

Oto skrótowy opis przekształcenia na przykładzie jednostki zdaniowej:

1. Dla każdego cyklu wprowadzamy wspólną jednostkę nieterminalną. Dla jednostek poziomu zdaniowego (`zr` → ... → `ze`) jest nią `zd`.

2. Listę parametrów nowej jednostki stanowi teoriomnogościowa suma argumentów wszystkich klauzul cyklu — oprócz parametru blokującego cykl. W przypadku jednostki zdaniowej zwróćmy uwagę, że mimo tej samej arności funktorów reprezentujących zdanie elementarne i zdanie jednorodne obie jednostki zawierają różny zestaw parametrów (zdanie elementarne posiada dodatkowy parametr ograniczenia zewnętrznego, zdanie jednorodne — oznaczenie spójnika; paradoksalnie oznaczone tym samym skrótem `0z`), zatem w jednostce nowej oprócz wspólnych musimy reprezentować oba argumenty[5].

3. Dodajemy nowy parametr umożliwiający zapis ograniczenia dystrybucji jednostek, którego wartością jest dla lewej strony reguły poprzednia nazwa jednostki, zaś dla reguł składowych po stronie prawej — lista jednostek danego lub niższego rzędu.

   Przykładowo, dla zdania składowego opisywanego oryginalnie jako jednorodne wartość parametru musi być równa `zj`, `zp` lub `ze`, gdyż istnienie cyklu sprawiłoby, że w wyniku przejścia obliczenia przez klauzule tworzące cykl zdanie mogłoby zostać zrealizowane jako jednorodne, proste lub elementarne.

   Wartość dodatkowego parametru w regułach używających jednostki najwyższego rzędu (tu: zdania równorzędnego, reguły dla wypowiedzenia — `w1` i frazy zdaniowej — `zd43` do `zd50`) nie jest ustalona, gdyż wystąpienie mogłoby zostać zrealizowane przez dowolną jednostkę cyklu.

4. Identyfikujemy w cyklu reguły, które nazywam *interweniującymi* — są to klauzule zawierające oprócz zwykłego przepisywania parametrów pewne dodatkowe warunki „rozwidlające", które przed wyłączeniem klauzul cyklu należy rozdystrybuować między jednostki cyklu w sposób zapewniający gramatyce identyczną siłę wyrazu. Naturalnie, podczas tej operacji należy wziąć pod uwagę wartości parametrów w pozostałych regułach na danym poziomie, by nie pominąć pewnych nieoczywistych zależności zapisanych w ten sposób w gramatyce.

   W przypadku zdań i wspomnianego parametru ograniczenia wewnętrznego ważne jest, że jest on używany w jednostce reprezentującej złożone („wielozdaniowe") zdanie proste (jako wyraz wpływu spójnika stanowiącego centrum tego zdania na charakterystykę gramatyczną zdania elementarnego), natomiast dla realizacji pojedynczej zdania prostego (reguła `p1`) jego wartość jest ustalana na `br`; w obu przypadkach parametr ten nie jest przekazywany do jednostki nadrzędnej. Aby zdanie elementarne mogło być bezpośrednim składnikiem zdania jednorodnego, szeregowego, równorzędnego, całego wypowiedzenia lub frazy zdaniowej, należy ustalić wartość parametru ograniczenia wewnętrznego na `br` w miejscu użycia jednostki zdaniowej w definicji każdej z wymienionych jednostek, natomiast nie przekazywać wartości w górę w definicji zdania prostego.

[5] Oczywiście, w przypadku większej liczby parametrów operacja dehierarchizacji straciłaby uzasadnienie, gdyż oznaczałaby sztuczne przeniesienie własności składniowych na poziom, gdzie nie mają one sensu; przekształcane jednostki cechuje jednak duża zbieżność parametrów.

5. Usuwamy reguły tworzące cykl — nie są już potrzebne, gdyż jednostki każdego poziomu mogą być użyte bezpośrednio w jednostkach nadrzędnych.

Oto przykład reguły oryginalnej:

```
zj(Wf, A, C, T, Rl, O, Neg, I, z(SwZ,Z), przec, @ @ @ @0)
--> s(j3),
    zp(Wf, A, C, T, Rl, O, Neg, I, z(pnpp(SwZ),Z2), _),
    { oblpnp(Z2, NZ) },
    spoj(sz, przec, ni),
    zj(Wf1, A1, C1, T1, Rl1, O1, Neg1, ni, z(SwZ,Z1), przec, _),
    { zrowne(Z1, NZ, Z) }.
```

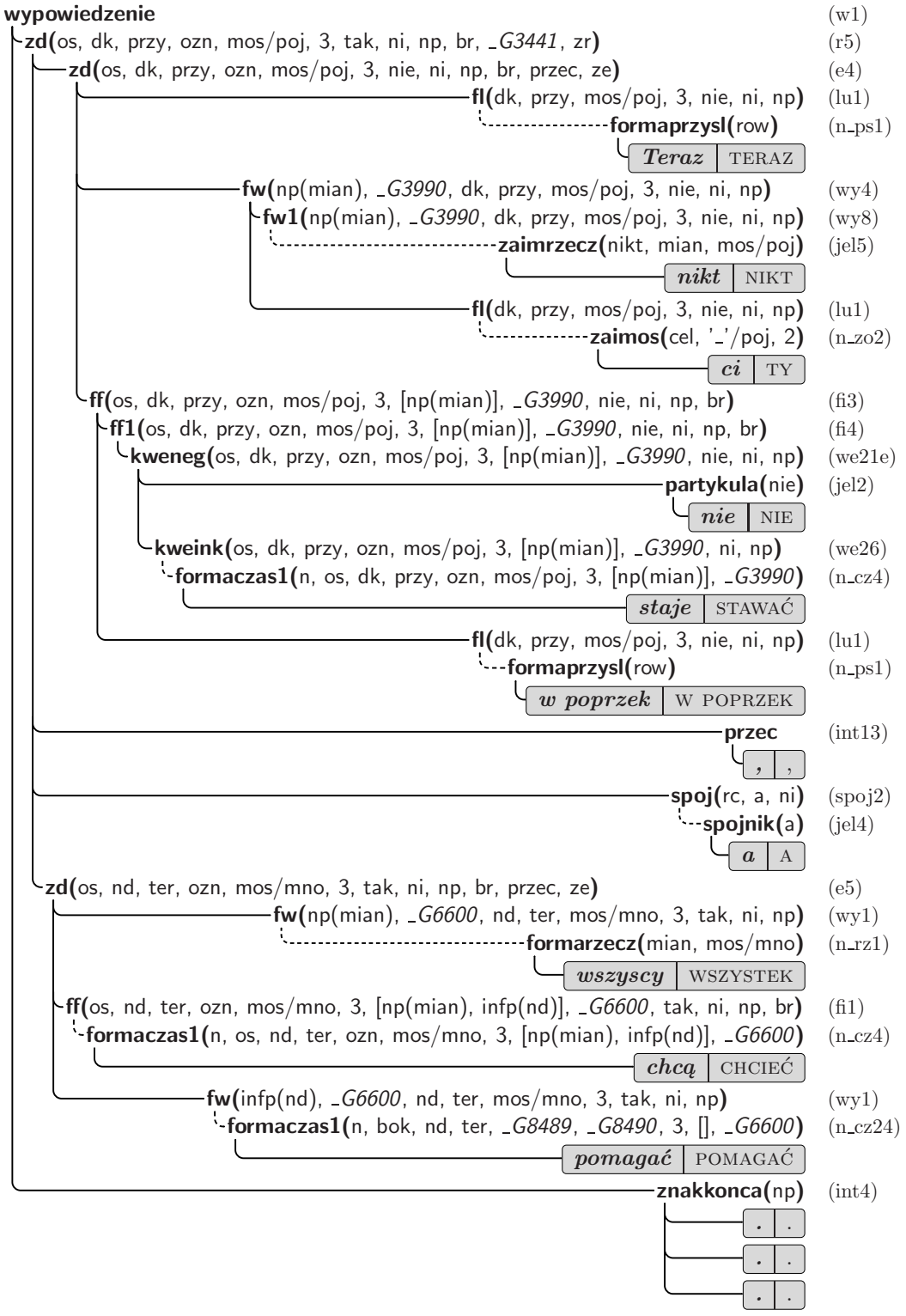
i jej nowego wariantu:

```
zd(Wf, A, C, T, Rl, O, Neg, I, z(SwZ,Z), _, przec, zj)
--> s(j3), zd(Wf, A, C, T, Rl, O, Neg, I, z(pnpp(SwZ),Z2), br, _, Zp),
    { oblpnp(Z2, NZ),
      rowne(Zp, [zp, ze]) },
    spoj(sz, przec, ni),
    zd(Wf1, A1, C1, T1, Rl1, O1, Neg1, ni, z(SwZ,Z1), br, przec, Zj),
    { zrowne(Z1, NZ, Z),
      rowne(Zj, [zj, zp, ze]) }.
```

Wprowadzoną zmianę należy traktować jako alternatywną dla wersji oryginalnej; jej zaletą wydaje się zmniejszenie hierarchii jednostek poprzez eliminację poziomów nie zawierających rozgałęzień i ujednolicenie zapisu jednostek o tym samym znaczeniu (co jest chyba zgodne z intencją Świdzińskiego, numerującego w sposób ciągły reguły odpowiadające jednostkom cyklu), wadą — możliwe zmniejszenie przejrzystości opisu. Jest to jednak kwestia do dyskusji, gdyż wydaje się, że zastosowany zabieg upraszcza rozumienie interweniujących warunków poprzez ich przeniesienie na wyższy poziom hierarchii. Co ciekawe, takie ujęcie tematu sprawia, że stwierdzenie Świdzińskiego o konieczności dopuszczenia *realizacji jednostki składniowej poziomu najniższego przez jednostkę poziomu najwyższego* staje się prawdziwe, gdyż hierarchia staje się pojedynczą jednostką.

Przykładowe drzewo dla wypowiednika `3040` (*„Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...”*) zanalizowanego z wykorzystaniem przekształconej wersji gramatyki (drzewo dla wersji oryginalnej znajduje się w rozdziale 7.2.3, s. 71) przedstawia rys. 11.2.

Poniżej opisuję szczegóły przekształceń reguł pozostałych cykli, w tym sposób przeniesienia dodatkowych warunków. Z braku miejsca cytuję postać reguł wyłącznie dla konstrukcji nominalnej, w pozostałych przypadkach ograniczając się do wskazania newralgicznych punktów opisu.



Rysunek 11.2: Drzewo rozbioru zdania *„Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać...”* wygenerowane dla gramatyki w wersji zmodyfikowanej

Fraza nominalna

Fraza nominalna umieszcza warunki wykraczające poza proste przepisywanie parametrów w klauzulach budujących *konstrukcję nominalną z inkorporacją* (knoink) z jednostek typu *konstrukcja nominalna* (knom):

```
knoink(P, Rl, 0, Neg, I, z(SwZ,NZ), Kl, @ @ @ @ @0)
--> s(no39),
    knom(P, Rl, 0, Neg, z(SwZ,Z), Kl, _),
    { zrowne(Z, ['np','p','pz'], NZ) },
    spoj(Tsp, I, ni),
    { rowne(I, ['bowiem','natomiast','więc','zaś']),
      rowne(Tsp, [pi,ri]) }.

knoink(P, Rl, 0, Neg, ni, z(SwZ,NZ), Kl, X)
--> s(no40),
    knom(P, Rl, 0, Neg, z(SwZ,Z), Kl, @X),
    { zrowne(Z, ['np','p','pz'], NZ) }.

knoink(P, Rl, 0, Neg, I, z(SwZ,NZ), Kl, X)
--> s(no41),
    knom(P, Rl, 0, Neg, z(SwZ,Z), Kl, @X),
    { rozne(I, ['bowiem','natomiast','ni','więc','zaś']),
      zrowne(Z, ['co','jaki','kto','który'], NZ) }.
```

Reguły interweniujące to no40 i no41[6], natomiast wszystkie trzy ograniczają i uzależniają od siebie wartości parametrów inkorporacyjności i zależności. Eliminacja reguł no40 i no41 jako tworzących cykl musi wiązać się z przeniesieniem tego uzależnienia do klauzul niższego rzędu, co zapewni poprawność dystrybucyjną konstrukcji użytych bezpośrednio na wyższym poziomie hierarchii.

Dla reguł dla konstrukcji nominalnej (wybrany zestaw):

```
knom(P, Rl, 3, Neg, z(_,Z), Kl, @ @ @ @ @ 0)
--> s(no42),
    zaimpyt(rzecz, P, Rl, 3, Kl),
    { Z = [p,pz],
      rowne(Kl, ['co','kto'])}.

knom(P, Rl, 3, Neg, z(_,[np]), Kl, @ @ @ @ @ 0)
--> s(no43),
    zaimno(rzecz, P, Rl, 3, Kl),
    { rowne(Kl, ['co','kto'])}.
```

[6]W przekazanej wersji gramatyki parametr inkorporacyjności uzgadniany był błędnie z parametrem rodzaju/liczby; tu podaję jej treść już w wersji poprawionej, zgodnej z postacią z pracy [Świdziński, 1992a]. Identyczne błędne uzgodnienie zawierały reguły no2 i no3.

```
knom(P, Rl, 3, Neg, z(_,[Kl]), wz, @ @ @ @ @ 0)
--> s(no47),
    zaimwzg(rzecz, P, Rl, 3, Kl),
    { rowne(Kl, ['co','kto','który']) }.
```

oznacza to konieczność ustalenia nieinkorporacyjności konstrukcji będącej wynikiem reguł `no42` i `no43` (na mocy przeniesienia wartości parametru inkorporacyjności z reguły `no40` dla wartości parametru zależności równego wartości niepytajnej, pytajnej lub pytajnozależnej, który to warunek spełniają łącznie właśnie wymienione reguły[7]) oraz zapisania warunku na parametr inkorporacyjności dla reguły `no47`, parametr zależności jest już bowiem ustalony poprzez ograniczenie wewnątrz reguły — bardziej szczegółowe niż lista dopuszczalnych wartości z reguły `no41`, nie trzeba go zatem dodatkowo zapisywać.

Oto nowa postać opisywanych reguł, po wprowadzeniu wspólnej jednostki `fno1`, usunięciu parametru blokującego cykl, ujednoliceniu listy argumentów i przeniesieniu warunków:

```
fno1(P, Rl, O, Neg, I, z(SwZ,NZ), Kl, KlGN, knoink)
--> s(no39),
    fno1(P, Rl, O, Neg, z(SwZ,Z), Kl, KlGN, knom),
    { zrowne(Z, ['np','p','pz'], NZ) },
    spoj(Tsp, I, ni),
    { rowne(I, ['bowiem','natomiast','więc','zaś']),
      rowne(Tsp, [pi,ri]) }.

fno1(P, Rl, 3, Neg, ni, z(_,Z), Kl, zaimpyt, knom)
--> s(no42),
    zaimpyt(rzecz, P, Rl, 3, Kl),
    { Z = [p,pz],
      rowne(Kl, ['co','kto'])}.

fno1(P, Rl, 3, Neg, ni, z(_,[np]), Kl, zaimno, knom)
--> s(no43),
    zaimno(rzecz, P, Rl, 3, Kl),
    { rowne(Kl, ['co','kto'])}.

fno1(P, Rl, 3, Neg, I, z(_,[Kl]), wz, _KlGN, knom)
--> s(no47), zaimwzg(rzecz, P, Rl, 3, Kl),
    { rowne( Kl, ['co','kto','który']),
      rozne(I, ['bowiem','natomiast','ni','więc','zaś']). }
```

[7]Mówiąc ściślej, w warunku `zrowne` obliczane jest przecięcie zbiorów podawanych jako dwa pierwsze argumenty — patrz [Woliński, 2004], s. 91–93 — w naszym przypadku wartość z reguły podrzędnej jest podzbiorem wartości z reguły nadrzędnej, a zatem wynik zostanie obliczony poprawnie.

Fraza przymiotnikowa

Definicja frazy przymiotnikowej zawiera warunki interweniujące w klauzulach `pt25` i `pt26`. Oprócz koniecznego do przeniesienia na niższy poziom wartości parametru inkorporacyjności reguła `pt26` służy jedynie ustawieniu określonej wartości parametru klasy (równej `wz`) na podstawie danej wartości parametru zależności (równej zaimkowi `jaki`) i jest nadmiarowa w świetle treści reguły `pt31`, która już tę wartość ustawia. Bliźniacza reguła `pt25` przepisuje jedynie wartości parametrów, obie można zatem bez obaw usunąć.

Fraza przysłówkowa

Definicja frazy przysłówkowej zawiera warunek interweniujący w klauzuli `ps18`; jest on jednak dodatkowym ograniczeniem wartości poszczególnych klauzul uszczegóławiających, całą klauzulę można zatem usunąć bez dodawania warunku do reguł podrzędnych; jedynym zabiegiem jest ich uzupełnienie o parametr inkorporacyjności o wartości `ni`.

Fraza zdaniowa

Fraza zdaniowa zdefiniowana jest w sposób ustalający oznaczenie spójnika szeregowego dla realizacji typowej na wartość odpowiadającą przecinkowi (reguła `zd21`). W przeciwieństwie do pozostałych cykli dodatkowy parametr pojawia się „u góry" struktury, musi zostać zatem odpowiednio obsłużony na wszystkich niższych poziomach hierarchii fraz zdaniowych. W będących składnikami fraz jednorodnych (reguły `zd22` – `zd37`) frazach zdaniowych z korelatem nie posiadających wcześniej parametru oznaczenia przecinka otrzymuje on wartość dowolną; w regułach definiujących frazę z korelatem (`zd38` – `zd40`) — bezpośrednio wartość przecinkową.

11.2 Inne drobne modyfikacje

Na bazie materiału korpusowego i na jego potrzeby wprowadziłem do gramatyki Świdzińskiego kilka innych drobnych zmian opisu o różnym charakterze. Przytaczam je poniżej, bez zachowania żadnego szczególnego porządku.

11.2.1 Zanegowane formy trybu warunkowego

Problem z analizą zdań zawierających nieciągłe formy trybu warunkowego (np. *„Ty byś nie pił."*) wynika z wstępnego założenia o rezygnacji z analizy nieciągłości w GFJP. Ogólne rozwiązanie tej kwestii nie jest trywialne z racji niemal dowolnej budowy członu znajdującego się pomiędzy pseudoimiesłowem a formą aglutynantu warunkowego (połączenia partykuły `by` z aglutynantem), Woliński ogranicza się zatem do konstrukcji, w których aglutynant warunkowy występuje bezpośrednio przed lub bezpośrednio po formie pseudoimiesłowu:

```
formaczas1(S, os, A, _C, war, R/L, O, Wym, _K)
--> s(n_cz12),
    [morf(_, H, praet:Num:Gend:AsAgl)],
    { asagl(AsAgl, As, nagl), aspekt(As, A),
      liczba(Num, L), rodzaj(Gend, R), rekcja(H, S, Wym) },
      condaglt(L, O).

formaczas1(S, os, A, _C, war, R/L, O, Wym, _K)
--> s(n_cz13),
    condaglt(L, O),
    [morf(_, H, praet:Num:Gend:AsAgl)],
    { asagl(AsAgl, As, nagl), aspekt(As, A),
      liczba(Num, L), rodzaj(Gend, R), rekcja(H, S, Wym) }.
```

Na mocy powyższych reguł i sposobu obsługi zaprzeczenia w GFJP możliwa jest analiza zdań zawierających frazy w rodzaju *„jedlibyście”*, *„byście jedli”* i *„nie jedlibyście”*, nie jest jednak możliwa analiza frazy *„byście nie jedli”*. Proste rozszerzenie opisu umożliwiające akceptację zdań zawierających tego rodzaju formy zaprzeczone jest jednak proste i może być zrealizowane w sposób zbliżony do opisywanej osobnymi regułami obsługi możliwości wystąpienia zaimka *się* pomiędzy pseudoimiesłowem a aglutynantem warunkowym.

Wprowadzam zatem następującą regułę, stanowiącą nieznaczną modyfikację reguły `n_cz13`:

```
formaczas1(S, os, A, _C, war, R/L, O, Wym, _K)
--> s(n_cz13a),
    condaglt(L, O),
    partykula(nie),
    [morf(_, H, praet:Num:Gend:AsAgl)],
    { asagl(AsAgl, As, nagl), aspekt(As, A),
      liczba(Num, L), rodzaj(Gend, R), rekcja(H, S, Wym) }.
```

11.2.2 Spójnik *a więc*

Wypowiednik `3868` zawiera zdanie ze spójnikiem *a więc*, nie akceptowanym przez gramatykę Świdzińskiego: *„Będą mogły być krzywoliniowe, a więc będą dyskami.”*. Jest to szczególny rodzaj podrzędnego spójnika typu *więc* typu centralnego (nie inkorporacyjnego).

Aby umożliwić jego analizę, dodałem regułę:

```
spoj1(pc, 'więc')
--> s(spoj28a),
    przec,
    spojnik(a),
    spojnik(F),
    { rowne(F, ['przeto','więc','zatem'])}.
```

11.2.3 Konstrukcje typu *nie najgorzej*

GFJP ogranicza udział parametru negacji w opisie hierarchii frazy przysłówkowej do konstrukcji przysłówkowej realizowanej jako zaimek negatywny, nie jest zatem wprost możliwa analiza zdań w rodzaju *„Poczynasz sobie nie najgorzej, chłopcze.”*. Aby ją umożliwić, dodajemy regułę postaci:

```
fps(naj, Neg, z(_,[np]), Kl)
--> s(ps22a),
    partykula(nie),
    formaprzysl(naj).
```

11.2.4 Formy gerundialne z *się*

Zdania zawierające formy rzeczownika odczasownikowego z *się*, takie jak np. *„Obrazują one zachowanie się organizmu kosmonauty.”* wymagają do poprawnej analizy dodania reguły:

```
formarzecz(P, R/L)
--> s(n_rz4),
    [morf(_, _, ger:Num:Cases:Gend:_:_)],
    [morf('się', 'się', qub)],
    { liczba(Num, L), przypadki(Cases, P), rodzaj(Gend, R) }.
```

11.2.5 *Niech*, *niechaj*, *niechże*

Jedna z realizacji właściwej formy czasownikowej opisuje formy trybu rozkazującego z partykułą *niech* (pierwsza i trzecia osoba liczby pojedynczej oraz trzecia mnogiej):

```
formaczas1(S, os, A, przy, roz, _R/L, O, Wym, _K)
--> s(n_cz17),
    [morf(_, niech, qub)],
    [morf(_, H, fin:Num:Per:As)],
    { (Num=sg, Per\=sec; Num=pl, Per=ter),
      aspekt(As, A), osoba(Per, O),
      liczba(Num, L), rekcja(H, S, Wym) }.
```

Na potrzeby analizy przykładów korpusowych dodałem warianty analizy dla *niechaj* i *niechże* oraz uwzględniłem wystąpienie partykuły negatywnej:

```
formaczas1(S, os, A, przy, roz, _R/L, O, Wym, _K)
--> s(n_cz17),
    [morf(_, Niech, qub)],
    [morf(_, H, fin:Num:Per:As)],
    { (Num=sg, Per\=sec; Num=pl, Per=ter),
```

```
      aspekt(As, A),
      osoba(Per, O), liczba(Num, L), rekcja(H, S, Wym),
      rowne(Niech, [niech, niechaj, niechże]) }.

formaczas1(S, os, A, przy, roz, _R/L, O, Wym, _K)
--> s(n_cz17a),
    [morf(_, Niech, qub)],
    partykula(nie),
    [morf(_, H, fin:Num:Per:As)],
    { (Num=sg, Per\=sec; Num=pl, Per=ter),
      aspekt(As, A),
      osoba(Per, O), liczba(Num, L), rekcja(H, S, Wym),
      rowne(Niech, [niech, niechaj, niechże]) }.
```

11.2.6 Zanegowana fraza przyimkowa

Fraza przyimkowa może bez straty ogólności wystąpić w postaci zanegowanej:

(109) *Nie o to chodzi.* *[2950]*

(110) *Gadali o Broni i nie o Broni.* *[2956]*

Aby obsłużyć tego rodzaju przypadek, w definicji opisującej ją jednostki dodałem regułę:

```
fpm(Pm, P, Neg, I, Z, Kl, 2, fpm)
--> s(pm2),
    partykula(nie),
    przyimek(Pm, P),
    fno(P, Rl, O, Neg, I, Z, Kl, _, _).
```

11.2.7 Konstrukcje przymiotnikowe i przysłówkowe z *coraz*

Konstrukcje z jednostką *coraz* i przymiotnikiem lub przysłówkiem w stopniu wyższym zostały dodane na najniższym poziomie jednostek elementarnych:

```
formaprzym(P, R/L, St)
--> s(n_pt2),
    [morf(coraz, coraz, qub)],
    formaprzym1(P, R/L, wyz).

formaprzysl(wyz)
--> s(n_ps2),
    [morf(coraz, coraz, qub)],
    [morf(_, _, adv:comp)].
```

Z punktu widzenia gramatyki jest to miejsce optymalne, gdyż uniemożliwia wystąpienie konstrukcji w rodzaju *coraz interesujące* czy *coraz do dziś*, pozwala natomiast na analizę zdań w rodzaju:

(111) *Drzewa stawały się coraz wyższe.* *[4900]*

(112) *Jesteśmy coraz dalej od brzegu.* *[2172]*

Fraza *coraz bardziej czerwone* analizowana jest jako konstrukcja przymiotnikowa z frazą przysłówkową.

11.2.8 Konstrukcje przymiotnikowe i przysłówkowe typu *za mało*

Akceptacja zdań w rodzaju

(113) *Dolałam za mało oliwy...* *[2250]*

(114) *Zrobi się za duża luka.* *[2107]*

czyli zawierających konstrukcje złożone z przyimka *za* i przymiotnika lub przysłówka w stopniu równym została zapewniona poprzez dodanie do gramatyki specyficznej konstrukcji obsługującej tego rodzaju złożenia. Rozwiązanie takie zostało również wprowadzone na najniższym poziomie jednostek elementarnych z wykorzystaniem udostępnianej przez Morfeusza kublikowej interpretacji jednostki:

```
formaprzym(P, R/L, St)
--> s(n_pt3),
    [morf(za, za, qub)],
    formaprzym1(P, R/L, row).

formaprzysl(row)
--> s(n_ps3),
    [morf(za,za,qub)],
    [morf(_,_,adv:pos)].
```

Podobnie jak w przypadku konstrukcji z *coraz* powyższy zapis wydaje się optymalny ze względny na dopuszczenie w drugim członie złożenia wyłącznie „przysłówków i przymiotników prostych”, co jest zgodne z jego użyciem w języku.

11.2.9 Imiesłowy przymiotnikowe i przysłówkowe

Imiesłowy przymiotnikowe są w GFJP traktowane jak przymiotniki[8].

Woliński konsekwentnie uzupełnia realizację formy przymiotnikowej o reguły dla imiesłowów czynnych i biernych[9]:

```
formaprzym1(P, R/L, row)
--> s(n_pt5),
    [morf(_, _, ppas:Num:Cases:Gend:_)].

formaprzym1(P, R/L, row)
--> s(n_pt6),
    [morf(_, _, pact:Num:Cases:Gend:_)].
```

zdając sobie sprawę z konieczności ulepszenia fragmentu GFJP dotyczącego imiesłowów przymiotnikowych.

Warto dodać bliźniacze reguły dla form zaprzeczonych:

```
formaprzym1(P, R/L, row)
--> s(n_pt5),
    partykula(nie),
    [morf(_, _, ppas:Num:Cases:Gend:_)].

formaprzym1(P, R/L, row)
--> s(n_pt6),
    partykula(nie),
    [morf(_, _, pact:Num:Cases:Gend:_)].
```

11.2.10 Rozszerzenie zakresu frazy luźnej

Aby móc zanalizować zdania w rodzaju

(115) *Wydostali się w dolinę, na uprawne suche pola.* *[4902]*

(zawierające frazy luźne oddzielone przecinkiem), konieczne jest rozszerzenie definicji frazy luźnej. Świdziński ogranicza listę realizacji tego typu frazy[10] do frazy przyimkowej, frazy nominalnej, frazy przysłówkowej, aglutynantu, frazy werbalnej imiesłowowej, frazy zdaniowej, wskazując możliwości rozszerzenia tego opisu i jednocześnie rezygnując z nich z zastrzeżeniem, że *fraza luźna jest tylko jednym z typów rozważanych w tej pracy konstrukcji, wcale nie najważniejszym.*

Rozszerzenie definicji polega na dodaniu nowej reguły:

[8] Jest to zgodne z rewolucyjnym jak na swoje czasy podejściem zastosowanym już w słowniku frekwencyjnym polszczyzny współczesnej — patrz [Kurcz i in., 1990] i [Nazarczuk, 1997].

[9] Patrz [Woliński, 2004], s. 62.

[10] Za [Świdziński, 1992a], s. 203.

```
fl(A, C, Rl, O, Neg, I, z(SwZ,NZ))
--> s(lu4a),
    fl1(A, C, Rl, O, Neg, I, z(SwZ,Z)),
    { zrozne(Z, ['p','pz'], NZ) },
    przec,
    fl(A, C, Rl, O, Neg, ni, z(SwZ1,Z1)),
    { zrowne(Z1, [np], SwZ1) }.
```

11.2.11 Przymiotniki poprzyimkowe

W specjalny sposób obsługuję nie adresowane przez GFJP konstrukcje złożone z przyimka i tzw. przymiotnika poprzyimkowego[11], czyli wyrażenia w rodzaju *po polsku*, *po prostu*. Dla nieodmiennego członu „przymiotnikowego" Morfeusz udostępnia specjalny znacznik klasy fleksemu, dzięki czemu jest możliwe łatwe rozszerzenie definicji GFJP o specjalny rodzaj przysłówka złożonego:

```
formaprzysl(row)
--> s(n_ps2),
    [morf(_, _, prep:_)],
    [morf(_, _, adjp)].
```

11.2.12 Zaimek zwrotny

Przez GFJP nie są obsługiwane dość licznie występujące w przykładach korpusowych formy zaimka zwrotnego:

(116) *Rozmawiali ze sobą.* *[2400]*

(117) *Scyzoryki starsi panowie mają zawsze przy sobie.* *[5207]*

Morfeusz rezerwuje dla tego zaimka specjalną klasę *siebie*, zawierającą pojedynczy leksem odmienny wyłącznie przez przypadek i ograniczony do form *siebie*, *sobie* i *sobą* (wartość *się* stanowi osobny leksem nieodmienny).

Definicja umożliwiająca obsługę użycia zaimka zwrotnego w analizowanych wypowiedzeniach jest następująca:

```
fno(P, _Rl, _O, _Neg, z(_,[np]), os, zaimzwrot)
--> s(no45a),
    zaimzwrot(P).

zaimzwrot(P)
--> s(n_zz),
    [morf(_, _, siebie:Cases)],
    { przypadki(Cases, P) }.
```

[11] W terminologii wprowadzonej w artykule [Woliński, 2003]; we wcześniejszym artykule [Woliński i Przepiórkowski, 2001] forma tego rodzaju nazywana jest przysłówkiem przyprzyimkowym.

Kwestię pokrewną, obsługiwaną z wykorzystaniem powyższej definicji, stanowi omawiany niżej (w rozdziale 8.1.2, s. 79) przypadek nietypowej realizacji frazy podmiotowej w postaci mianownikowej formy zaimka *się*:

(118) *Spis wyborców układa się oddzielnie dla każdego obwodu głosowania.* *[6492]*

(119) *Aparat wysuwa się do sieni.* *[5904]*

Do poprawnej analizy zdań zawierających tego rodzaju frazę wymaganą wystarczy już tylko rozszerzenie definicji Morfeusza:

```
zamien([i(A, B, się, się, qub) | Tail],
       [i(A, B, się, się, qub),
        i(A, B, się, się, siebie:nom) | NewTail])
:- zamien(Tail, NewTail), !.
```

11.3 Uzupełnienie słownika wymagań czasownikowych

W GFJP typy fraz wymaganych przez frazę finitywną stanowią jej parametr i określane są w *słowniku wymagań czasownikowych* — pliku zawierającym przyporządkowanie typów i parametrów fraz wymaganych danemu leksemowi czasownikowemu. Świgra dopuszcza sześć rodzajów fraz wymaganych, sparametryzowanych charakterystyką morfologiczno-składniową zależną od typu frazy (dla każdego z typów fraz podaję przykład reguły ze słownika wymagań):

- frazę bezokolicznikową o zadanym aspekcie (`infp(A)`),
  `slowczas('nakazywać', n, [[infp(_)]]).`
- frazę przyimkową z danym przyimkiem wymagającym określonego przypadka nominalnej frazy składnikowej (`prepnp(F, P)`),
  `slowczas('wracać', n, [[prepnp('do',dop)]]).`
- frazę nominalną w podanym przypadku (`np(P)`),
  `slowczas('wypominać', n, [[np(mian), np(bier), np(cel)]]).`
- frazę przymiotnikową w podanym przypadku (`adjp(P)`),
  `slowczas('udawać', n, [[np(mian), adjp(bier)]]).`
- frazę przysłówkową (`advp`),
  `slowczas('prowadzić', s, [[np(mian), advp]]).`
- frazę zdaniową danego typu (`sentp(T)`).
  `slowczas('przeczytać', n, [[np(mian), sentp('że')]]).`

Hasła słownika stanowią kolejne klauzule predykatu `slowczas` zawierającego 1) identyfikator leksemu złożony z formy bezokolicznika, 2) określenie, czy opisywany jest leksem z *się* czy bez *się* oraz 3) zestaw wymagań dla danego leksemu.

Dla celów analizy składniowej przykładów z korpusu wypowiedników dodałem do słownika 111 nowych wymagań (z czego 73 dotyczyło czasowników nie reprezentowanych wcześniej w słowniku). Użyte dane pochodziły w prostej linii ze słowniczka

stworzonego na etapie równoległych prac nad analizą charakterystyki składniowej fraz wymaganych przez centrum finitywne wypowiedników (patrz rozdział 4.2, s. 42); z opisanych bezpośrednio poniżej względów (arbitralna interpretacja zakresu wymagania) pełne włączenie do słownika wymagań opisów wyekstrahowanych z danych korpusu wypowiedników nie wydało się jednak pożyteczne.

11.4 Wykluczenie fraz luźnych dla analiz z frazą wymaganą

Jednym z najciekawszych aspektów składniowych jest z naszego punktu widzenia nierozstrzygalna chyba kwestia granicy między frazą wymaganą a luźną oraz wymaganiem czasownika a jego realizacją frazeologiczną.

Dla przykładu zobaczmy opis walencyjny czasownika *atakować* w opracowaniach Polańskiego [Polański, 1966] i Grenia [Greń, 2001]. Polański wymienia dla tego czasownika następujące dwa wymagania:

```
atakować - {(NPAcc) + (NPI)}
atakować - {(NPAcc) + (NPMod)}
```

oznaczające tyle co „atakować kogoś czymś" i „atakować kogoś w jakiś sposób". U Grenia znajdujemy natomiast (po ujednoliceniu notacji):

```
atakować - {(NPAcc)}
atakować - {(NPAcc) + (o NPAcc)}
```

czyli „atakować kogoś" i „atakować kogoś o coś". Jeszcze większe problemy może stwarzać frazeologia (wskutek braku wyraźnych granic między naturalnym wymaganiem składniowym a wymaganiem frazeologicznym): „atakować od lewego skrzydła", „atakować z boku"[12]. Problem ten dość wyraźnie zaobserwować można np. w opisie haseł słownika [Polański, 1980–1988], gdzie oprócz kwestii składniowych uwidaczniają się także wpływające na postać związków zjawiska z pogranicza semantyki i stylistyki.

Dla pełnego opisu należałoby zatem albo rozszerzyć pojemność realizującego dany typ wymagania schematu zdaniowego, wprowadzając dużą liczbę wymagań dla danego czasownika, albo rezygnując z reprezentacji dokładnych wymagań frazeologicznych ograniczyć jego definicję do pojedynczych wymagań i jednocześnie przenieść fragment opisu na poziom fraz luźnych. Rozwiązanie drugie nastąpiło właśnie w gramatyce Świdzińskiego, dzięki czemu możliwa jest analiza zdań w rodzaju *Jan pożycza od Marii książki dla Piotra.*[13] Inną pozytywną konsekwencją rezygnacji ze słownikowej reprezentacji wymagań frazeologicznych jest większa elastyczność gramatyki ze względu na ewoluujący język czy wariacje stylu językowego.

[12] Za [Skorupka, 1977], s. 90.

[13] Analiza ta daje jednak w wyniku ponad 200 drzew rozbioru — właśnie ze względu na alternatywę fraz luźnych i wymaganych.

Opisując technikę zastosowaną dla korpusu wypowiedników Świdziński pisze wprost[14]:

> Podejmując decyzję [czy dany składnik jest frazą wymaganą, czy może luźną] kierujemy się wyczuciem i zdrowym rozsądkiem, ale z preferencją dla interpretacji „dopełnieniowej". [...] Decyzję uznania zdania elementarnego za realizację frazeologizmu podejmowano intuicyjnie, bez kwerendy po słownikach polskich. W szczególności decydowano niejednokrotnie uznawać za frazę wymaganą składnik, który — strukturalnie — jest frazą luźną.

Biorąc pod uwagę złożoność problemu zdecydowałem się przeprowadzić eksperyment eliminujący frazy luźne na mocy hipotezy, że nie istnieją sensowne równorzędne analizy wariantowe z frazą luźną i frazą wymaganą — z dwóch drzew rozbioru, z których jedno zawiera łuk dla frazy wymaganej, drugie zaś dla frazy luźnej, możemy odrzucić analizę z frazą luźną jako nadmiarową.

Intuicyjnie wydaje się, że rozszerzeniem hipotezy mogłoby być badanie izomorficzności drzew rozbioru zawierających frazę wymaganą i wykluczającą frazę luźną, co nie jest zadaniem trudnym i zostało zrealizowane poprzez wykorzystanie funkcji obliczającej dla danego drzewa współczynnik skonstruowany tak, by jego wartość dla drzew izomorficznych była równa, zaś dla nieizomorficznych — różna[15]). Jest to jednak założenie fałszywe, gdyż zależy w całkowitym stopniu od przyjętej gramatyki. Przykładowo, w GFJP drzewa zawierające na tej samej pozycji frazę luźną i wymaganą nie są izomorficzne — fraza wymagana znajduje się na tym samym poziomie co finitywna, tworząc wraz z nią zdanie elementarne (reguły `e1`–`e12`), natomiast fraza luźna w definicji frazy finitywnej może zostać dołączona do frazy finitywnej właściwej (o poziom niżej — reguły `fi2` i `fi3`). Sytuację tę ilustruje rys. 11.3 prezentujący drzewa rozbioru dla zdania *„Kocham ją."*.

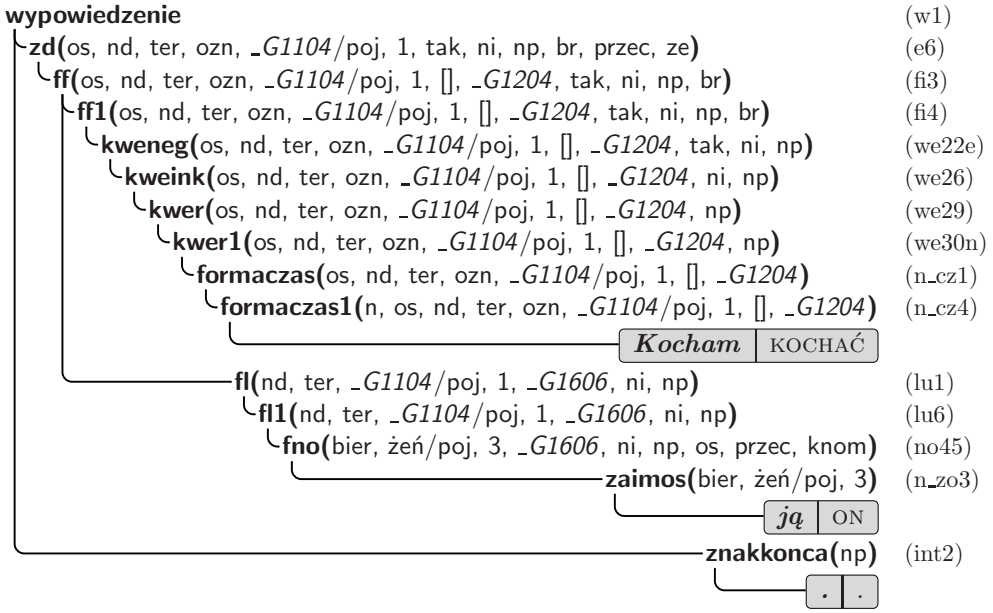
Przed przystąpieniem do eksperymentu warto także zauważyć, że w przypadku ogólnym postawiona hipoteza jest za słaba — np. zdanie *„Będzie jutro."* posiada dwie naturalne, równorzędne analizy oznaczające „dzień jutrzejszy nadejdzie", gdzie *jutro* jest formą rzeczownika, a więc frazą wymaganą (wg szkolnej terminologii: podmiotem) i „coś/ktoś będzie (dopiero) jutro", gdzie *jutro* jest przysłówkiem, a więc frazą luźną (szkolnym „okolicznikiem czasu"). Żadna z analiz nie powinna zatem zostać wykluczona. Poruszona kwestia ilustruje także problem wieloznaczności, dyskutowany szerzej w rozdziale 12.3 (s. 148).

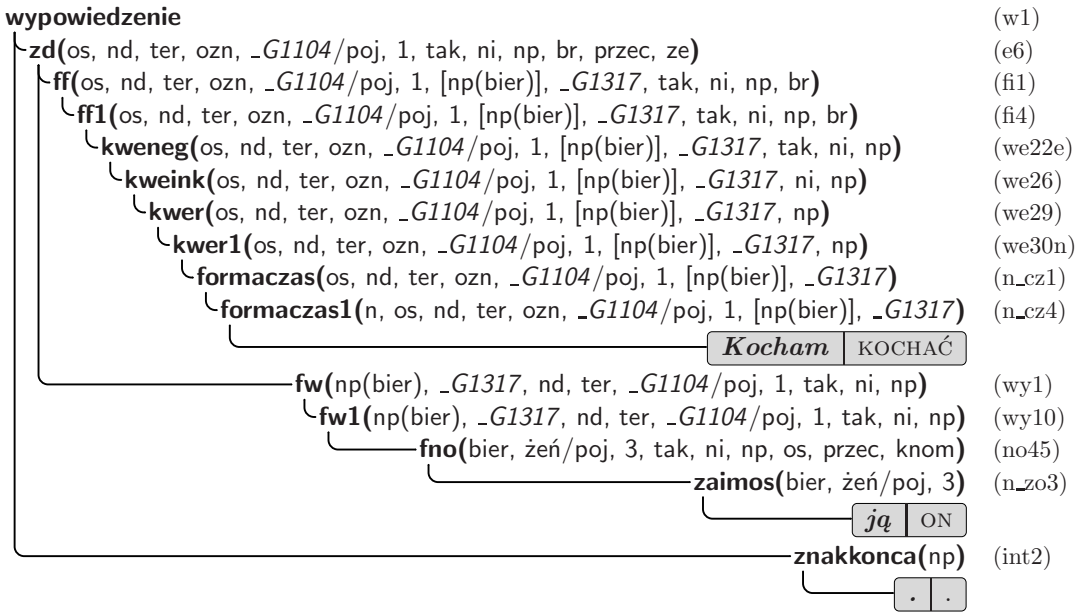
Przykład korpusu wypowiedników dowodzi jednak, że nawet najsłabsze kryterium — eliminacji analiz z frazą luźną, jeśli istnieje analiza z frazą wymaganą — daje w większości przypadków dobre przybliżenie intuicyjnego sposobu analizy zdania, ograniczając w prosty sposób liczbę drzew wynikowych bez odcięcia analiz poprawnych.

---

[14] [Świdziński, 1996], s. 59.

[15] Współczynnik ten jest po prostu napisem reprezentującym strukturę drzewa. Taka jego konstrukcja może wedle potrzeby uwzględniać lub zaniedbywać np. parametry jednostek nieterminalnych.





Rysunek 11.3: Warianty drzewa rozbioru dla zdania *„Kocham ją.”*: z frazą luźną i frazą wymaganą

Eliminacji wyników nadmiarowych dokonałem na etapie końcowym, tj. już po wygenerowaniu drzew rozbioru. Ze względu na zastosowaną metodę analizy (*bottom-up*, czyli wstępującą — od jednostek prostych do złożonych) jest to jedyny właściwy sposób, gdyż samo osiągane przed konstrukcją drzew zapewnienie, że elementy analizowanego napisu posiadają dane interpretacje wymagane nie wystarcza do eliminacji analiz luźnych ze względu na ich potencjalną poprawność. Dla przykładu, słowom *„ją"* i *„bardzo"* w zdaniu *„Kocham ją bardzo."* są na etapie analizy przypisywane równoprawne interpretacje luźne i wymagane. Przysłówek *„bardzo"* to na mocy reguły `ps22` fraza przysłówkowa, która jest z kolei frazą wymaganą właściwą (`wy16`), a ta — frazą wymaganą (`wy1`). To, że jego interpretacja jako frazy wymaganej nie ma sensu w analizowanym przykładzie rozstrzyga się dopiero w momencie tworzenia drzew rozbioru (dla naszego zdania dwóch, z frazą wymaganą *„ją"* i luźną *„bardzo"* oraz z dwoma frazami luźnymi) — przy zastosowaniu metody wykluczenia interpretacji luźnej słowa *„bardzo"* od razu po napotkaniu interpretacji wymaganej drzew analizy nie dałoby się utworzyć w ogóle.

Na mocy spostrzeżenia, że zadanie badania izomorfizmu drzew jest rozłączne z zadaniem eliminacji fraz luźnych, kryterium wykluczającym stało się rozpięcie fraz obu typów nad tym samym fragmentem analizowanego tekstu (zestawem segmentów w rozumieniu analizatora morfologicznego Wolińskiego). Dla przytoczonego wyżej przykładu (patrz rys. 11.3) segmentem tym oznaczonym w różnych (nieizomorficznych!) drzewach rozbioru jest słowo *„ją"*. Wyniki liczbowe uzyskane po zastosowaniu opisanej metody przytaczam w rozdziale 13.3.2 (s. 164).

Część IV

Dyskusja wyników weryfikacji

Rozdział 12

Porównanie wyników weryfikacji ręcznej i automatycznej

Porównanie wyników strukturalizacji „intuicyjnej" — dokonywanej przez edytorów na etapie analizy „z kartką papieru" — z wynikami analizy automatycznej zostało już w dużej mierze wykonane przy okazji szczegółowych dyskusji nad niektórymi własnościami obu definicji i postulowanymi zmianami umożliwiającymi przetwarzanie automatyczne. W bieżącym rozdziale przedstawiam ten problem od nieco innej strony, starając się nie powielać wątków, które omówiłem już wcześniej; w przypadku nawiązywania do już poruszonych tematów podaję odsyłacze do odpowiednich części pracy. Tam, gdzie uważam to za pożyteczne, niektóre kwestie dyskutuję w obszerniejszy sposób.

12.1 Dwie koncepcje weryfikacji

W ramach projektu korpusu wypowiedników zrezygnowano z analizy mechanicznej na rzecz prowadzonej ręcznie, która to decyzja uzasadniona jest następująco[1]:

> Odmiennie niż we wspomnianym wyżej eksperymencie pilotowym[2], który, jak pamiętamy, był kwerendą realizacji różnych schematów zdaniowych, zadanie miało polegać na przeprowadzeniu subtelnej analizy składniowej wszystkich jednostek zdaniowych. Narzędzi analizy dostarczyła GFJP. Analiza automatyczna, wykorzystująca wartości parametrów morfologicznych z korpusu SLF[3], nie wydała się łatwo osiągalna.

i dalej[4]:

[1] [Świdziński, 1996], s. 11.

[2] Patrz rozdział 1.2, s. 1.2.

[3] Słownika frekwencyjnego [Kurcz i in., 1990] (a dokładnie materiału list frekwencyjnych — por. [Kurcz i in., 1974a], [Kurcz i in., 1974b], [Lewicki i in., 1975], [Kurcz i in., 1976] i [Kurcz i in., 1977]) stanowiącego źródło tekstów korpusu wypowiedników; chodzi o dodane ręcznie szczątkowe kody morfologiczne, które mogłyby zostać użyte w procesie analizy składniowej bez potrzeby angażowania analizatora morfologicznego. Por. także przykład z rozdziału 9.2.2, s. 96.

[4] [Świdziński, 1996], s. 22.

> Ponieważ GFJP nie ma komponentu semantycznego, wszystkie możliwe interpretacje danego wypowiedzenia są równoważne, choć istnieją, być może, pary zróżnicowane semantycznie lub takie pary, że jedna z interpretacji, strukturalnie sensowna, jest intuicyjnie nie do przyjęcia. Dlatego właśnie z punktu widzenia celów niniejszego projektu nie wydawało się właściwe budowanie automatycznego analizatora.

Pomijając kwestie techniczne (niedostępności analizatora), drugie stwierdzenie wydaje się mieć znacznie większą wagę w kontekście weryfikacji opisu gramatycznego: w moim przekonaniu musi ona polegać na pełnym sprawdzeniu potencji opisu, na który można będzie dopiero nałożyć interpretacje semantyczne czy wykluczyć warianty „intuicyjnie nie do przyjęcia". Analiza ręczna zastosowana w korpusie wypowiedników jest natomiast zbudowana wokół koncepcji odrzucania pewnych wariantów na bazie przesłanek „pozamerytorycznych", jeśli za takie uważamy jedynie opis gramatyczny. Jeśli natomiast przyjmiemy, że metoda zakładała intuicyjny wybór jednej z dróg analizy z pełną świadomością odrzucenia równoprawnych gramatycznie wariantów, zadanie weryfikacji opisu zostanie ograniczone do sprawdzenia, czy w gramatyce istnieje co najmniej jeden sposób rozbioru danego zdania.

Takie rozumienie pojęcia weryfikacji empirycznej opisu gramatycznego jest nawet węższe niż zaproponowane w artykule opisującym wstępne wyniki projektu korpusu wypowiedników[5]:

> Ponieważ zadaniem językoznawcy — przynajmniej tego, który zajmuje się tekstami — jest tylko zdawanie sprawy z ich budowy, nie zaś odkrywanie jakiejś mitycznej natury języka, problem weryfikacji opisu sprowadza się do sprawdzenia jego adekwatności obserwacyjnej (czy obejmuje wszystkie wyrażenia poprawne i *tylko* takie wyrażenia) i, być może, adekwatności opisowej (czy interpretacje wyrażeń są zgodne z intuicją strukturalną właściwą rodzimemu użytkownikowi).

W moim przekonaniu dla pełnej weryfikacji zgodności modelu z empirią języka sprawdzenie adekwatności opisowej jest warunkiem koniecznym, gdyż jedynie wtedy może nastąpić wykrycie „nadinterpretacji" językowych wprowadzanych przez model. Ostrożność w stwierdzeniu Świdzińskiego bierze się zapewne z przekonania o możliwości istnienia „różnych intuicji językowych", a więc braku „intuicji wzorcowej" mogącej zweryfikować poprawność strukturalizacji. Konsekwencje tego faktu są jednak spore, gdyż podczas ręcznej weryfikacji gramatyki poprzez analizę tekstu Świdziński milcząco przyjmuje jedną, niejako „uprzywilejowaną" strukturalizację, czego wynikiem jest opis jednostki wypowiednikowej, dopuszczający co najwyżej po jednym drzewie rozbioru.

Ponadto weryfikacja modelu nie jest możliwa bez pełnej analizy składniowej, po której nastąpić winno sprawdzenie adekwatności opisowej zmierzające do wyeliminowania z modelu ścieżek wariantowych i niezgodnych z intuicją użytkownika, co nie tylko pozwoli usunąć analizy ewidentnie niepoprawne, ale także może przyczynić się do zwiększenia wskaźnika adekwatności obserwacyjnej (wielość analiz wariantowych

[5] [Świdziński, 1993a], s. 16.

niesie za sobą konsekwencje gramatyczne, które mogą objawić się w postaci akceptacji wyrażeń niepoprawnych, skonstruowanych zgodnie ze schematem ujawnionym przez wariant). Oczywiście, pozostaje jeszcze kwestia wariantowości opisów generowanych przez formalizm modelu (jak w przypadku opozycji fraza luźna/fraza wymagana), która musi zostać rozstrzygnięta przez twórcę gramatyki zgodnie z kryterium zdrowego rozsądku.

Wątek weryfikacji gramatyki poruszam tu jednak przede wszystkim w celu podkreślenia potrzeby i wagi dokonania pełnej analizy składniowej, która w historii opisywanego korpusu nie została wcześniej wykonana; jest on jednak poboczny do samego zadania weryfikacji korpusu, toteż w poprzednim i bieżącym rozdziale rozwijam go jedynie w takim zakresie, w jakim jest mi potrzebny do uzyskania zadowalającego opisu porównawczego.

12.2 Dyskusja metody tworzenia korpusu wypowiedników

Niektóre kwestie omawiające krytycznie metodę tworzenia i anotowania korpusu zostały poruszone w części prezentującej korpus (por. np. rozdziały 1.3.1 — s. 17, 1.3.2 — s. 19), poniżej przedstawiam zagadnienia nie adresowane wcześniej.

12.2.1 Dobór próbek

W opisie metody wyboru danych korpusu wypowiedników[6] silnie akcentowany jest sposób zapewnienia zrównoważenia zestawu opisywanych danych (poprzez wybór dokładnie co dziesiątej próbki z korpusu słownika frekwencyjnego, zastępowanie próbek niedostępnych lub uszkodzonych itp.) Założenie to zostało oczywiście spełnione w jednym z aspektów (udziału próbek źródłowych), chciałbym jednak zwrócić uwagę na jego konsekwencje.

W rozdziale podsumowującym projekt[7] podane zostały następujące liczby określające udział wypowiedzeń i wypowiedników z próbek poszczególnych stylów w powstałym korpusie:

Podkorpus (styl)	**Wypowiedzenia**		**Wypowiedniki**	
	Liczba	**Liczba słów**	**Liczba**	**Liczba słów**
popularnonaukowy	576	10335	1109	16826
wiadomości prasowych	551	9642	916	13737
publicystyczny	552	10208	1196	17760
proza	924	10994	2278	20832
dramat	1911	11119	3318	16918

Powyższy rozkład łącznej długości wypowiedzeń jest zrozumiały w kontekście metodyki tworzenia korpusu słownika frekwencyjnego: długość każdej z próbek wynosiła

[6] Patrz [Świdziński, 1996], s. 16.

[7] [Świdziński, 1996], s. 73 i s. 154 — dwa cytowania tabeli.

ok. 50 słów, co znalazło odzwierciedlenie w średniej łącznej długości wybranych danych (co dziesiąta, czyli 200 próbek z każdego stylu).

Będący wynikiem konsekwentnie zrealizowanej metody wyboru próbek rozkład liczby wypowiedzeń (więc i wypowiedników) sprawia jednak, że już źródłowy zestaw danych wydaje się konfliktować z podstawowym celem projektu — weryfikacją formalizmu gramatycznego stworzonego dla subkodu pisanego. Mimo „tekstowego" źródła danych, redaktorzy słownika włączali do transzy dramatu artystycznego „teksty oparte na dialogu i przeznaczone do realizacji scenicznej"[8] o charakterze mówionym, co wpłynęło na rodzaj ekstrahowanych wypowiedników — w znaczącym procencie oznajmieniowych (31% rekordów dramatu to wypowiedniki tego typu, dla porównania współczynniki dla pozostałych stylów wynoszą od 4,1% do 11,4% — średnio 6,8%).

Wybór 42,6% wypowiedzeń (co daje blisko 38-procentowy udział wypowiedników[9]) ze stylu dramatycznego mimowolnie staje więc w opozycji do zadania weryfikacji, a sam Świdziński przyznaje[10]:

> Nawet pobieżna analiza próbek dramatu pokazała, że aparat GFJP nie chwyta znacznej części wypowiedników; konstatacja ta spowodowała konieczność sięgnięcia po dodatkowe narzędzia badawcze.

Wspomnianymi narzędziami są zapewne metoda opisu wypowiedników oznajmieniowych, być może także sposób opisu niektórych cech ilościowych. Nie zostały one jednak wskazane jawnie.

12.2.2 Specyfika analizy ręcznej

Jak już wspomniałem w rozdziale 8.3 (s. 85), w pracy [Świdziński, 1993a] Świdziński podaje wysoką wartość udziału zdań akceptowanych w badanej próbce zdań korpusowych[11]:

> Przeprowadzona analiza materiału empirycznego pokazała, że stopień adekwatności obserwacyjnej testowanej gramatyki jest bardzo wysoki: z 855 zdań poddanych analizie ponad osiemset stanowi wyrażenia poprawne w sensie GFJP. Jest to wynik zadowalający, nawet jeśli zważyć, że analizę zatrzymywano na poziomie fraz: budowa wewnętrzna frazy nie była ujawniana, chyba że dana fraza była frazą zdaniową lub zawierała frazę zdaniową jako składnik.

Rezultat ten odbiega od współczynnika uzyskanego przeze mnie z oryginalną wersją gramatyki w procesie automatycznej analizy całego zestawu wypowiedników zdaniowych (patrz także rozdział 8.3); cytowana próbka Świdzińskiego miała jednak

[8] [Kurcz i in., 1977], s. 9.

[9] W posiadanej wersji korpusu, zawierającej pełen zestaw próbek dramatu i stylu popularnonaukowego oraz mniej więcej po połowie danych pozostałych stylów współczynnik ten wynosi aż 49,3%.

[10] [Świdziński, 1996], s. 154.

[11] Patrz s. 21.

objętość pięciokrotnie mniejszą niż cały zestaw dostępnych danych, na wynik mógł zatem wpłynąć sposób jej doboru, nigdzie nie opisany.

Oczywiście, zastosowana przeze mnie analiza automatyczna była prowadzona przy skrajnie odmiennych założeniach, mianowicie aż do osiągnięcia jednostek terminalnych, toteż o wykluczeniu zdań w wielu przypadkach mogła zadecydować ich podstruktura frazowa, niesłusznie zaniedbywana w eksperymencie Świdzińskiego. Nie wiadomo także, czy i w jaki sposób badano uzgodnienie parametrów jednostek frazowych (czy np. zatrzymywano analizę na poziomie struktury jednostki zdaniowej, badając zgodność parametrów reprezentanta jednostki frazowej, czy zarzucano uzgadnianie tych parametrów i akceptowano zdanie wyłącznie na podstawie jego zgodności z wzorcem reguły zdaniowej). Podawane przykłady wypowiedzeń nieakceptowanych przez GFJP zawierają się całkowicie w zbiorze konstrukcji programowo zaniedbywanych (patrz rozdział 7.1.4, s. 65), co pozwala wnioskować, że w przyjętym przeze mnie zestawie próbek procentowy udział zdań akceptowanych byłby jeszcze wyższy — przeczą temu wyniki empiryczne. Problemy naprawiane w poprzednim rozdziale nie wynikają też z niedostatków mechanizmu Świgry, co pozwala przypuszczać, że akceptacja jednostek frazowych (jak również eliminacja wieloznaczności, o czym piszę niżej) odbywała się zapewne przy milczącym założeniu pełnej zgodności z modelem GFJP, co nie było działaniem uprawnionym, czego dowodzą problemy napotkane przy przetwarzaniu w pełni automatycznym. Być może było to wynikiem projekcji prawdziwego założenia o pełnej poprawności przykładów korpusowych na założenie o pełności opisu GFJP, jednak w tym wypadku nie można by mówić o pełnej weryfikacji poprawności, wymienianej jako główny cel powstania korpusu.

W świetle powyższego uprawniony wydaje się wniosek, że projekt korpusowy należy traktować jako faktyczną weryfikację nie pełnego opisu GFJP, a jedynie jej fragmentu opisującego zdania złożone, co więcej, w postaci zmodyfikowanej w stosunku do opisu oryginalnego. Wbrew cytowanym na wstępie rozdziału intencjom przyjęta przy tworzeniu korpusu wypowiedników metoda położyła więc nacisk na analizę schematów zdaniowych, a nie rzeczywistego opisu składniowego z GFJP. Wydaje się też to pośrednio potwierdzać waga, z jaką Świdziński podchodzi do wyników ilościowych dotyczących rozkładu schematów.

Wysoki wynik Świdzińskiego można by zweryfikować prowadząc automatyczną symulację analizy z zaniedbywaniem poziomu frazowego, na przykład modyfikując tekst wypowiedników, tak by analiza elementów frazowych zawsze kończyła się sukcesem (wszystkie wystąpienia fraz zostałyby zastąpione reprezentantami odnośnych klas fraz — maksymalnie prostymi konstrukcjami, co do których wiadomo, że są akceptowane przez GFJP). Prowadzenie takiego eksperymentu zarzucam jednak ze względu na wątpliwą korzyść interpretacyjną: ze względu na wspomniany brak definicji użytej próbki uzyskane wyniki byłyby nieporównywalne z wynikiem Świdzińskiego. Co więcej, ze względu na niepełny opis parametrów jednostek frazowych zadanie to mogłoby okazać się niewykonalne (patrz rozdział 4.4, s. 43).

12.3 Wieloznaczność morfologiczna a wieloznaczność syntaktyczna

Najbardziej widoczną różnicą między wynikami analizy ręcznej i automatycznej jest wielość generowanych w tym drugim przypadku drzew rozbioru. Ogólnie rzecz biorąc wieloznaczność analizy może być wynikiem wariantowości *morfologicznej* lub czysto *syntaktycznej*.

Wariantowość morfologiczna jest wynikiem istnienia homonimów i synkretyzmów: analiza frazy *„odkupienie win"* będzie dawać w wyniku dwa drzewa rozbioru niezależnie od gramatyki, jeśli analizator morfologiczny zinterpretuje słowo *win* jako dopełniacz liczby mnogiej rzeczowników *wino* i *wina* (a dodatkowa wieloznaczność może pojawić się przecież także wokół homonimu *odkupienie*); podobnie w przypadku tożsamych postaci form o różnej charakterystyce należących do jednego leksemu[12] — *„Stoję za panią mecenas."* (za kim?/za kogo?).

Syntaktyczna wariantowość strukturalizacji może być natomiast zgodna z intuicją językową lub sztucznie „wymuszona" nadmiarowością gramatyki. Pomijam całkowicie aspekt semantyczny, który może wprowadzać dodatkową wieloznaczność dla pojedynczej strukturalizacji (jak np. dla *„stać w kolejce"*).

W przypadku korpusu wypowiedników analizy są ujednoznacznione, choć (potencjalnie) niekoniecznie jednoznaczne. Oto przykłady zdań korpusowych, na pierwszy rzut oka jednoznacznych, prezentujących niektóre z opisanych wyżej zjawisk:

(120) *Gałęzie drapały jak koty ostrzące pazury, kłuły kruchymi cierniami.* *[4890]*

(121) *Chodzi przede wszystkim o właściwą politykę inwestycyjną w portach.* *[6710]*

(122) *Zaczynam je ścielić.* *[4580]*

(123) *Miałem przez okno przypatrywać się twoim romansom...* *[949]*

Przykłady (120) i (121) ilustrują niejednoznaczność morfologiczną wynikającą z istnienia homonimu (przekraczającą granicę leksemu[13]), przykład (122) — niejednoznaczność morfologiczną wynikającą z synkretyzmu w obrębie leksemu `on`, zaś przykład (123) — niejednoznaczność składniową objawiającą się dwoma różnymi strukturalizacjami wokół dwóch różnych centrów finitywnych (*„miałem"* i *„przypatrywać się"*). We wszystkich przypadkach zapisana w korpusie analiza domyślna została wybrana przez edytora „z wykorzystaniem aparatu semantycznego i probabilistycznego", czyli podświadomie.

Na mocy cytowanego już wyżej stwierdzenia Świdzińskiego alternatywne rozbiory należy traktować jako równoważne, co ma usprawiedliwiać ich liczbę[14]:

[12] Skala wieloznaczności wewnątrzparadygmatycznej jest też pochodną przyjętej reprezentacji informacji morfologicznej — np. paradygmat przymiotnika zawiera nie więcej niż 15 różnych form.

[13] Postacie hasłowe dla formy *„koty"* to *„kot"* i *„kota"*, dla formy *„portach"* — *„port"*, *„porta"* i *„porto"*.

[14] [Świdziński, 1996], s. 22.

> Ponieważ GFJP nie ma komponentu semantycznego, wszystkie możliwe interpretacje danego wypowiedzenia są równoważne, choć istnieją, być może, pary zróżnicowane semantycznie lub takie pary, że jedna z interpretacji, strukturalnie sensowna, jest intuicyjnie nie do przyjęcia.

Jest to zrozumiałe w przypadku przykładów (120) i (122), dla których mimo wieloznaczności morfologicznej otrzymujemy izomorficzne drzewa rozbioru, nie daje się natomiast uzasadnić dla przykładu (123), posiadającego co najmniej dwie nierównoważne strukturalizacje.

Być może właśnie w wyniku powyższego przekonania gramatykę Świdzińskiego cechuje tak duży stopień swobody — dla wielu reguł można podać przykład konstrukcji niepoprawnej akceptowanej przez daną regułę, co sprawia, że wieloznaczności ujawniają się w niespodziewanych momentach. Dla ilustracji, realizacje złożone konstrukcji przymiotnikowej dopuszczają zbyt dużą niezależność parametrów składników, co sprawia, że na bazie przykładu podanego przez autora gramatyki można łatwo znaleźć przykład wypowiedzenia niepoprawnego akceptowanego przez gramatykę[15], np. dla reguł `pt2` i `pt6`:

(124) *On jest taki, jakby nas nie lubił, od dawna.*

(125) **On jest takiego, jakby nas nie lubiła, od dawna.*

(126) *Znam poświęconą Piotrowi książkę.*

(127) **Znam poświęconą Piotrze książkę.*

Dla kwestii weryfikacji korpusu wypowiedników podejście takie oznacza konieczność zaniedbania problemu nadmiarowości analiz, o ile istnieje co najmniej jedna analiza zgodna z rozbiorem dokonanym ręcznie.

Wielokrotnie wyrażane przypuszczenie, że tego rodzaju „nadrzędna" analiza jest pierwszą napotkaną przy parsowaniu wypowiedzenia metodą *top-down* (co mogłoby wynikać z faktu, że gramatyka była pisana z myślą o strategii zstępującej) zostało sprawdzone podczas prac nad przetwarzaniem wyników analiz. W tym celu stworzyłem program do sortowania drzew rozbiorów na podstawie numerów reguł GFJP, tak by analizy korzystające z reguł o niższych numerach były traktowane jako uprzywilejowane w stosunku do analiz o numerach wyższych. Hipoteza o istnieniu analizy „wyróżnionej" nie znalazła jednak potwierdzenia.

Oczywiście, technika porządkowania drzew analizy mogłaby zostać wykorzystana w bardziej zaawansowany sposób i niezależnie od powyższych założeń do ujednoznacznienia lub uprawdopodobnienia analizy: w wariancie skrajnym danej interpretacji morfologicznej czy potencji składniowej mogłoby zostać przypisane prawdopodobieństwo wystąpienia (forma słowa *„koty"* odnosi się raczej do zwierzaka niż rzędnej wysokości punktu, a to z kolei raczej forma mianownika/biernika niż wołacza; w stylu neutralnym zdanie *„Real pokonał AC Milan."* oznacza zazwyczaj zwycięstwo tego pierwszego), które mogłoby dodatkowo szeregować wyniki. Podobna

[15] Problem ten zauważa także Woliński, patrz [Woliński, 2004], s. 113–114.

w założeniach metoda ujednoznaczniania na poziomie morfologicznym wykorzystująca reguły lingwistyczne opisywana jest np. w pracy [Rudolf, 2004][16].

Dane liczbowe z zakresu jednego z aspektów analizy wieloznaczności prezentuję w rozdziale 13.3.1 (s. 162).

12.4 Wypowiedniki a GFJP

12.4.1 Gramatyka Świdzińskiego a schematy zdaniowe

Korpus wypowiedników za jedną z podstawowych metod opisu próbki danych przyjmuje jej klasyfikację względem schematów zdaniowych (patrz rozdział 4.1, s. 39). Jest to koncepcja zbyt prosta do bezpośredniego włączenia do dokładnego opisu języka naturalnego, skupia się bowiem wyłącznie na jednym aspekcie tego opisu, mianowicie otoczeniu frazy finitywnej i z tego powodu może być wyłącznie punktem wyjścia dla pełnej definicji formalnej. Co więcej, schematy rozumiane ściśle ograniczają liczbę fraz wymaganych i nie uwzględniają fraz luźnych, utrudniając klasyfikację nierzadkich w języku naturalnym zdań w rodzaju *Jan pożycza od Marii książki dla Piotra.* (typ schematu: fpn + fno + fpn). Podobnie, nadmiarowe może się wydawać wprowadzenie osobnych schematów dla zdań *Jan wygląda na zmęczonego.* (fpp) i *Jan wygląda nam na zmęczonego.* (fno + fpp), który notabene jako jedyny z całego zestawu nie był reprezentowany w korpusie (patrz rozdział 4.1). W tym sensie schematy stanowią raczej repertuar konstrukcji cząstkowych niż pełny wykaz konstrukcji dopuszczalnych.

Świdziński sam przyznaje[17]:

> Schemat zdaniowy jest — w rozumieniu niniejszego opisu — pewną potencją. Oznacza to, że jego zapis nie zawsze odtwarza rzeczywisty kształt danego zdania elementarnego. Jest on raczej projekcją własności słownikowo-składniowych leksemu, którego forma finitywna stanowi centrum.

Między innymi z powyższych przyczyn opisana poniżej gramatyka formalna nie zawiera bezpośrednio definicji schematów zdaniowych, lecz rozwija i uszczegóławia ich ideę poprzez wprowadzenie szerszego zbioru typów struktur składniowych, odzwierciedlających wewnętrzną budowę składników. Koncepcja schematów znajduje jednak odbicie w wymaganiach czasownikowych definiowanych za pomocą reguł słownikowych[18].

[16] Patrz s. 93–103.

[17] Patrz [Świdziński, 1996], s. 53.

[18] Co ciekawe, nie wszystkie rodzaje schematów mają już w tej gramatyce swoje odpowiedniki — dotyczy to np. konstrukcji z frazą przyimkowo-przymiotnikową w rodzaju *Jan wygląda na zmęczonego.*

12.4.2 Różnice między GFJP a opisem korpusowym i ich konsekwencje

Gramatyka formalna języka polskiego różni się znacznie stopniem szczegółowości od rodzajów struktur reprezentowanych w korpusie wypowiedników. Świdziński zdaje się przekonywać, że uszczegółowienie to nie jest wystarczająco istotne dla idei weryfikacji, sprawia ono jednak, że oba opisy składniowe posługują się praktycznie różnymi formalizmami.

Różnice te w zasadzie nie stoją w sprzeczności z rozumieniem weryfikacji jako próby zastosowania formalizmu leżącego u źródeł GFJP do zawartości korpusu, ale bez bezpośrednich konsekwencji wynikających z zawartego w gramatyce Świdzińskiego opisu (jak zauważa sam autor, GFJP jest jedynie *źródłem narzędzi analitycznych projektu*[19], którymi są zapewne analiza na składniki bezpośrednie czy ograniczenie sposobów zespalania jednostek składniowych do współrzędnego i podrzędnego). Sam formalizm wydaje się być jednak używany w opisie korpusu dość swobodnie; jest także rozszerzany w przypadkach zauważalnych rozbieżności między zawartym w GFJP modelem polszczyzny a empirycznymi faktami językowymi (takimi jak np. częste nieciągłości fraz, w ogóle nie uwzględnione w gramatyce Świdzińskiego). W pewnym sensie opis korpusowy zdaje więc sprawę z większej liczby konstrukcji typowych niż sama GFJP, lecz jest on jedynie punktem wyjścia dla dalszych badań, bez konsekwencji dla ewentualnego dalszego rozszerzania gramatyki.

W opisie korpusowym została przejęta z GFJP typologia spójników współrzędnych (prawy, lewy, centralny, inkorporacyjny, szeregowy), choć sama hierarchia zdań została wyraźnie spłaszczona. Jest to efektem rezygnacji z binarności struktury zdania złożonego poprzez wprowadzenie szeregowości już na poziomie spójnikowego centrum struktury, dla spójników tożsamych leksykalnie często opartego o więcej niż jeden rodzaj spójnika, jak np. w przykładzie[20]:

(128) *Szarpnął obrusem, zerwał z kosy, rzucił pod nogi i biegiem dołączył do towarzyszy.* *[5259]*

którego centrum jest trójka spójników `PRZECINEK-PRZECINEK-I`.

W opisie członów współrzędnych jest to metoda znacznie bardziej naturalna, właśnie ze względu na reprezentację współrzędności szeregu poprzez dokładnie jeden poziom hierarchii, co nie jest możliwe w GFJP. Charakterystyka jednostek współrzędnych (informacja o wypowiedniku współrzędnym początkowym, środkowym i końcowym w ramach parametru *statusu*) wydaje się nadmiarowa, gdyż informacja o kolejności zespolenia składników grupy wypowiedników jest możliwa do odtworzenia na bazie tekstowej odpowiedniości fragmentów treści elementów grupy (co więcej, taka analiza informacji jest konieczna, bowiem wypowiedniki podrzędne nie są opatrzone tego rodzaju klasyfikacją).

Znacząco różna jest też w obu gramatykach reprezentacja zdania elementarnego ze względu na obecność fraz luźnych. W GFJP frazy luźne dołączane są do fraz

[19] Patrz [Świdziński, 1996], s. 19.

[20] [Świdziński, 1996], s. 48.

wymaganych lub zdania elementarnego, wyłącznie na jego początku, w korpusie wypowiedników — zawsze bezpośrednio do zdania elementarnego. Opis korpusowy zakłada pełną pomijalność strukturalną fraz luźnych podczas gdy w GFJP fraza luźna jest składnikiem wchodzącym w uzgodnienia (np. fraza luźna pytajna, pytajnozależna, względna lub aglutynacyjna).

Spore rozbieżności wykazuje również stopień szczegółowości opisu fraz — klasyfikacja w tekście wypowiedników wyróżnia ich cztery rodzaje: finitywną, podmiotową, wymaganą i luźną, podczas gdy GFJP uszczegóławia je (głównie w zakresie realizacji morfologicznej) wyróżniając obok finitywnej frazy przyimkowo-nominalną, nominalną, przymiotnikową, przysłówkową, werbalną i zdaniową. Wątpliwość tę częściowo wyjaśnia parametr przypisujący wypowiednikowi numer schematu zdaniowego (patrz rozdział 4.1, s. 39), posługującego się jawnie bardziej szczegółową klasyfikacją fraz. W porównaniu z repertuarem fraz GFJP różnicę stanowi więc praktycznie wyłącznie fraza przyimkowo-przymiotnikowa, wyłączona z zakresu konstrukcji akceptowanych przez GFJP. Poziom fraz składnikowych zdania elementarnego jest punktem końcowym analizy; składniki terminalne reprezentowane są w opisie korpusowym wyłącznie przez spójniki — centra zespoleń współrzędnych.

Co ciekawe, artykuł zarysowujący projekt korpusu wypowiedników jeszcze przed jego rozpoczęciem [Świdziński, 1993a] przewiduje użycie hierarchii fraz niemal identycznej z opisem GFJP:

WYPOWIEDZENIE
ZDANIE ZŁOŻONE
ZDANIE ELEMENTARNE
FRAZA: FINITYWNA, WYMAGANA, LUŹNA
  FRAZY: WERBALNA, NOMINALNA, PRZYIMKOWA,
  PRZYMIOTNIKOWA, PRZYSŁÓWKOWA, ZDANIOWA
    FRAZA ZDANIOWA:
      ZDANIE (ZŁOŻONE lub ELEMENTARNE)
                lub
      SPÓJNIK + ZDANIE (ZŁOŻONE lub ELEMENTARNE)

Rysunek 12.1: Opis zakładanej pierwotnie hierarchii fraz

Przykład ten pokazuje jednak, jak z formalnego punktu widzenia różnią się oba opisy, uzasadnione wydaje się więc przypuszczenie, że wyniki ręcznej analizy mają z GFJP związek jedynie pośredni, a mianowicie w sensie wykorzystania w obu przypadkach tego samego opisu schematów zdaniowych: w przypadku GFJP jako podstawy tworzenia formalnego opisu polszczyzny, natomiast w projekcie korpusu wypowiedników — jako zestawu reguł wspierających analizę dokonywaną intuicyjnie.

Formalny opis algorytmu analizy wypowiedzenia, jakim jest GFJP, nie mógł być stosowany przez osoby pracujące nad opisem wypowiedników — zapewne jako zbyt trudny do masowego użycia przez człowieka. W korpusie wypowiedników nie ma np. śladu po wynikach uzgodnienia parametrów poszczególnych fraz. Oczywiście mógł to być zabieg celowy, gdyż zadanie anotacji składniowej korpusu jest znacząco inne niż postawiony ogólnie problem analizy składniowej, który ma rozwiązywać GFJP. Możemy założyć (i tak zapewne przyjęto), że tekst wypowiedników stanowi

poprawne wypowiedzenia, zatem uzgodnienia parametrów składniowych są już dokonane na mocy założenia o poprawności składniowej tekstu (pochodzącego ze źródeł pisanych korpusu słownika frekwencyjnego). Jeśli pełne drzewo rozbioru zawierające wartości tych parametrów nie ma być wynikiem projektu, postanowiono zebrać jedynie te najistotniejsze, mianowicie odpowiadające centrum struktury. W tym kontekście schematy zdaniowe stały się wygodnym skrótem umożliwiającym prostą klasyfikację fraz przy ograniczeniu liczby oznaczeń w tekście (choć można by sobie także wyobrazić większą liczbę symboli kategoryzujących frazy, przez co dodatkowe pole z numerem schematu mogłoby zostać wyliczone automatycznie).

Taka interpretacja założeń leżących u początku projektu korpusu wypowiedników sprawia, że oryginalne zadanie weryfikacji należy traktować raczej jako próbę ilustracji pewnych zjawisk gramatycznych na bazie niektórych mechanizmów gramatyki Świdzińskiego niż pełne, formalne sprawdzenie akceptowalności tekstów poszczególnych wypowiedników w ramach GFJP.

12.4.3 Korpusowe składniki frazowe a frazy GFJP

W rozdziale 8.1 (s. 75) zamieściłem dyskusję odpowiedniości między frazami korpusowymi a frazami GFJP oraz zakres realizacji członów wszystkich oznaczonych typów; poniżej krótko podsumowuję statystyczny aspekt tej różnicy. Wszystkie wyniki odnoszą się do zmodyfikowanej wersji gramatyki powstałej w wyniku tej pracy.

Rodzaj składnika	*fraza fin.*	*fraza podm.*	*fraza wym.*	*fraza luźna*	*człon inny*
Liczba wystąpień	4807	2292	4483	2125	1154
Członów różnych	2851	1693	3716	1655	318
Akceptowanych przez GFJP	2671	1290	2428	992	247
Nieakceptowanych przez GFJP	180	403	1288	663	71

12.5 Kwestia wypowiedników niezdaniowych

Na krótki komentarz zasługuje też — moim zdaniem — obecność wypowiedników oznajmieniowych w korpusie; poniżej prezentuję mój pogląd na tę kwestię i opisuję wynik eksperymentu służącego analizie jednego z rodzajów oznajmień.

12.5.1 Oznajmienia w korpusie wypowiedników

Choć Skibicki pisze[21]:

> W trakcie wprowadzania informacji, a więc również w procesie analizy kolejnych wypowiedników, okazało się, że proponowany zestaw reguł syntaktycznych wystarcza do opisu bardziej regularnej części polszczyzny (zdań), a jest niewystarczający dla form innych (m. in. oznajmienia).

[21] [Skibicki, 2000], rozdział 8, podrozdział *Wnioski na temat informacji z bazy*, s. 33.

wydaje się, że obecność w korpusie wypowiedników oznajmieniowych jest celowa — o ile samodzielne oznajmienia nie odegrały żadnej roli w realizacji podstawowego celu badań, a ich opisy zawierają oprócz kategorii dystrybucyjnych prawie wyłącznie wartości cech ilościowych, to składnikami niezdaniowych wypowiedników złożonych były nierzadko wypowiedniki zdaniowe. W tego rodzaju przypadku wypowiedniki składowe były wypisywane i poddawane analizie.

Oprócz zapewnienia kompletności opisu nie można ponadto wykluczyć hipotezy o włączeniu oznajmień do korpusu na potrzeby późniejszego rozszerzania GFJP, gdyż Świdziński dużo miejsca poświęca zajmującej się problemem klasyfikacji jednostek oznajmieniowych monografii Marka Wiśniewskiego [Wiśniewski, 1994]. Z tego właśnie opracowania przejęta została wykorzystana w korpusie typologia oznajmień (nieznacznie rozszerzona o cytaty obce i konstrukcje urwane), a pośrednio także sama koncepcja wypowiednika jako jednostki składniowej o realizacji zdaniowej lub oznajmieniowej.

Dyskutując pracę Wiśniewskiego Świdziński wyróżnia ponadto kilka innych typów oznajmień, jak polskie *question tags* (*„Mają rację, nie?”*, *„Przyjdź, dobrze?”*), samodzielne pytajniki (*„Jak?”*, *„Czyżby?”*) czy oznajmienia puste (*„Przyszedł i...”*), które nie posiadają jednak realizacji w korpusie i nie są dodatkowo oznaczane.

12.5.2 Oznajmienia jako składowe wypowiedników nadrzędnych

Osobny ustęp warto poświęcić składni oznajmień, gdyż wobec tak silnej reprezentacji jednostek nie posiadających wewnętrznej struktury w korpusie weryfikującym opis składniowy (łącznie 1269 rekordów) najciekawszym aspektem ich istnienia jest sposób, w jaki tworzą związki składniowe z innymi wypowiednikami.

Świdziński pisze z jednej strony[22]:

> W niniejszej pracy (...) oznajmieniom nie przypisujemy jednak żadnej struktury.

podczas gdy oznajmienia nieelementarne są strukturyzowane[23]:

> Nie strukturyzuje się dalej oznajmień elementarnych. Jeśli jednak zawierają one jakiś składnik będący wypowiednikiem, ujawnia się jego obecność i własności, w wypadku zaś, gdy jest wypowiednikiem złożonym lub zdaniem elementarnym — także jego strukturę.

i nie chodzi tu wyłącznie o wyodrębnienie wypowiedników podrzędnych, stających się osobnymi rekordami korpusu, ale także o ujawnienie strukturalizacji odpowiedników zdań złożonych i dokonanie ich opisu przy zachowaniu podobnych założeń (wskazane zostaje centrum struktury, zachowana typologia spójników itp.).

[22] [Świdziński, 1996], s. 34.
[23] [Świdziński, 1996], s. 37.

Wstępna analiza oznajmień różnych typów wykazuje, że jedynie składnia równoważników zdań i elips, czyli konstrukcji bez formy finitywnej czasownika powielających schemat zdaniowy części równoważnej jest zbliżona do składni pełnego zdania elementarnego — właśnie poprzez uzupełnienie członu finitywnego (zawartego w poprzedniku lub „wydedukowanego"). Ich analiza składniowa byłaby możliwa po dopełnieniu konstrukcji odpowiednią frazą finitywną. Rozważania te nie są wyłącznie teoretyczne, gdyż analiza taka została przeze mnie przeprowadzona dla jednego z typów oznajmień; opisuję ją bezpośrednio poniżej.

Zbliżonych zabiegów analitycznych można by dokonywać także dla wypowiedzeń dopowiedzeniowych:

(129) [*Mama*] <*się zdenerwowała*>... *[2588]*

(130) *Rzeczywiście...* *[2589]*

czy wypowiedzeniowych członów syntaktycznych (usamodzielnionych fraz — a w ogólnym przypadku raczej grup fraz):

(131) *(Ludwik Bartoch)* <*jestem*>. *[5586]*

(132) *Znad Srebrnego Potoku.* *[5587]*

a także dla niereprezentowanych w korpusie, aczkolwiek wymienianych przez Świdzińskiego[24] usamodzielnionych fraz zdaniowych czy nawet polskich *question tags* czy samodzielnych pytajników, jednak wymagałoby to znacznych ingerencji w strukturę wypowiednika (*„Rzeczywiście."* = *„Zdenerwowała się."*, *„Znad Srebrnego Potoku."* = *„Jestem znad Srebrnego Potoku."*), uzasadnionych raczej dopiero na etapie badania struktury wypowiedzi, a nie składni pojedynczego wypowiednika.

12.5.3 Analiza wypowiedników niezdaniowych

W ramach eksperymentu przeprowadziłem analizę wypowiedników niezdaniowych jednego typu — równoważników zdań — po uzupełnieniu ich treści brakującą frazą finitywną[25]. Zabieg ten stosowany jest także przez Świdzińskiego, który wprowadza poprawki do tekstów korpusu uzupełniając wypowiedniki urwane[26].

Operacja miała na celu sprawdzenie (co prawda, na bardzo ograniczonej próbce danych), w jakim stopniu konstrukcje oznajmieniowe wpisują się w formalizm GFJP i jak mogłaby przebiegać ich analiza przy wykorzystaniu już dostępnych mechanizmów. Zakres weryfikacji wybieram arbitralnie: zgodnie z wyjaśnieniem z poprzedniego rozdziału nie prowadzę analizy tekstów wypowiedników oznaczonych jako usamodzielnione frazy, również możliwej po ich uzupełnieniu o brakujący składnik, *który*

[24] [Świdziński, 1996], s. 33.

[25] Przykład jest to o tyle prosty, że pominięta fraza pochodzi z danego wypowiednika (jest to podstawa rozróżnienia między równoważnikiem a elipsą, która musiałaby zostać dopełniona frazą z poprzednika lub „wydedukowaną").

[26] Patrz [Świdziński, 1996], s. 44–45.

— *intuicyjnie — należy strukturalnie do danego wypowiednika*[27]; analizę wypowiedników eliptycznych, równie prostą do przeprowadzenia, pomijam ze względu na duży rozmiar danych koniecznych do ręcznego przetworzenia (niespełna 800 rekordów).

Konkurencyjnym pomysłem mogłaby być analiza oznajmień jako grup jednostek frazowych, jednak opisana próba wydaje się ciekawsza w kontekście zalążka analizy większych fragmentów wypowiedzi, przekraczających granice znaku kończącego zdanie.

Teksty równoważników zostały uzupełnione składnikiem pochodzącym z wypowiednika zależnego przy zachowaniu zasady minimalnej ingerencji w tekst oryginalny. Do tego celu wystarczała zazwyczaj sama ograniczona do czasownika fraza finitywna, jednak w przypadku czasowników o określonych wymaganiach składniowych zachodziła czasem potrzeba uzupełnienia wypowiednika o pozostałe frazy wymagane, jak np. w przykładzie (134):

(133) *<Nie zależy> (1 mi 1) (2 na pedagogicznych wywodach 2), =lecz= ...* *[4886]*

(134) *..., =lecz= na partyjnej ocenie sytuacji w hotelach.* *[4887]*

dla którego analizujemy tekst:

(135) *Zależy mi na partyjnej ocenie sytuacji w hotelach.* *[4887m]*

lub o partykułę negacji, jak w przykładzie (137):

(136) *...=że= do czterech plemników, które dały początek czworgu jego dzieciom, trafił za każdym razem allel zero, ani razu zaś allel a.* *[4076]*

(137) *...=że= ... ani razu zaś allel a.* *[4079]*

dla którego analizujemy tekst:

(138) *Ani razu nie trafił zaś allel a.* *[4079m]*

Nieliczne przykłady wymagały nieco głębszych modyfikacji stosowanych już wcześniej dla wypowiedników zdaniowych (por. np. rozdział 9.1.5, s. 91), jak w przykładzie (140):

(139) *To nie tylko studentom potrzebna jest porządna szkoła średnia, potrzebna jest w ogóle społeczeństwu i tego, co całemu społeczeństwu, potrzeba również studentowi.* *[5881]*

(140) *..., co całemu społeczeństwu,...* *[5887]*

dla którego analizujemy tekst:

[27] [Świdziński, 1996], s. 118.

(141) *Tego potrzeba całemu społeczeństwu.* *[5887m]*

Na 33 zmodyfikowane w ten sposób równoważniki analiza składniowa trzydziestu zakończyła się pomyślnie.

Pełną listę poddanych analizie równoważników wraz z informacją o przekształceniach, jakim zostały poddane oraz wyniku analizy zamieszczam na płycie CD (patrz rozdział D.6, s. 211).

Rozdział 13

Omówienie wyników liczbowych

W bieżącym rozdziale przedstawiam wyniki finalnej analizy składniowej dokonanej po wprowadzeniu opisanych wyżej modyfikacji oraz prezentuję niektóre dane liczbowe porównujące opis składniowy dokonany ręcznie z opisem dokonanym automatycznie przy użyciu Świgry.

13.1 Złożoność procesu analizy

W celu poprawy złożoności czasowej i pamięciowej algorytmu Woliński dzieli proces analizy na dwa zasadnicze etapy: właściwą analizę składniową, której wynikiem jest upakowany las analiz wynikowych (ang. *packed* lub *shared parse forest*, patrz [Woliński, 2004], s. 37 i dalsze) i proces generowania pełnych drzew rozbioru na podstawie wyników parsera. Uzyskanie informacji o pomyślnym przebiegu procesu jest więc możliwe już po zakończeniu jego pierwszego etapu, podczas gdy pełna liczba wyników jest dostępna dopiero po ich przetworzeniu algorytmem o większej złożoności. Jak zauważa Woliński[1], własności gramatyki Świdzińskiego nie eliminują w ogólnym przypadku wykładniczego wzrostu liczby analiz. W świetle powyższego oraz uzyskanych dla pierwszego etapu procesu analizy danych liczbowych (dla niektórych przykładów czas trwania procesu analizy wynosi wiele godzin) rezygnuję z zamieszczenia na dołączonej płycie pełnego zestawu rozbiorów wynikowych. Ich uzyskanie jest możliwe w opisanym w rozdziale D.1 (s. 206) trybie wykorzystania środowiska analizy składniowej.

13.2 Końcowe wyniki analizy automatycznej z nową wersją gramatyki

Po dokonaniu opisanych w poprzednim rozdziale usprawnień przeprowadziłem ponowną analizę treści wypowiedników, by zmierzyć zakres wprowadzonych optymalizacji. Przedtem jednak przeprowadziłem test opisu na używanych w tym przypadku

[1] Patrz [Woliński, 2004], s. 36.

standardowych przykładach testowych stworzonych na potrzeby projektu Janusza S. Bienia[2].

13.2.1 Analiza przykładów testowych

Weryfikacja gramatyki na przykładach testowych okazuje się niezwykle potrzebna przede wszystkim ze względu na obecność w zestawie przykładów testowych zdań niepoprawnych, z definicji nie reprezentowanych w korpusie wypowiedników. Ma to znaczenie dla zapewnienia adekwatności rozwijanej gramatyki, gdyż proces dostosowania gramatyki do akceptacji zdań korpusu może zachęcać do nadmiernego rozluźniania opisu.

Użyty zestaw testowy jest identyczny z wykorzystanym przez Wolińskiego[3], może zatem służyć do porównania siły wyrazu obu gramatyk — otrzymanej z oryginalną wersją Świgry i używanej do inicjalnej weryfikacji składniowej korpusu wypowiedników (patrz 8.3, s. 85) oraz jej rozszerzenia powstałego w wyniku niniejszej pracy.

Poniższa tabela przedstawia wyniki testów weryfikacyjnych, dla porównania z otrzymanymi przez Wolińskiego w formacie identycznym z użytym w jego pracy. Podaję liczby przykładów w kategoriach uwzględniających z jednej strony klasyfikację poprawności nadaną im przez autora gramatyki, z drugiej — akceptowalność przez analizator składniowy. Podobnie jak Woliński przez wartości przeciętne rozumiem medianę:

	poprawnych		**niepoprawnych**	
	akc.	**nieakc.**	**akc.**	**nieakc.**
liczba przykładów	515		145	
	459 89,13%	56 10,87%	52 35,86%	93 64,14%
przeciętny czas (s)	0,27	0,22	0,30	0,15
przec. liczba kroków	272150	305322	242285	169613

Adekwatność obserwacyjna[4], czyli stosunek liczby poprawnych odpowiedzi do rozmiaru danych testowych wynosi 83,64%.

Obliczeń dokonano na komputerze z procesorem Intel Pentium 4M o częstotliwości taktowania 2.2 GHz wyposażonym w 1 GB pamięci RAM. Łączny czas analizy przykładów testowych wyniósł 492 sekundy (ponad 10 minut[5]).

13.2.2 Analiza tekstów wypowiedników

Po dołączeniu zdań zawierających liczebniki analizie mogło zostać poddanych 4716 przykładów (w tym 4540 różnych; największą frekwencję równą 9 ma wypowiednik *„Tak jest.”*).

[2][Bień, 2000]; o formacie pliku testowego wspominałem już wcześniej w rozdziale 10.2.1 (s. 110); sam plik umieszczony jest na dołączonej do pracy płycie, patrz rozdział D.6, s. 211.

[3]Patrz [Woliński, 2004], s. 107.

[4]Patrz [Bańko, 1990].

[5]U Wolińskiego — 11,5 minuty ze względu na mniejszą moc obliczeniową używanego procesora

Oto końcowe wyniki analizy:

Wypowiedniki	**Liczba**	**Udział %**
akceptowane	3962	84,01 %
nieakceptowane	732	15,52 %
nie dające się zanalizować (analiza trwa dłużej niż 8 godzin)	22	0,47 %

W porównaniu z trzydziestoprocentowym wynikiem bazowym[6] wynik ten wydaje się zadowalający; jego poprawienie, z pewnością możliwe, wymagałoby dużego nakładu pracy, zapewne również ze strony Autora gramatyki.

Zbiór wypowiedników o długim czasie analizy cechuje duża średnia długość (średnio 35 słów) oraz obecność konstrukcji współrzędnych powodujących wspomniany wyżej w rozdziale 13.1 znaczny wzrost złożoności wynikający z dopuszczenia wszystkich „nawiasowań" konstrukcji współrzędnej. Oto kilka przykładów wypowiedników z tego zbioru o długości zbliżonej do obserwowanej średniej:

(142) *Kampanie drugiej wojny światowej w sposób nie budzący żadnych wątpliwości wykazały, a niedawne działania wojenne w Korei, Wietnamie, Egipcie i Algerii potwierdziły, że głównym zadaniem samolotu nie jest już przewożenie bomb, lecz że należy go traktować jako nowy środek transportu, który może wpłynąć na zmianę formy prowadzenia wojny, co jest szczególnie ważne w erze broni jądrowych.* *[3392]*

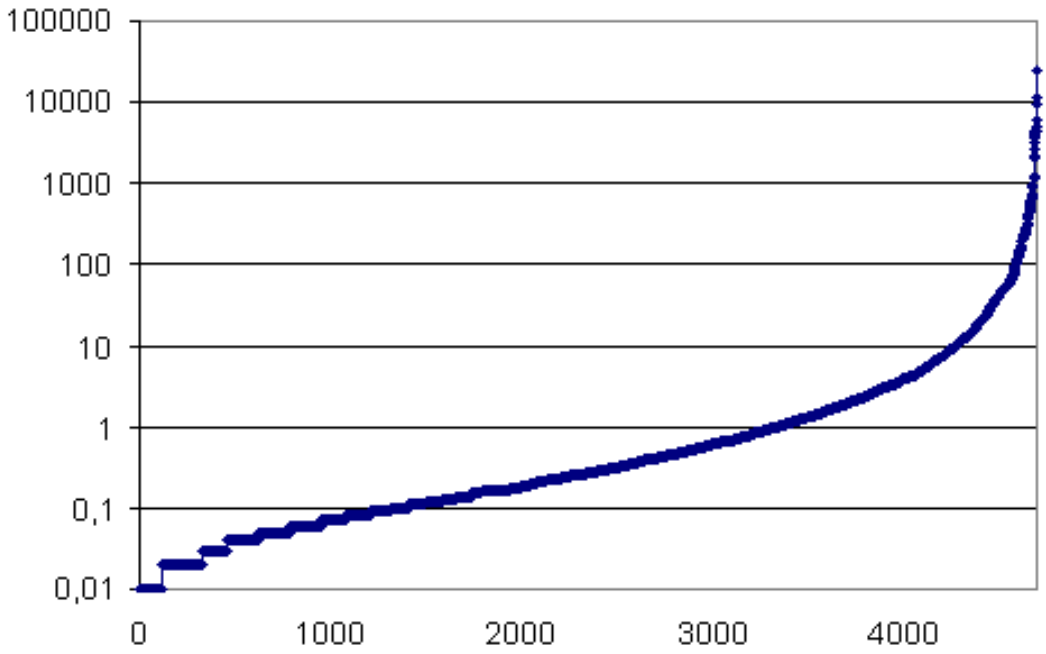
(143) *To, co ja widzę dokładnie, przymykam oczy i widzę przecież wyraźnie skrzypiącą bramę, podwójne druty kolczaste, wartownika na wieży, korowód wychudzonych cieni, ja to widzę, a oni mogą co najwyżej odtwarzać na podstawie tego, co znajdą w ziemi.* *[5361]*

(144) *Parametrami, które ilościowo mogą określić przydatność paliwa pod względem mechanicznym, są moduł Younga, dopuszczalne naprężenie, udarność i tym podobne, które ilościowo mogą określić przydatność paliwa pod względem mechanicznym, ostatecznego jakiegoś ogólnego ujęcia zagadnień wytrzymałościowych paliw i wpływu tych własności na balistykę wewnętrzną silników dotychczas brak.* *[3383]*

Rys. 13.1 prezentuje czasy analizy wszystkich pozostałych wypowiedników posortowane rosnąco. Teksty 2863 wypowiedników (60,71%) zostały zanalizowane w czasie poniżej 0,5 sekundy, 3343 wypowiedników (70,89%) — poniżej 1 sekundy, 4275 wypowiedników (90,65%) — poniżej 10 sekund. Łączny czas analizy wszystkich wypowiedników wyniósł ponad 38 godzin.

Długość najdłuższego analizowanego wypowiednika wynosi 61 słów (w całym korpusie — 72 słowa), średnia arytmetyczna długości wypowiednika — 8,61 słów, podczas gdy mediana — 6 słów. Rys. 13.2 przedstawia zależność czasu analizy wypowiednika od jego długości, zaś rys. 13.3 — mediany czasów przetwarzania dla wypowiedników o danej długości.

[6]Patrz rozdział 8.3, s. 85.



Rysunek 13.1: Czasy analizy posortowane rosnąco

13.3 Kwestia wieloznaczności

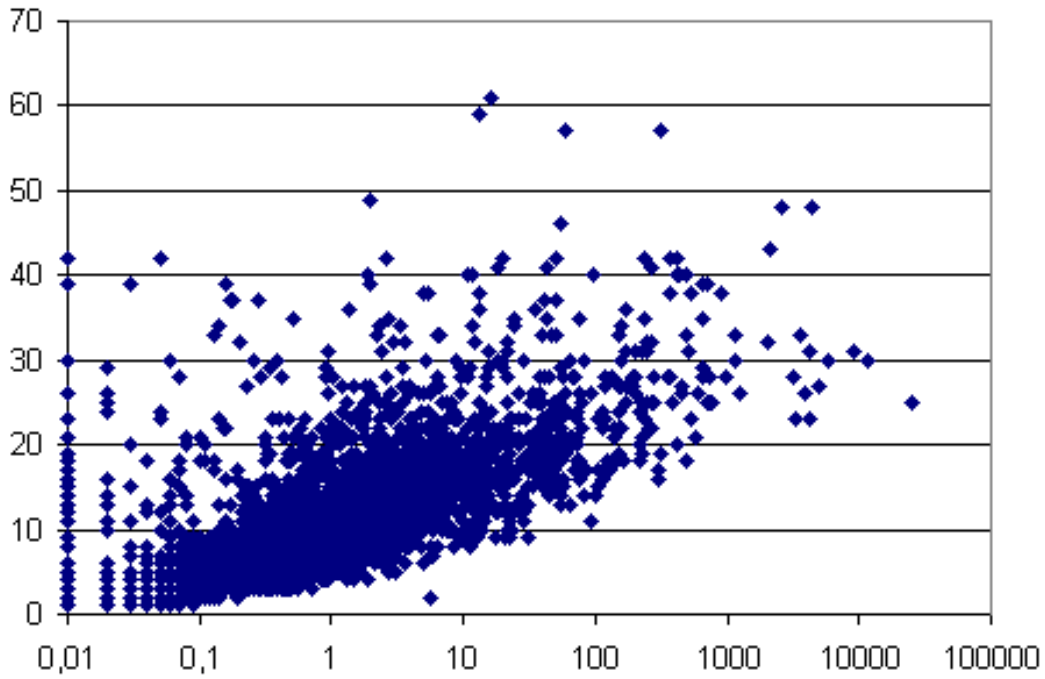
Poruszona w rozdziale 12.3 (s. 148) kwestia wieloznaczności może być również źródłem wielu interesujących zadań weryfikacyjnych. W niniejszym rozdziale przedstawiam dwa z nich zdając sobie sprawę, że nie wyczerpuje to tematu; kilka innych pomysłów weryfikacyjnych w tym aspekcie sygnalizuję w rozdziale 14.2 (s. 168).

Ze względu na dużą złożoność procesu konstrukcji drzew prezentowane niżej obliczenia przeprowadziłem wyłącznie dla wypowiedników akceptowanych o czasie analizy nie przekraczającym pół sekundy (łącznie 2559 wypowiedników, co daje ponad 54% analizowanej próbki).

13.3.1 Liczba izomorficznych drzew rozbioru

Jednym z ciekawszych kwestii z zakresu badania wpływu wieloznaczości na otrzymywane wyniki z punktu widzenia optymalizacji gramatyki jest zbadanie liczby typowych wieloznaczności składniowych wyrażających się podziałem zestawu wynikowych drzew rozbioru na klasy abstrakcji pod względem kształtu drzewa. Każdy rodzaj wieloznaczności tego rodzaju wskazuje znacząco różną drogę analizy.

W przeprowadzonym eksperymencie zaniedbywałem wieloznaczności morfologiczne oraz niektóre syntaktyczne, wprowadzane np. przez potencjalnie nadmiarowe zestawu wartości parametrów klauzul, skupiłem się natomiast na analizie liczby drzew istotnie różnych pod względem kształtu. Przykładowo, zdanie *„Mam akwarium.”* ma wg Świgry 13 drzew rozbioru (z frazą finitywną o formach hasłowych *„mieć”*



Rysunek 13.2: Czasy analizy wypowiedników o danej długości (w słowach)

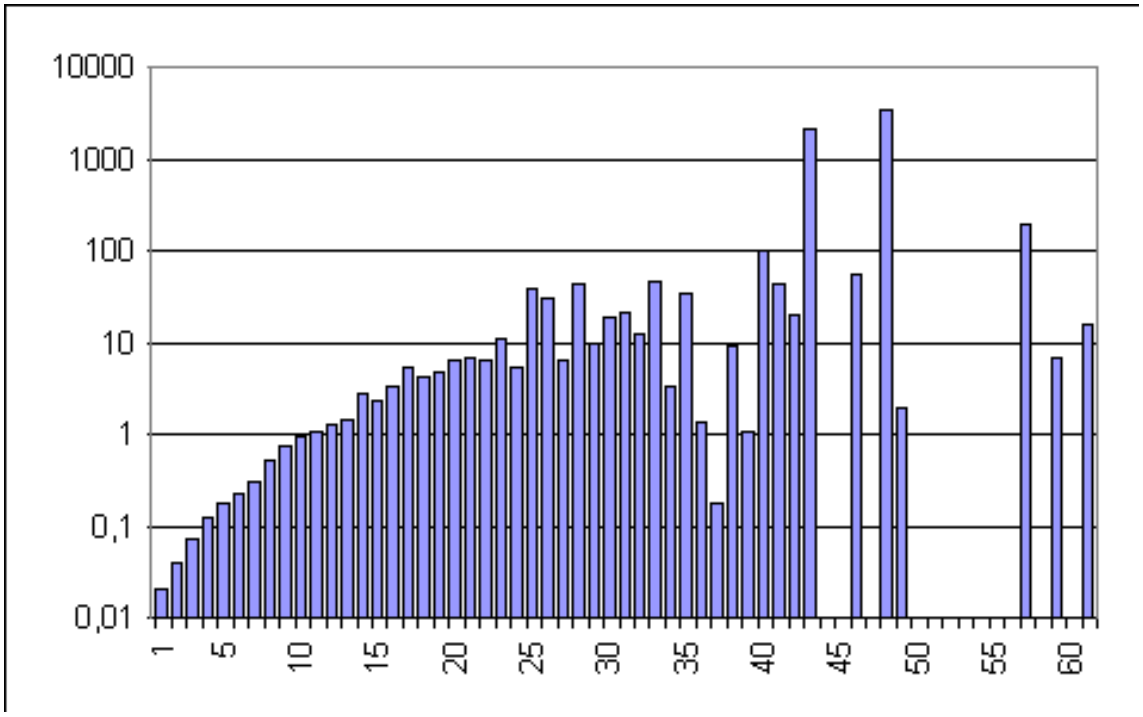
i *„mamić”* oraz z frazą luźną *„akwarium”* w celowniku, dopełniaczu, bierniku, narzędniku i odpowiadającymi im analizami z frazą wymaganą), lecz mają one jedynie dwa istotnie różne kształty — z frazą wymaganą i frazą luźną.

Rys. 13.4 prezentuje różnicę między średnią arytmetyczną liczby drzew uzyskiwanych w procesie analizy składniowej a średnią liczbą drzew nieizomorficznych. Dane użyte do generacji tego i następnego wykresu zawiera w zbiorczej formie tabela 13.6 umieszczona na końcu rozdziału (s. 166).

Pierwszą obserwację stanowi stosunkowo duży udział drzew o identycznym kształcie — różniących się wyłącznie wartościami parametrów klauzul. Mógłby być on wynikiem wieloznaczności morfologicznej, jednak w przypadku GFJP wydaje się być konsekwencją zbyt słabych niekiedy ograniczeń nakładanych na parametry konstrukcji składniowych[7].

Drugie ważne spostrzeżenie stanowi fakt, że mimo występowania w wyniku dużej liczby drzew o identycznym kształcie liczba drzew nieizomorficznych — a więc odpowiadających znacząco różnym ścieżkom analizy — wciąż pozostaje spora. Sytuacja ta odpowiada zapewne, jak w przypadku długiej analizy niektórych wypowiedników, powieleniu pewnych nadmiarowych konstrukcji składniowych, zatem przytoczone dane mogą wskazywać optymalny sposób wykorzystania wyników do poprawy jakości gramatyki: poprzez weryfikację przykładów najkrótszych. Do tego celu nadaje się zapewne najlepiej zestaw testowy wykorzystany w rozdziale 13.2.1 (s. 160) i zadanie to wydaje się ważniejsze niż możliwość wykorzystania Świgry do przetwarzania dużych zbiorów tekstów.

[7]Patrz przykłady konstrukcji niepoprawnych w rozdziale 12.3, s. 148.



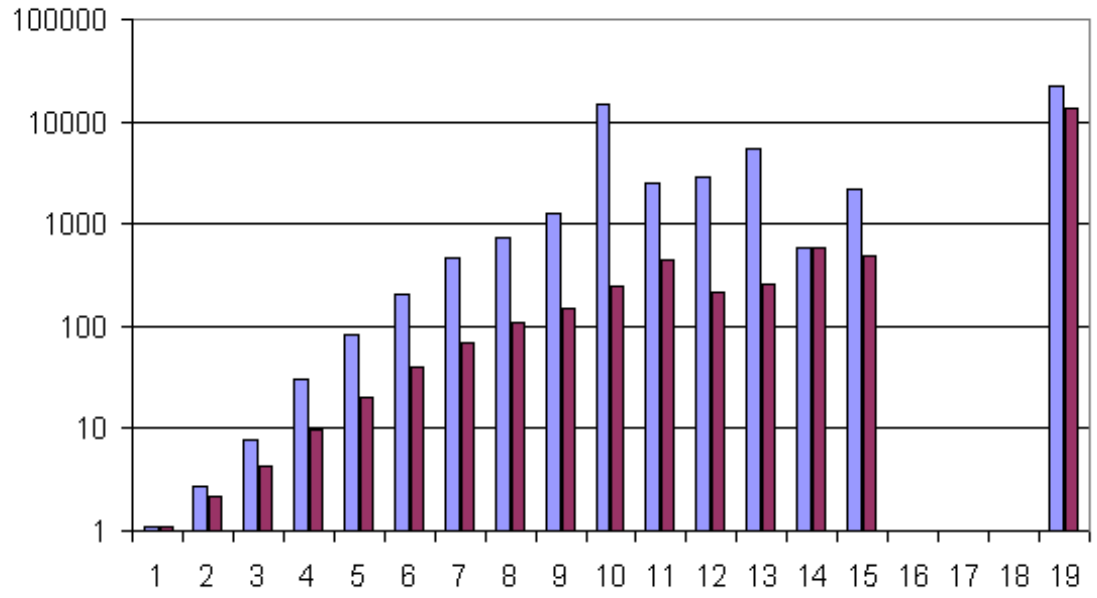
Rysunek 13.3: Mediany czasów przetwarzania dla wypowiedników o danej długości (w słowach)

13.3.2 Wyniki eliminacji fraz luźnych

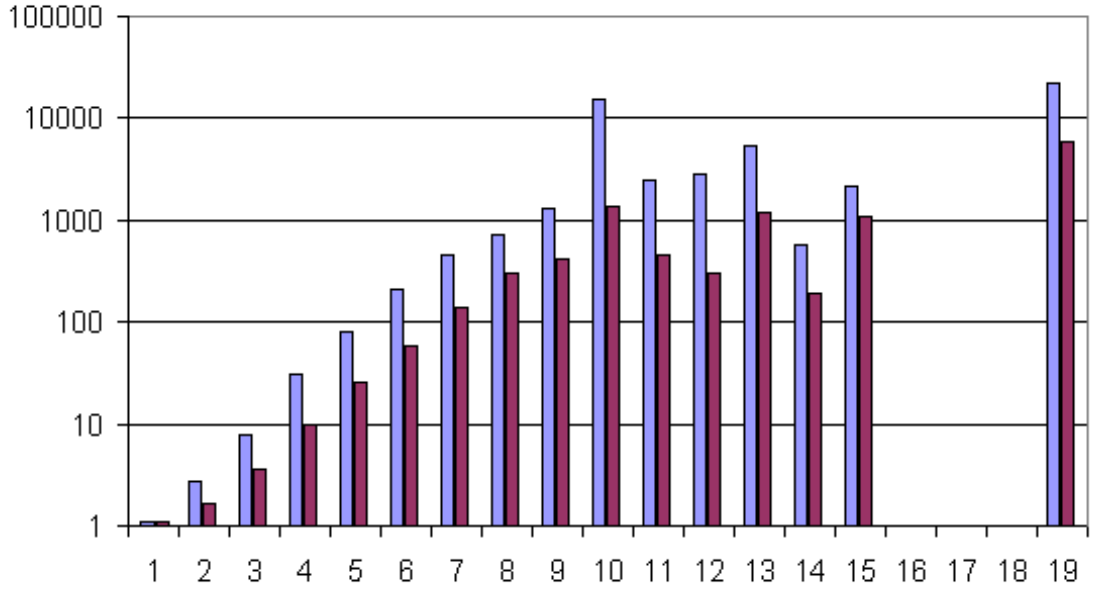
Dla opisanego w rozdziale 11.4 (s. 136) eksperymentu polegającego na wykluczeniu analizy z frazą luźną gdy istnieje odpowiadająca mu analiza z frazą wymaganą kluczowa staje się kwestia sprawdzenia, w jakim stopniu ulega zmniejszeniu liczba drzew rozbioru po zastosowaniu takiego wykluczenia.

Rys. 13.5 prezentuje różnicę między średnią arytmetyczną liczby drzew uzyskiwanych w procesie analizy składniowej z wykorzystaniem Świgry a średnią liczbą drzew po eliminacji analiz zawierających frazy luźne.

Przeprowadzony eksperyment wykazał także, podobnie jak w przypadku eliminacji drzew izomorficznych, znaczący udział w ogólnej liczbie rozbiorów wynikowych konstrukcji nieobejmujących fraz luźnych — liczba wieloznaczności, jakkolwiek zmniejszyła się znacząco po usunięciu fraz luźnych, wciąż pozostała duża. Sytuację tę można zinterpretować podobnie, jako świadectwo dużego stopnia swobody cechującego gramatykę Świdzińskiego.



Rysunek 13.4: Średnia liczba drzew dla wypowiedników danej długości (kolor jasny — liczba drzew uzyskiwana podczas analizy, kolor ciemny — liczba drzew nieizomorficznych)



Rysunek 13.5: Średnia liczba drzew dla wypowiedników danej długości (kolor jasny — liczba drzew uzyskiwana podczas analizy, kolor ciemny — liczba drzew po eliminacji analiz zawierających frazy luźne)

Długość wypowiednika w słowach	**Liczba wypowiedników o danej długości**	**Średnia liczba drzew uzyskiwanych Świgrą**	**po eliminacji fraz luźnych**	**Procentowe zmniejszenie liczby drzew**	**Średnia liczba różnych kształtów drzew**	**Różnica procentowa między liczbą drzew a liczbą drzew różnych**
1	211	1,01	1,01	0,00%	1,00	1,40%
2	377	2,69	1,67	38,01%	2,13	20,83%
3	483	7,90	3,60	54,38%	4,20	46,80%
4	482	30,30	9,88	67,40%	9,63	68,23%
5	367	82,04	25,12	69,38%	19,99	75,64%
6	238	204,95	59,40	71,02%	40,64	80,17%
7	175	460,31	135,90	70,48%	67,66	85,30%
8	93	731,55	299,12	59,11%	107,57	85,30%
9	57	1281,96	414,56	67,66%	149,77	88,32%
10	38	15046,68	1370,84	90,89%	246,68	98,36%
11	17	2488,47	460,76	81,48%	445,59	82,09%
12	9	2892,89	304,67	89,47%	217,44	92,48%
13	6	5426,00	1164,67	78,54%	255,67	95,29%
14	1	576,00	192,00	66,67%	576,00	0,00%
15	4	2146,50	1066,50	50,31%	483,00	77,50%
16	0	—	—	—	—	—
17	0	—	—	—	—	—
18	0	—	—	—	—	—
19	1	21780,00	5940,00	72,73%	13344,00	38,73%
Dla wszystkich wypowiedników:		737,17	231,77	57,47%	122,10	60,53%

Tabela 13.6: Wyniki liczbowe operacji eliminacji fraz luźnych i badania liczby drzew nieizomorficznych w zależności od długości wypowiednika

Rozdział 14

Perspektywy dalszych prac

Niniejsza praca stanowi eksperyment badawczy dostarczający danych mogących posłużyć dalszej analizie definicji polszczyzny zawartej w Gramatyce Formalnej Języka Polskiego. Pierwszy krok tego procesu stanowi właśnie dokonana automatyczna analiza morfologiczno-składniowa tekstu skonfrontowana z efektami analizy manualnej, co stało się możliwe dopiero niedawno, po powstaniu analizatora Marcina Wolińskiego. Dalsza praca nad uzupełnieniem formalnego opisu polszczyzny i rozwojem mechanizmów służących jego przetwarzaniu jest jednak niezbędna, dlatego poniżej przedstawiam kilka możliwych, niezależnych kierunków działań.

14.1 Rozwój gramatyki Świdzińskiego

Osiągnięta w tej pracy akceptowalność przykładów testowych na poziomie 85% wydaje się do bieżącego zastosowania wystarczająca. Gramatykę warto jednak rozwijać dalej, a prace w tym kierunku są już prowadzone zarówno przez jej Autora, jak i niezależnie od niego, m. in. w Instytucie Podstaw Informatyki Polskiej Akademii Nauk — dostępność Świgry znacznie je ułatwia.

Od mozolnych prób (i błędów) uzupełniania gramatyki ważniejsze może się jednak okazać spełnienie potrzeby stworzenia ogólnej koncepcji jej rozwoju, gdyż wiele alternatywnych ścieżek wprowadzania niezbędnych zmian może łatwo doprowadzić do zatracenia ducha (względnej) prostoty opisu, do czego przyznaje się po części i autor tej pracy. Poniżej odnoszę się do kilku poruszanych na etapie optymalizacji gramatyki zagadnień, zdając sobie jednak sprawę, że jest to jedynie ułamek potencjału GFJP.

Pierwszym z nich jest wykorzystanie sporych możliwości drzemiących w pełnym opisie grupy nominalnej, przygotowanym do rozszerzenia definicji o pozostałe klasy[1] — samodzielny zaimek przymiotny czy zaimek rzeczowny. Ciekawą i potrzebną modyfikacją mogłaby być także rozbudowa interpunkcyjnych cech grupy, a więc i całej gramatyki, o konstrukcje apozycyjne (patrz rozdział 10.3.1, s. 114).

Podobnie uproszczony jest zaproponowany obecnie opis liczebników. Jego największą wadą jest brak rozwiązania problemu uzgodnień nielokalnych, właściwych stwier-

[1] Patrz [Szpakowicz i Świdziński, 1990], s. 5-7.

dzeniom w rodzaju *„Pięciu skrzypków było zmęczonych."* czy *„Dwie śpiewaczki przyjadą chore."*. Do obsługi tego rodzaju konstrukcji konieczne byłoby wprowadzenie do GFJP nowego schematu zdaniowego z wymaganą frazą liczebnikowo-nominalną o którym wiem, że jest w fazie opracowania przez Autora gramatyki.

Opis nie uwzględnia ponadto np. możliwych do realizacji konstrukcji z artykułu [Saloni i Gruszczyński, 1978], takich jak rozszerzenie zbioru leksemów liczebnikowych o niektóre rzeczowniki[2] (*ćwierć*, *paręset* itp.) czy wspomniane już wyżej konstrukcje z „niestandardowymi" podrzędnikami dystrybucyjnymi. Wchodzą one w specyficzne uzgodnienia z pozostałymi elementami trójek podstawowych[3], których analiza nie była przedmiotem niniejszej pracy.

Nie zostało też wykorzystanych wiele szczegółowych, acz nie mniej ważnych własności liczebników, jak m. in. kategorii deprecjatywności (formy deprecjatywne rzeczowników rodzaju męskoosobowego zachowują się jak formy rodzaju męskozwierzęcego), propozycja rozciągnięcia której na klasę gramatyczną liczebników pojawiła się w artykule [Bień i Saloni, 1982]. W końcu przydatny byłby także opis grup liczebnikowych.

14.2 Dalsza weryfikacja gramatyki i wyników analizy składniowej

Pożytecznym zadaniem wskazywanym także przez Świdzińskiego[4] mogłoby stać się objęcie analizą większego zakresu materiału korpusowego — niekoniecznie w ścisłym rozumieniu zastosowania technik analizy manualnej opisanej w pracy [Świdziński, 1996], lecz choćby jedynie poprzez użycie analizatora automatycznego i późniejszą weryfikację jego wyników.

Ważnym zagadnieniem jest też poruszona już w rozdziale 12.3 (s. 148) kwestia wieloznaczności syntaktycznej i morfologicznej oraz jej wpływu na liczbę otrzymywanych wyników. Opisany w rozdziale 13.3.1 (s. 162) eksperyment obliczeniowy można by rozszerzyć na badanie pełnych list argumentów klauzul wynikowych w celu określenia liczby drzew różniących się wyłącznie parametrami morfologicznymi czy analizę zakresu interwencji parametrów nieukonkretnionych. W tym kontekście warto zwrócić uwagę na pozostałe możliwości analizy wyników liczbowych, w szczególności np. sprawdzenie:

- typów klas abstrakcji podziału wyników analizy składniowej względem izomorficzności drzew rozbioru,
- rozkładu wystąpień w zbiorze drzew analiz określonych węzłów, z uwzględnieniem lub zaniedbaniem wartości zawartych w nich parametrów składniowych,
- „wspólnych mianowników" rozbioru, czyli fragmentów drzew powtarzających się w wynikach i reprezentujących „stabilne" analizy cząstkowe.

[2] Patrz pkt 0.2.7, s. 27.

[3] Por. rozdział 10.1.1, s. 100. Saloni wskazuje w przypadku leksemu *paręset* konieczność jego samodzielnego wystąpienia, wydaje się jednak, że — choć skrajnie rzadka — poprawna składniowo byłaby forma *paręset dziesięć*.

[4] [Świdziński, 1996], s. 155.

14.3 Rozwój narzędzi analizy

W kwestii narzędzi analizy warto zwrócić uwagę na problemy zauważone już przez Wolińskiego[5]: konieczność rozbudowy słownika analizatora morfologicznego oraz słownika wymagań czasownikowych. Oba te postulaty znalazły oddźwięk w przeprowadzonych zadaniach weryfikacyjnych, a zebrane dane zostały udostępnione do wykorzystania.

Wciąż nie zrealizowanym pomysłem pozostaje dodanie do analizatora składniowego mechanizmu reprezentacji informacji frazeologicznej (jako — wg Szpakowicza i Świdzińskiego[6] — podlegającej składniowym metodom opisu) i jej wykorzystanie w procesie analizy składniowej. Szczątkowa obsługa związków frazeologicznych w bieżącej wersji analizatora ograniczyła się do uzupełnienia ogólnego słownika wymagań czasownikowych, co w oczywisty sposób może prowadzić do ich przeciążenia. Frazeologizmy inne niż czasownikowe są reprezentowane w jeszcze prymitywniejszy sposób, mianowicie jedynie poprzez sklejanie analiz morfologicznych wielu słów w jedną analizę wielowyrazową. Jednym z najprostszych rozwiązań tego nietrywialnego problemu mogłoby być utworzenie słownika frazeologizmów, zawierającego z jednej strony wykaz związków o stałej łączliwości, a z drugiej wzorcowe drzewa cząstkowego rozbioru danego związku. Jest to idea pokrewna do użytej obecnie, ograniczona do stałych połączeń wyrazowych realizowanych jako elementarne jednostki wynikowe, jednak reprezentacja danych i sposób ich użycia wymaga na pewno głębszego przemyślenia.

[5]Patrz [Woliński, 2004], s. 115.

[6]Patrz [Świdziński i Szpakowicz, 1989].

Podsumowanie

W pracy dokonana została wieloaspektowa weryfikacja korpusu wypowiedników polskich Marka Świdzińskiego — bazy zdań i oznajmień z naniesioną charakterystyką i strukturą składniową jednostek elementarnych, stworzonej do różnych przedsięwzięć lingwistycznych, w tym badań nad polską składnią.

Inicjalna czynność procesu weryfikacji reprezentowanych danych, weryfikacja grafemiczna, pozwoliła na wyeliminowanie większości błędów zapisu oraz przygotowała materiał do dalszej pracy. Ważny etap weryfikacji stanowiło też porównanie dostępnej wersji korpusu z materiałem źródłowym oraz innymi zbiorami danych tworzonymi na jego bazie, a przez to niezależnie przejrzanymi i poprawionymi. Weryfikacja morfologiczna korpusu pozwoliła z jednej strony na dokonanie dodatkowego sprawdzenia warstwy typograficznej, z drugiej — na rozbudowę i korektę zasobów źródłowych użytego analizatora morfologicznego.

Główną część pracy wypełniły wnioski z procesu weryfikacji składniowej korpusu wykorzystującej gramatykę formalną języka polskiego Świdzińskiego oraz analizator składniowy Świgra Marcina Wolińskiego. Na szczególną uwagę zasługuje fakt, że zadanie to było pierwszym zastosowaniem pełnej postaci gramatyki do autentycznych tekstów. Korzystając z dostępnych wyróżnień jednostek składniowych poziomu frazowego oprócz analizy składniowej pełnych tekstów wypowiedników dokonano osobnej weryfikacji składni fraz.

Etap weryfikacji składniowej wymagał rozszerzenia gramatyki Świdzińskiego o konstrukcje językowe używane w korpusie wypowiedników, a nie reprezentowane do tej pory w gramatyce, takie jak konstrukcja liczebnikowa czy grupy składniowe. Dokonano także wielu drobnych, ale niezbędnych usprawnień w zakresie akceptowanych konstrukcji językowych oraz zweryfikowano hipotezy o domniemanej kolejności drzew analizy oraz o równoważności dystrybucyjnej jednostek zdaniowych. Proces weryfikacji składni dostarczył także danych do porównania gramatyki Świdzińskiego z jej uproszczonym wariantem użytym do reprezentacji struktur składniowych w korpusie oraz do analizy aspektu wieloznaczności danych korpusowych.

Osobny etap stanowiła analiza pochodnych danych lingwistycznych pozyskanych na bazie próbek korpusu, mianowicie rozkładu schematów zdaniowych, realizacji fraz poszczególnych typów, porządku składników zdania elementarnego czy typologii oznajmień. Nawiązując do wcześniejszego, opartego na słownikowej kwerendzie projektu składniowego słownika czasowników Świdzińskiego, na danych korpusu powstał słownik czasowników z informacją składniową w identycznym formacie, wykorzystany także do uzupełnienia słownika wymagań czasownikowych analizatora składniowego.

W ramach pracy dane bazy wypowiedników zostały zapisane w postaci korpusu rozbiorów gramatycznych w formacie XML-owym. Próbkę korpusu stanowi jednostka poziomu wypowiedzenia zawierająca komplet informacji składniowej, oryginalnie dostępnej wyłącznie dla jednostek elementarnych.

Ważny aspekt pracy stanowi także udostępnienie analizatora składniowego Świgra Marcina Wolińskiego w środowisku Windows oraz stworzenie na jego bazie zestawu narzędzi do przetwarzania korpusu wypowiedników, które mogą — po niewielkim dostosowaniu — okazać się przydatne do analizy morfologicznej i składniowej dowolnych korpusów tekstów.

Bibliografia

[Bańko, 1990] Bańko, M., 1990. *Niektóre problemy oceny adekwatności gramatyk (na przykładzie fragmentu gramatyki Szpakowicza)*. Studia gramatyczne IX, Wrocław.

[Bański, 2001] Bański, P., 2001. *The proposed encoding scheme for the IPI PAN corpus (7 T11C 043 20)*. Warszawa, grudzień 2001. `http://nlp.ipipan.waw.pl/CORPUS/banski_raport.pdf`.

[Bień, 1996] Bień, J. S., 1996. *Komputerowa weryfikacja opisu składni polskiej*. Raport Instytutu Informatyki Uniwersytetu Warszawskiego TR 96-06 (227), maj 1996.

[Bień, 1997a] Bień, J. S., 1997a. *Komputerowa weryfikacja formalnej gramatyki Świdzińskiego*. Biuletyn Polskiego Towarzystwa Językoznawczego, zeszyt LII (1997), s. 147–164.

[Bień, 1997b] Bień, J. S., 1997b. *Konstrukcje werbalne z aż w gramatyce Świdzińskiego*. Tom dedykowany prof. Marii Szupryczyńskiej.

[Bień, 2000] Bień, J. S., 2000. *Zestaw testów do weryfikacji i oceny parserów języka polskiego. Sprawozdanie merytoryczne (nieznacznie zmodyfikowane) z projektu KBN 8 T11C C 002 13*. `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/tajp/`.

[Bień i Saloni, 1982] Bień, J. S. i Z. Saloni, 1982. *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*. Prace Filologiczne, tom XXXI, s. 31–45, Warszawa 1982.

[Colmerauer, 1978] Colmerauer, A., 1978. *Metamorphosis grammars*. [w:] L. Bolc (red.), *Natural Language Communication with Computers*, Lecture Notes in Computer Science 63, Springer-Verlag 1978, s. 133–189.

[Derwojedowa, 2000] Derwojedowa, M., 2000. *Porządek linearny składników zdania elementarnego w języku polskim*. Dom Wydawniczy Elipsa, Warszawa 2000.

[Derwojedowa i Rudolf, 2003] Derwojedowa, M. i M. Rudolf, 2003. *Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu*. Poradnik Językowy, zeszyt 5.

[Derwojedowa i in., 2003] Derwojedowa, M., M. Rudolf, i M. Świdziński, 2003. *Two formal approaches to Polish numeral phrases implemented*. [w:] *Studia z gramatyki i leksykologii języka polskiego*, red. M. Gębka-Wolak, I. Kaproń-Charzyńska, M. Urban. s. 93–108, Uniwersytet Mikołaja Kopernika, Toruń.

[Gazdar i in., 1985] Gazdar, G., E. Klein, G. Pullum, i I. Sag, 1985. *Generalized Phrase Structure Grammar*. Blackwell Publishing, Oxford, England, and Harvard University Press, Cambridge, Massachusetts, 1985.

[Gazdar i Mellish, 1989] Gazdar, G. i Ch. Mellish, 1989. *Natural Language Processing in Prolog*. Brighton, Edinburgh, March 1989. `http://www.informatics.sussex.ac.uk/research/nlp/gazdar/nlp-in-prolog/`.

[Greń, 2001] Greń, Z., 2001. *Słowniki składniowe języka czeskiego i polskiego*.

[Ide i in., 2000] Ide, N., P. Bonhomme, i L. Romary, 2000. *XCES: An XML-based standard for linguistic corpora*. [w:] Proceedings of the linguistic resources and evaluation conference. Athens, Greece. Patrz `http://www.cs.vassar.edu/XCES/`.

[Ide i in., 1996] Ide, N., G. Priest-Dorman, i J. Véronis, 1996. *Corpus encoding standard*. Maszynopis. `http://www.cs.vassar.edu/CES/`.

[Kallas, 1980] Kallas, K., 1980. *Grupy apozycyjne we współczesnym języku polskim*. Rozprawy Uniwersytetu Mikołaja Kopernika, Toruń 1980.

[Kay, 2005] Kay, M., 2005. *XSLT Programmer's Reference*. Wrox Press Ltd. Wydanie drugie. ISBN 18-6100-506-7.

[Kopcińska, 1997] Kopcińska, D., 1997. *Strukturalny opis składniowy zdań z podmiotem–mianownikiem we współczesnej polszczyźnie*. Dom Wydawniczy Elipsa, ISBN 83–7151–228–7. Warszawa 1997.

[Kurcz i in., 1990] Kurcz, I., A. Lewicki, J. Sambor, K. Szafran, i J. Woronczak, 1990. *Słownik frekwencyjny polszczyzny współczesnej*. Kraków, 1990. Instytut Języka Polskiego PAN.

[Kurcz i in., 1974a] Kurcz, I., A. Lewicki, J. Sambor, i J. Woronczak, 1974a. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom I. Teksty popularnonaukowe*. Warszawa 1974. Uniwersytet Warszawski. s. 4 nlb., 428 + 2 nlb., 429–858.

[Kurcz i in., 1974b] Kurcz, I., A. Lewicki, J. Sambor, i J. Woronczak, 1974b. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom II. Drobne wiadomości prasowe*. Warszawa 1974. Uniwersytet Warszawski. s. 2 nlb., 398 + 2 nlb., 399–792.

[Kurcz i in., 1976] Kurcz, I., A. Lewicki, J. Sambor, i J. Woronczak, 1976. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom IV. Proza artystyczna*. Warszawa 1976. Uniwersytet Warszawski. s. 282 + 2 nlb., 283–557 + 1 nlb., 558–885.

[Kurcz i in., 1977] Kurcz, I., A. Lewicki, J. Sambor, i J. Woronczak, 1977. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom V. Dramat artystyczny*. Warszawa 1977. Uniwersytet Warszawski. s. 320 + 2 nlb., 321–631.

[Lewicki i in., 1975] Lewicki, A., W. Masłowski, J. Sambor, i J. Woronczak, 1975. *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom III. Publicystyka*. Warszawa 1975. Uniwersytet Warszawski. s. 4 nlb., 340 + 2 nlb., 341–684.

[Nazarczuk, 1997] Nazarczuk, M., 1997. *Wstępne przygotowanie korpusu „Słownika frekwencyjnego polszczyzny współczesnej" do dystrybucji na CD-ROM*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa 1997. Instytut Języka Polskiego Uniwersytetu Warszawskiego. 59 s., płyta CD.

[Ogrodniczuk, 2000] Ogrodniczuk, M., 2000. *Wykorzystanie SGML i TEI do zapisu polskich danych lingwistycznych*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa 2000. Instytut Informatyki Uniwersytetu Warszawskiego. 83 s., płyta CD.

[Ogrodniczuk, 2001] Ogrodniczuk, M., 2001. *DTD i XML Schema, czyli XML pod kontrolą*. Artykuł dla czasopisma Software 2.0, 6/2001.

[Ogrodniczuk, 2003a] Ogrodniczuk, M., 2003a. *Nowa edycja wzbogaconego korpusu słownika frekwencyjnego*. [w:] Stanisław Gajda (red.), Językoznawstwo w Polsce. Stan i perspektywy. Polska Akademia Nauk — Komitet Językoznawstwa, Uniwersytet Opolski — Instytut Filologii Polskiej. Opole 2003, s. 181–190. `http://www.mimuw.edu.pl/jsbien/MO/JwP03/`.

[Ogrodniczuk, 2003b] Ogrodniczuk, M., 2003b. *Rozszerzenie opisów morfologicznych w tekstach korpusu* Słownika frekwencyjnego polszczyzny współczesnej. [w:] Jadwiga Linde-Usiekniewicz, Romuald Huszcza (red.) Prace językoznawcze dedykowane Profesor Jadwidze Sambor. Wydział Polonistyki Uniwersytetu Warszawskiego. Warszawa 2003, s. 164–168.

[Ogrodniczuk, 2004] Ogrodniczuk, M., 2004. *From SGML to XML with TEI: Automated conversion of a corpus of Polish from P3 to P4 format*. Investigationes Linguisticae. Poznań, grudzień 2004. Instytut Lingwistyki, Uniwersytet Adama Mickiewicza. `http://www.staff.amu.edu.pl/~inveling/maciej_ogrodniczuk_inve11.pdf`.

[Ogrodniczuk, 2005a] Ogrodniczuk, M., 2005a. *An extension of Świdziński's grammar of Polish*. [w:] *Archives of Control Sciences*, volume 15 (LI), nr 3, s. 251–261.

[Ogrodniczuk, 2005b] Ogrodniczuk, M., 2005b. *Restructuring Świdziński's grammar of Polish*. [w:] Proceedings of 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Wydawnictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza. Poznań 2005, s. 177–181.

[Polański, 1966] Polański, K., 1966. *Główne typy struktur zdaniowych w języku polskim*. Zeszyty Naukowe WSP w Katowicach. Prace Językoznawcze III. Katowice, s. 83–99.

[Polański, 1980–1988] Polański, K., 1980–1988. *Słownik syntaktyczno-generatywny czasowników polskich*. Pod red. Kazimierza Polańskiego, Wrocław 1980 — t. I, 1984 — t. II, 1988 — t. III.

[Prinke, 2000] Prinke, Rafał T., 2000. *Fontes ex machina: komputerowa analiza źródeł historycznych*. Polska Akademia Nauk, Biblioteka Kórnicka, Centrum Elektronicznych Tekstów Humanistycznych PAN, Poznań 2000. 342 s.

[Przepiórkowski, 2004] Przepiórkowski, A., 2004. *Korpus IPI PAN — wersja wstępna*. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2004. ISBN 83-910948-8-X. `http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_pl.pdf`.

[Przepiórkowski i Świdziński, 1997] Przepiórkowski, A. i M. Świdziński, 1997. *Polish Verbal Negation Revisited: A Metamorphosis vs HPSG Account*. Prace IPI PAN 829. Instytut Podstaw Informatyki Polskiej Akademii Nauk. Warszawa, 2004. `http://nlp.ipipan.waw.pl/~adamp/Papers/1997-829/neg-fgp-rep.pdf`.

[Rudolf, 2004] Rudolf, M., 2004. *Metody automatycznej analizy korpusu tekstów polskich. Pozyskiwanie, wzbogacanie i przetwarzanie informacji lingwistycznych*. Praca doktorska. Uniwersytet Warszawski, Wydział Polonistyki.

[Saloni, 1974] Saloni, Z., 1974. *Klasyfikacja gramatyczna leksemów polskich*. Język Polski LIV, z. 1, s. 3–13, z. 2, s. 93–101.

[Saloni, 1976] Saloni, Z., 1976. *Cechy składniowe polskiego czasownika*. Ossolineum, Wrocław 1976.

[Saloni, 1977] Saloni, Z., 1977. *Kategorie gramatyczne liczebników we współczesnym języku polskim*. Studia gramatyczne I. Wrocław 1977.

[Saloni, 1982] Saloni, Z., 1982. *Uwagi o opisie fleksyjnym tzw. zaimków rzeczownych*. *Acta Universitatis Lodziensis*, Folia Linguistica II, Łódź 1982, s. 243–253.

[Saloni, 2001] Saloni, Z., 2001. *Czasownik polski. Odmiana, słownik*. Wiedza Powszechna, Warszawa 2001.

[Saloni i Gruszczyński, 1978] Saloni, Z. i W. Gruszczyński, 1978. *Składnia grup liczebnikowych we współczesnym języku polskim*. [w:] Studia gramatyczne II. Wrocław, s. 17–42, Ossolineum 1978.

[Saloni i Świdziński, 1981] Saloni, Z. i M. Świdziński, 1981. *Składnia współczesnego języka polskiego*. Warszawa 1981. Wydanie II poprawione i zmienione: Warszawa 1985. Wydanie III: Warszawa 1987. Wydanie IV, zmienione: Warszawa 1998. Wydawnictwo Naukowe PWN.

[SGML, 1986] SGML, 1986. *International Standard ISO 8879 Information Processing — Text and Office Systems — Standard Generalized Markup Language (SGML)*. ISO (International Organization for Standardization). Genewa, 1986.

[SJP, 2002] SJP, 2002. *Słownik języka polskiego pod red. M. Szymczaka*. Wydawnictwo Naukowe PWN, Warszawa 2002.

[Skibicki, 2000] Skibicki, K., 2000. *Komputerowa weryfikacja wybranych zasobów lingwistycznych*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa 2000. Instytut Informatyki Uniwersytetu Warszawskiego.

[Skorupka, 1977] Skorupka, S., 1977. *Słownik frazeologiczny języka polskiego*. Wiedza Powszechna, Wydanie III. Warszawa 1977.

[Szafran, 1996] Szafran, K., 1996. *Analizator morfologiczny SAM-95 — opis użytkowy*. Raport Instytutu Informatyki Uniwersytetu Warszawskiego TR 96–05 (226), maj 1996. `ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/tr226.ps`.

[Szpakowicz, 1978] Szpakowicz, S., 1978. *Automatyczna analiza składniowa polskich zdań pisanych*. Niepublikowana rozprawa doktorska.

[Szpakowicz, 1986] Szpakowicz, S., 1986. *Formalny opis składniowy zdań polskich*. Wydanie drugie, Wydawnictwa UW, Warszawa 1986. `ftp://ftp.mimuw.edu.pl/pub/user/polszczyzna/Szpakowicz/`.

[Szpakowicz i Świdziński, 1981] Szpakowicz, S. i M. Świdziński, 1981. *Zarys klasyfikacji schematów zdaniowych we współczesnej polszczyźnie pisanej*. Polonica VII, s. 5–35.

[Szpakowicz i Świdziński, 1990] Szpakowicz, S. i M. Świdziński, 1990. *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej*. [w:] Studia Gramatyczne IX, s. 9–54. Polska Akademia Nauk, Instytut Języka Polskiego, Ossolineum. Wrocław 1990. ISBN 83-04-03303-8.

[Świdziński, 1987] Świdziński, M., 1987. *Formalny opis składniowy polskich zdań o składniku zdaniowym*. Praca habilitacyjna (maszynopis powielony). Wydział Polonistyki UW, Warszawa 1987.

[Świdziński, 1992a] Świdziński, M., 1992a. *Gramatyka formalna języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 1992.

[Świdziński, 1992b] Świdziński, M., 1992b. *Od interpretacji do faktów językowych: weryfikacja empiryczna gramatyki formalnej*. Rozprawy Uniwersytetu Warszawskiego, Warszawa 1992.

[Świdziński, 1993a] Świdziński, M., 1993a. *Od interpretacji do faktów językowych: weryfikacja empiryczna gramatyki formalnej*. Biuletyn Polskiego Towarzystwa Językoznawczego, zeszyt XLIX (1993), s. 15–24.

[Świdziński, 1993b] Świdziński, M., 1993b. *Schematy zdaniowe w słowniku i w tekście*. [w:] Prace z dejin slavistiky XVI. Ceske a polske srovnavaci studie. Praha.

[Świdziński, 1993c] Świdziński, M., 1993c. *Verb patterns in the Polish vocabulary and texts*. [w:] R. Köhler, B. Rieger (red.), Contributions to Quantitative Linguistics, Dordrecht – Boston – London.

[Świdziński, 1994a] Świdziński, M., 1994a. *Instrukcja opisu wypowiedników z korpusu „Słownika frekwencyjnego polszczyzny współczesnej"*. Warszawa 1994. Wydruk komputerowy powielony, stron 24 (tekst nie zachował się w całości — brak pierwszej strony).

[Świdziński, 1994b] Świdziński, M., 1994b. *Syntactic Dictionary of Polish Verbs*. Warszawa 1994.

[Świdziński, 1996] Świdziński, M., 1996. *Własności składniowe wypowiedników polskich*. Dom Wydawniczy Elipsa, ISBN 83-7151-178-7. Warszawa 1996.

[Świdziński, 1997] Świdziński, M., 1997. *Projekt badawczy nr P104 0 30 04 Ukierunkowana gramatycznie tekstowa baza danych: korpus wypowiedzeń współczesnej polszczyzny pisanej — raport końcowy*. 17 lutego 1997 r.

[Świdziński i Szpakowicz, 1989] Świdziński, M. i S. Szpakowicz, 1989. *Non-Typical Linguistic Phenomena in the Universal Basic Dictionary of Contemporary Polish*. University of Ottawa, 1989, s. 20–24.

[Świdziński i Szpakowicz, 1994] Świdziński, M. i S. Szpakowicz, 1994. *Sentence schemata in the Universal Basic Dictionary of Contemporary Polish*. International Journal of Lexicography vol. 7 No 1, Oxford, s. 1–30.

[TEIP4, 2001] TEIP4, 2001. *Sperberg-McQueen, C. M. i Burnard, L. (red.) TEI P4. Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition.*. Chicago, Oxford, 2001, 2002, 2004. The Association for Computers and the Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Lingustic Computing (ALLC). `http://www.tei-c.org/release/doc/tei-p4-doc/html/`.

[TEIP5, 2005] TEIP5, 2005. *Sperberg-McQueen, C. M. i Burnard, L. (red.) TEI P5. Guidelines for Electronic Text Encoding and Interchange. Revised and re-edited.*. Oxford – Providence – Charlottesville – Nancy 2005. The Association for Computers and the Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Lingustic Computing (ALLC). `http://www.tei-c.org/release/doc/tei-p5-doc/html/`.

[Tokarski, 2002] Tokarski, J., 2002. *Schematyczny indeks a tergo polskich form wyrazowych*. Opracowanie i redakcja: Zygmunt Saloni. Wydanie drugie. Wydawnictwo Naukowe PWN, Warszawa 2002.

[Wall i in., 2001] Wall, L., T. Christiansen, R. L. Schwartz, i S. Potter, 2001. *Perl — programowanie*. Warszawa 2001. ReadMe, wydanie drugie. ISBN: 83-7101-382-5.

[Wiśniewski, 1994] Wiśniewski, M., 1994. *Strukturalna charakterystyka polskich wypowiedzeń niezdaniowych*. Toruń 1994.

[Woliński, 2003] Woliński, M., 2003. *System znaczników morfosyntaktycznych w korpusie IPI PAN*. Polonica XXII–XXIII, s. 39–55. Kraków 2003. `http://dach.ipipan.waw.pl/CORPUS/znakowanie.pdf`.

[Woliński, 2004] Woliński, M., 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. Rozprawa doktorska. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2004. `http://www.ipipan.waw.pl/~wolinski/publ/mw-phd.pdf`.

[Woliński i Przepiórkowski, 2001] Woliński, M. i A. Przepiórkowski, 2001. *Projekt anotacji morfosyntaktycznej korpusu języka polskiego*. Prace IPI PAN 938. Instytut Podstaw Informatyki Polskiej Akademii Nauk. ISSN 0138-06-48. Warszawa, grudzień 2001. `http://www.ipipan.waw.pl/~wolinski/publ/ipi938.pdf`.

[XML, 2004] XML, 2004. *Extensible Markup Language (XML) 1.0 (Third Edition)*. World Wide Web Consortium Recommendation, 4 February 2004. `http://www.w3.org/TR/REC-xml/`.

Dodatek A

Charakterystyka opisowa i liczbowa korpusu wypowiedników

A.1 Szczegóły opisu wypowiedników

W tabeli prezentuję w skondensowanej formie szczegółowe informacje dotyczące opisu pojedynczego rekordu korpusu wypowiedników, spisane na podstawie pozycji [Świdziński, 1996] (s. 45–63). Do niej też odnoszą się wszystkie podane niżej numery punktów i stron.

Kod	Opis	Wartości parametrów
Lokalizatory		
ID	identyfikator wypowiednika	liczba z zakresu 1–6721
STYL	oznaczenie transzy korpusu słownika frekwencyjnego (pkt 5.6, s. 47)	`PU` — publicystyka `WP` — wiadomości prasowe `PO` — teksty popularnonaukowe `PR` — proza `DR` — dramat
PR	numer próbki (pkt 5.6, s. 47)	liczba z zakresu 1–2000 (w praktyce 1–1991)
WYP	numer wypowiedzenia w próbce (pkt 5.6, s. 47)	liczba z zakresu 1–22
ZD	numer wypowiednika w wypowiedzeniu (pkt 5.6, s. 47)	liczba z zakresu 1–16
Cechy jakościowe: dystrybucja		
TW	typ wypowiednika (pkt 5.7.2, s. 48)	`Z` – zdanie `D` – dopowiedzenie `R` – równoważnik `E` – elipsa `V` – wołacz `W` – wykrzyknienie `F` – fraza `C` – cytat obcy

Kod	Opis	Wartości parametrów `U` – konstrukcja urwana
WSP	koordynacja (pkt 5.7.2, s. 48)	`K` – wypowiednik złożony brak wartości – elementarny
ST	status (pkt 5.7.3, s. 48)	`S` – samodzielny `Wp` – współrzędny początkowy `Ws` – współrzędny środkowy `Wk` – współrzędny końcowy `P` – podrzędny `D` – dostawiony `R` – zdanie złożone–reszta
TYP	charakterystyka kontekstowa (pkt 5.7.4, s. 48)	dla wypowiedników samodzielnych: wartość pusta
		dla współrzędnych: nazwa spójnika–centrum
		dla względnych: wartość postaci `R: x`, `x` jest nazwą zaimka względnego
		dla pytajnozależnych: wartość postaci `PZ: x`, `x` to nazwa pytajnika
		dla przytoczeń: wartość `OR`
		dla wypowiedników pospójnikowych: nazwa spójnika podrzędnego
		dla wypowiedników dostawionych: wartość pusta, `DWUKR` (dwukropek), `PARENT` (parenteza) i inne
Cechy jakościowe: morfologia		
CEN	centrum — forma czasownika lub spójnik (pkt 5.8.1, s. 50)	forma spójnika, czasownika lub quasi-czasownika albo symbol `R` oznaczający zdanie złożone–resztę
HAS	hasło — postać hasłowa centrum (pkt 5.8.2, s. 50)	nazwa spójnika, leksemu czasownika lub quasi-czasownika, wartość pusta dla zdań złożonych–reszt
NEG	wystąpienie negacji centrum (pkt 5.8.3, s. 51)	wartość pusta dla wypowiedników nie będących zdaniem elementarnym `N` dla centrum postaci *nie v* `T` w pozostałych wypadkach
KL	klasa gramatyczna centrum (pkt 5.8.4, s. 51)	`V` — czasownik `Q` — quasi-czasownik `C` — spójnik

Kod	Opis	Wartości parametrów `Q(V)` — quasi-czasownik „składniowy”
ASP	aspekt (pkt 5.8.5, s. 51)	`i` — niedokonany `p` — dokonany
CHAR	charakterystyka fleksyjna centrum (pkt 5.8.6, s. 52)	czwórka wartości: osoba, liczba, rodzaj, czas lub tryb osoba: `1` — pierwsza `2` — druga `3` — trzecia
		liczba: `p` — pojedyncza `m` — mnoga
		rodzaj: `m` — męski w liczbie pojedynczej lub męskoosobowy w mnogiej `ż` — żenski `n` — nijaki `-m` — niemęskoosobowy
		czas: `pe` — przeszły `pe2` — przeszły drugi `te` — teraźniejszy `tpy` — przyszły prosty `py` — przyszły złożony
		tryb: `roz` — rozkazujący `war` — warunkowy `war2` — warunkowy drugi
		wartości specjalne: `BOS` — bezosobnik `BOK` — bezokolicznik
Cechy jakościowe: opis struktury		
SCH	schemat zdaniowy (pkt 5.9.1, s. 53, dodatek D)	liczba jedno-, dwu- lub trzycyfrowa; pierwsza cyfra odpowiada liczbie miejsc, kolejne — numerowi schematu
OPIS	charakterystyka frazy wymaganej (pkt 5.9.2, s. 55)	oddzielone znakiem plusa kody wymagań, ew. uzupełnione znakiem dolara dla realizacji frazeologizmu: fraza podmiotowa: symbol `m`
		fraza nominalna: symbol przypadka (pierwsza litera polskiej nazwy)

Kod	Opis	Wartości parametrów
		fraza przyimkowo-nominalna: przyimek, minus, przypadek
		fraza przymiotnikowa: symbol przypadka ze znakiem cudzysłowu
		fraza przyimkowo-przymiotnikowa: przyimek, minus, przypadek, cudzysłów
		fraza przysłówkowa: wartość `PS`
		fraza bezokolicznikowa: symbol `BOK` z ew. oznaczeniem `i` lub `p`, jeśli czasownik ogranicza aspekt wymagania
		fraza zdaniowa: oznaczenie typu frazy (nazwa spójnika lub `PZ`) i ew. korelatu (oddzielone znakiem hasza)
		fraza zdaniowa w pozycji podmiotu-mianownika — symbol `m`, a po nim oznaczenie typu frazy zdaniowej w nawiasach klamrowych
TPSU	typ frazy podmiotowej (pkt 5.9.3, s. 60)	`1` — pusty podmiot pierwszoosobowy `2` — pusty podmiot drugoosobowy `3w` — pusty podmiot trzecioosobowy mający koreferenta wewnątrz danego wypowiednika `3z` — pusty podmiot trzeciooosobowy nie mający koreferenta wewnątrz danego wypowiednika `BOK` — brak podmiotu w zdaniu z centrum bezokolicznikowym `BOS` — brak podmiotu w zdaniu z centrum bezosobnikowym `się` — podmiot będący formą zaimka *się*
TPI	typ członu innego (pkt 5.9.4, s. 61)	rodzaj oznajmienia (kody jak dla pola `TW`), symbol członu lub oznaczenie trzeciej frazy wymaganej (`OB3`) lub luźnej (`LU3`)
SZYK	porządek składników zdania (pkt 5.9.5, s. 62)	ciąg symboli `S`, `V` i `O`, ew. z indeksami w postaci apostrofów w liczbie zgodnej z numerem składnika nieciągłego

Kod	Opis	Wartości parametrów
Cechy ilościowe (liczby słów)		
DL	długość wypowiednika (pkt 5.10, s. 62)	liczba z zakresu 0–72
VF	długość frazy finitywnej (pkt 5.10, s. 62)	liczba z zakresu 0–59
SU	długość frazy podmiotowej (pkt 5.10, s. 62)	liczba z zakresu 0–43
OB1	długość pierwszej frazy wymaganej (pkt 5.10, s. 62)	liczba z zakresu 0–50
OB2	długość drugiej frazy wymaganej (pkt 5.10, s. 62)	liczba z zakresu 0–38
LU1	długość pierwszej frazy luźnej (pkt 5.10, s. 62)	liczba z zakresu 0–40
LU2	długość drugiej frazy luźnej (pkt 5.10, s. 62)	liczba z zakresu 0–34
IN	długość członu innego (pkt 5.10, s. 62)	liczba z zakresu 0–61
Zapis struktury		
TEKST	tekst wypowiednika	tekst przeplatany oznaczeniami składniowymi

A.2 Rozkład typów wypowiedników

Oto klasyfikacja typów wypowiedników występujących w korpusie:

Rodzaj	Liczba	Udział %
Wypowiedniki zdaniowe		
zdanie	5452	81,12%
Wypowiedniki oznajmieniowe		
elipsa	795	11,83%
wołacz	130	1,93%
wykrzyknienie	119	1,77%
fraza	110	1,64%
dopowiedzenie	58	0,86%
równoważnik	46	0,69%
cytat obcy	6	0,09%
konstrukcja urwana	5	0,07%

A.3 Rozkład stopnia złożoności wypowiedników

Oto statystyka wystąpień poszczególnych rodzajów wypowiedników:

Rodzaj	Liczba	Udział %
wypowiedniki samodzielne	3512	52,25%
wypowiedniki podrzędne	1223	18,20%
wypowiedniki współrzędne początkowe	701	10,43%
wypowiedniki współrzędne końcowe	700	10,41%
wypowiedniki dostawione	356	5,30%
zdania złożone – reszty	137	2,04%
wypowiedniki współrzędne środkowe	92	1,37%

Ciekawie wypada także ocena „stopnia złożoności" korpusu — stosunku liczby wypowiedników wchodzących w skład grup zbudowanych z wypowiednika nadrzędnego i wszystkich podrzędnych do łącznej liczby wypowiedników. Wypowiedników tworzących 1518 grup jest aż 4784, czyli ponad 71% łącznej liczby wypowiedników. Wypowiedniki złożone i ich składniki stanowią więc zdecydowaną większość korpusu.

A.4 Podsumowanie korelacji parametrów gramatycznych

Poniższa tabela podsumowuje rozkład liczby wypowiedników ze względu na ich składnikowość, obecność frazy finitywnej i związki ze strukturą nadrzędną:

Rodzaj	Typ	Liczba	Udział %
Wypowiedniki złożone			
zdanie	samodzielne	406	6,04%
	reszta	137	2,04%
	podrzędne	46	0,68%
	współrzędne końcowe	19	0,28%
	współrzędne początkowe	11	0,16%
	dostawione	8	0,12%
elipsa	samodzielna	55	0,82%
	współrzędna końcowa	4	0,06%
	podrzędna	3	0,04%
	dostawiona	2	0,03%
	współrzędna początkowa	1	0,01%
dopowiedzenie	samodzielne	5	0,07%
fraza	samodzielna	1	0,01%
wołacz	samodzielny	1	0,01%
wykrzyknienie	samodzielne	1	0,01%

Wypowiedniki elementarne			
zdanie	samodzielne	2372	35,29%
	podrzędne	1092	16,25%
	współrzędne początkowe	610	9,08%
	współrzędne końcowe	551	8,20%
	dostawione	119	1,77%
	współrzędne środkowe	81	1,21%
elipsa	samodzielna	486	7,23%
	współrzędna końcowa	78	1,16%
	podrzędna	75	1,12%
	współrzędna początkowa	70	1,04%
	dostawiona	15	0,22%
	współrzędna środkowa	6	0,09%
wołacz	dostawiony	108	1,61%
	samodzielny	17	0,25%
	współrzędny środkowy	2	0,03%
	współrzędny końcowy	1	0,01%
	współrzędny początkowy	1	0,01%
fraza	samodzielna	98	1,46%
	dostawiona	6	0,09%
	współrzędna końcowa	2	0,03%
	współrzędna początkowa	2	0,03%
	podrzędna	1	0,01%
wykrzyknienie	dostawione	80	1,19%
	samodzielne	35	0,52%
	współrzędne końcowe	1	0,01%
	współrzędne początkowe	1	0,01%
	podrzędne	1	0,01%
równoważnik	współrzędny końcowy	41	0,61%
	współrzędny środkowy	3	0,04%
	podrzędny	2	0,03%
dopowiedzenie	samodzielne	26	0,39%
	dostawione	17	0,25%
	współrzędne początkowe	5	0,07%
	współrzędne końcowe	3	0,04%
	podrzędne	2	0,03%
konstrukcja urwana	samodzielna	5	0,07%
cytat obcy	samodzielny	4	0,06%
	podrzędny	1	0,01%
	dostawiony	1	0,01%

Dodatek B

Parametry GFJP

Gramatyka będąca wynikiem tej pracy zawiera 508 reguł, 56 symboli nieterminalnych.

Poniżej wyjaśniam szczegółowo używaną notację oraz zamieszczam listy symboli nieterminalnych z oznaczeniem jednostek dodanych i ich parametrów z krótkimi opisami.

B.1 Szczegóły notacji

Reguły gramatyki zapisane są w notacji DCG:

```
fl1(A, C, Rl, O, Neg, I, z(SwZ, NZ))
--> s(lu7),
    fps(St, Neg, I, z(SwZ, Z), Kl, _, _),
    { zrozne(Z, ['byxx', 'choćby', 'czyżby',
                 'gdyby', 'jakby', 'jakoby', 'żeby'], NZ) }.
```

Strony reguły oddzielone są znakiem strzałki (`-->`). Jednostki składniowe reprezentowane są poprzez termy, których funktor stanowi nazwę jednostki, zaś argumenty wyznaczają parametry składniowe jednostki. Powiązanie między wartościami parametrów zapewniane jest poprzez mechanizm unifikacji Prologu. W treść reguły mogą także wchodzić dodatkowe warunki — akcje prologowe ujęte w nawiasy klamrowe.

Przytoczona reguła definiuje frazę luźną właściwą `fl1` jako frazę przysłówkową `fps` o określonych parametrach, przy czym zachodzi uzgodnienie wartości parametrów negacji `Neg` i inkorporacji `I`. Warunek nałożony na parametr zależności `z(SwZ, NZ)` wymusza spełnienie predykatu `zrozne`, który zachodzi, gdy wartość parametru `Z` nie znajduje się na liście będącej drugim argumentem funktora warunku.

B.2 Lista jednostek nieterminalnych

`agl(Rl, O, I)`
— aglutynant

`agl1(Rl, O)`
— aglutynant właściwy

`condaglt(L, O)`
— warunkowy morfem aglutynacyjny

`ff(Wf, A, C, T, Rl, O, Wym, K, Neg, I, Z, Ow)`
— fraza finitywna

`ff1(Wf, A, C, T, Rl, O, Wym, K, Neg, I, Z, Ow)`
— fraza finitywna właściwa

`fl(A, C, Rl, O, Neg, I, Z)`
— fraza luźna

`fl1(A, C, Rl, O, Neg, I, Z)`
— fraza luźna właściwa

`fno(P, Rl, O, Neg, I, Z, Kl, Typ)`
— fraza nominalna

`formaczas(Wf, A, C, T, Rl, O, Wym, K)`
— forma czasownikowa

`formaczas1(S, Wf, A, C, T, Rl, O, Wym, K)`
— forma czasownikowa właściwa

`formalicz(P, Rl, Ak)`
— forma liczebnikowa

`formalicz10(P, Rl, Ak)`
— forma liczebnikowa typu 10 (1–99)

`formalicz100(P, Rl, Ak)`
— forma liczebnikowa typu 100 (1–999)

`formaprzym(P, Rl, St)`
— forma przymiotnikowa

`formaprzym1(P, Rl, St)`
— forma przymiotnikowa właściwa

`formaprzysl(St)`
— forma przysłówkowa

`formarzecz(P, Rl)`
— forma rzeczownikowa

`fpm(Pm, P, Neg, I, Z, Kl, Typ)`
— fraza przyimkowa

`fps(St, Neg, I, Z, Kl, Typ)`
— fraza przysłówkowa

`fpt(P, Rl, St, Neg, I, Z, Kl, Typ)`
— fraza przymiotnikowa

`fw(Tfw, K, A, C, Rl, O, Neg, I, Z)`
— fraza wymagana

`fw1(Tfw, K, A, C, Rl, O, Neg, I, Z)`
— fraza wymagana właściwa

`fwe(Wf, A, C, T, Rl, O, Wym, K, Neg, I, Z)`
— fraza werbalna

`fzd(Tfz, K, A, C, T, Neg, I, Typ)`
— fraza zdaniowa

`fzde(Tfz, A, C, T, Neg, I)`
— fraza zdaniowa elementarna

`kor(K, I)`
— korelat

`kor1(P)`
— korelat właściwy

`kweink(Wf, A, C, T, Rl, O, Wym, K, I, Z)`
— konstrukcja werbalna z inkorporacją

`kweneg(Wf, A, C, T, Rl, O, Wym, K, Neg, I, Z)`
— konstrukcja werbalna z negacją

`kwer(Wf, A, C, T, Rl, O, Wym, K, Z)`
— konstrukcja werbalna

`kwer1(Wf, A, C, T, Rl, O, Wym, K, Z)`
— konstrukcja werbalna właściwa

`morfagl(F, Rl, O)`
— morfem aglutynacyjny

`partykula(F)`
— partykuła

`przec`
— przecinek

`przecsp`
— przecinek przedspójnikowy

`przyimek(Pm, P)`
— przyimek

`przyzlo(S, Rl, Wym, K)`
— składowa czasu przyszłego

`pyt(F, I)`
— pytajnik partykułowy

`spoj(Tsp, Oz, I)`
— spójnik

`spoj1(Tsp, Oz)`
— spójnik właściwy

`spojnik(Oz)`
— spójnik

`wypowiedzenie`
— wypowiedzenie

`zaimneg(Kgz, P, Rl, O, Kl)`
— zaimek negatywny

`zaimno(Kgz, P, Rl, O, Kl)`
— zaimek nieokreślony

`zaimos(P, Rl, O)`
— zaimek osobowy

`zaimprzym(F, P, Rl)`
— zaimek przymiotny

`zaimprzys(F)`
— zaimek przysłowny

`zaimpyt(Kgz, P, Rl, O, Kl)`
— zaimek pytajny

`zaimrzecz(H, P, Rl)`
— zaimek rzeczowny

`zaimwzg(Kgz, P, Rl, O, Kl)`
— zaimek względny

`zaimzwrot(P)`
— zaimek zwrotny

`zd(Wf, A, C, T, Rl, O, Neg, I, Z, Ow, Oz, Typ)`
— zdanie

`znakkonca(Z)`
— znak końca

B.3 Lista parametrów i ich wartości

Parametr		Wartość	Opis
`A`	(aspekt)	`dk` `nd`	dokonany niedokonany
`Ak`	(akomodacyjność)	`congr` `rec`	uzgadniająca nieuzgadniająca

`C`	(czas)	`ter`
		`prze`
		`przy`
			teraźniejszy
			przeszły
			przyszły

`F`	(forma/segment)	wartość słownikowa	
`H`	(identyfikator leksemu)	wartość słownikowa	

`I`	(inkorporacja)	`bowiem`
		`natomiast`
		`więc`
		`zaś`
			inkorporacyjna
			inkorporacyjna
			inkorporacyjna
			inkorporacyjna

Parametr		Wartość	Opis
		ni	nieinkorporacyjna
K	(korelatywność)	nk	niekorelatywna
		bp	bezprzecinkowa
Kgz	(klasa gramatyczna zaimka)	przym	przymiotna
		przysl	przysłowna
		rzecz	rzeczowna
Kl	(klasa)	co	zaimek typu co
		kto	zaimek typu kto
		os	zaimek osobowy
		przym	przymiotnik
		przysl	przysłówek
		rzecz	rzeczownik
		tk	wartość specjalna
		wz	zaimek względny
		zaim	zaimek
L	(liczba)	poj	pojedyncza
		mno	mnoga
Neg	(negacja)	tak	niezaprzeczona
		nie	zaprzeczona
		ani	zaprzeczona odspójnikowa
O	(osoba)	1	pierwsza
		2	druga
		3	trzecia
Ow	(ograniczenie wewnętrzne)	br	brak ograniczenia
		choćby	spójnikowa
		dopóki	spójnikowa
		gdy	spójnikowa
		gdyby	spójnikowa
		więc	
Oz	(oznaczenie spójnika)	wartość słownikowa	
P	(przypadek)	mian	mianownik
		dop	dopełniacz
		cel	celownik
		bier	biernik
		narz	narzędnik
		miej	miejscownik
		wol	wołacz
Pm	(przyimek)	wartość słownikowa	
R	(rodzaj)	m	męski (osobowy, żywotny lub nieżywotny)
		mos	męski osobowy
		mzw	męski zwierzęcy
		mżyw	męski żywotny (osobowy lub zwierzęcy)
		mnż	męski nieżywotny

Parametr	Wartość	Opis
	`mnos`	męski nieosobowy (żywotny lub nieżywotny)
	`mn`	męski (dowolny) lub nijaki
	`nmo`	nie męski osobowy (dowolny z wyjątkiem `mos`)
	`żeń`	żeński
	`nij`	nijaki
	`pt`	przymnogi
	`np`	nijaki lub przymnogi
`S` (wymaganie *się*)	`s`	*się* wymagane
	`n`	*się* niedopuszczalne
`St` (stopień)	`row`	równy
	`wyz`	wyższy
	`naj`	najwyższy
`T` (tryb)	`ozn`	oznajmujący
	`roz`	rozkazujący
	`war`	warunkowy
`Tfw` (typ frazy wymaganej)	`np(P)`	fraza nominalna w przypadku `P`
	`adjp(P)`	fraza przymiotnikowa w przypadku `P`
	`advp`	fraza przysłówkowa
	`infp(A)`	fraza czasownikowa w bezokoliczniku o aspekcie `A`
	`prepnp(Pm,P)`	fraza przyimkowa
	`sentp(Tfz)`	fraza zdaniowa typu `Tfz`
`Tfz` (typ frazy zdaniowej)	`aż1`	spójnikowa
	`aż2`	spójnikowa
	`bowiem`	spójnikowa
	`bo`	spójnikowa
	`chociaż`	spójnikowa
	`choćby`	spójnikowa
	`co`	względna
	`czy`	spójnikowa
	`dopóki`	spójnikowa
	`gdyby`	spójnikowa
	`gdy`	spójnikowa
	`jakby`	spójnikowa
	`jaki`	względna
	`jakoby`	spójnikowa
	`jak`	spójnikowa
	`jeśli`	spójnikowa
	`kto`	względna
	`który`	względna
	`mie1`	mieszana 1
	`mie2`	mieszana 2
	`mie3`	mieszana 3
	`podczas`	spójnikowa

Parametr	Wartość	Opis
	ponieważ	spójnikowa
	pz	pytajnozależna
	zanim	spójnikowa
	żeby	spójnikowa
	że	spójnikowa
Tsp (typ spójnika)	rc	równorzędny centralny
	rl	równorzędny lewy
	rp	równorzędny prawy
	ri	równorzędny inkorporacyjny
	sz	szeregowy
	szk	szeregowy końcowy
	po	podrzędny początkowy
	pc	podrzędny centralny
	pl	podrzędny lewy
	pp	podrzędny prawy
	pi	podrzędny inkorporacyjny
Wf (wyróżnik fleksyjny)	os	osobowy
	bos	bezosobowy
	bok	bezokolicznikowy
	psu	imiesłowowy uprzedni
	psw	imiesłowowy współczesny
Z (zależność)	aż1xx	typu aż z ograniczeniem czasu, aspektu i negacji
	aż1	spójnikowa
	aż2	spójnikowa
	bowiem	spójnikowa
	boxx	z ograniczeniem czasu, aspektu i negacji
	box	z ograniczeniem czasu
	bo	spójnikowa
	byxx	z ograniczeniem czasu, aspektu i negacji
	chociaż	spójnikowa
	choćby	spójnikowa
	co	względna
	czy	spójnikowa
	czyżby	kontekstowa
	dopóki	spójnikowa
	gdyby	spójnikowa
	gdyxx	z ograniczeniem czasu, aspektu i negacji
	gdy	spójnikowa
	jakby	spójnikowa
	jaki	względna
	jakoby	spójnikowa
	jak	spójnikowa

Parametr	Wartość	Opis
	jeśli	spójnikowa
	kto	względna
	który	względna
	npt	niepytajna z ograniczeniem trybu
	npxt	niepytajna z ograniczeniem czasu i trybu
	npxxt	niepytajna z ograniczeniem czasu, trybu, aspektu i negacji
	npxx	niepytajna z ograniczeniem czasu, aspektu i negacji
	npx	niepytajna z ograniczeniem czasu
	np	niepytajna
	podczas	spójnikowa
	ponieważ	spójnikowa
	pxx	pytajna z ograniczeniem czasu, aspektu i negacji
	px	pytajna z ograniczeniem czasu
	pz	pytajnozależna
	p	pytajna
	zanimxx	zanim z ograniczeniem czasu, aspektu i negacji
	zanim	spójnikowa
	żeby	spójnikowa
	że	spójnikowa

Dodatek C

Modyfikacje korpusu

Poniżej przedstawiam listę błędnych oznaczeń, która jest jednocześnie zestawem odesłań do zamieszczonego na dołączonej do pracy płycie zbiorczego wykazu poprawek dokonanych przeze mnie w korpusie wypowiedników na etapie weryfikacji typograficznej warstwy opisu korpusowego (patrz rozdział 5, s. 49).

C.1 Usterki typograficzne

Rodzaj błędu i przykład	**Lista wystąpień**
zbędny tekst (na końcu próbki, dopisek lub tekst opuszczony oznaczony jak fraza, informacje lokalizacyjne w treści próbki): „*0.*", „[*Urwane*]", „*291, 3-2 …, /gdzie/ (powstać) <mają>*"	2443, 2449, 2453, 2691, 3884, 4060, 6376, 6382, 6385, 6621
brak oznaczenia znaku nie należącego do wypowiednika (występującego w treści samodzielnie, a zaburzającego postać tekstową poddawaną analizie): „:", „#"	2882–83, 5012, 5123, 5326–27, 5645, 5657–58, 5660, 5726, 5773, 5800, 5836, 5839, 5863, 5911–13, 6003, 6022, 6027, 6037, 6082, 6100, 6128, 6186, 6216, 6229, 6239–40, 6482, 6532, 6645
błędna postać wielokropka: „*sumienia..*", „(*do Petersburga*),...."	90, 188, 342, 995, 1380, 1874, 1880, 1882, 1888–89, 2214, 2298, 2400, 2541, 2546, 2567, 2637, 2890, 3271, 3273, 3276, 3277, 3385, 3665, 3884, 4056, 4083, 4125, 4219–20, 4288, 4584–86, 4588, 4668, 4921, 5268, 5357, 5623, 5711, 6243, 6283, 6488, 6490, 6533

Rodzaj błędu i przykład	**Lista wystąpień**
znaki interpunkcyjne nie należące do frazy znajdujące się w jej wnętrzu: „*panu...)*”, „[*Helenko, Helenko,*]”	18, 30, 55, 172, 226, 324, 428, 553, 570, 589, 648, 671, 733, 885, 1059, 1086, 1089, 1092, 1101, 1107, 1157, 1179, 1185, 1194, 1331, 1365, 1883, 1893, 2154, 2157, 2214, 2236, 2383, 2501, 2502, 2541, 2564, 2599, 2650, 2685, 2758, 2856, 2865, 2970, 3043, 3084, 3388, 3544, 3598, 3599, 3632, 3635, 3678, 3737, 3772, 3788, 3877, 3910, 3932, 3934, 3943, 4008, 4099, 4143, 4165, 4184, 4217, 4228, 4282, 4330, 4335, 4428, 4456, 4482, 4565, 4957, 5046, 5059, 5102, 5178, 5230, 5293, 5305, 5346, 5426, 5483, 5496, 5588, 5609, 5700, 5719, 5779, 5918, 5982, 5994, 6030, 6075, 6101, 6279, 6304, 6358, 6542, 6546, 6554, 6598, 6638, 6681, 6686, 6688, 6699
brak lub błędny znak interpunkcyjny na końcu wypowiednika: „*(Nic) <nie wiesz>*”, „*biżuterii*],”	22, 24, 27, 28, 45, 46, 56, 57, 72, 77, 80, 81, 82, 90, 185, 428, 573, 591, 626, 723, 786, 848, 861, 863, 943, 945, 950, 957, 958, 959, 963, 964, 965, 966, 968, 969, 971, 972, 975, 976, 977, 981, 982, 986, 988, 990, 991, 992, 997, 998, 1000, 1001, 1002, 1003, 1013, 1014, 1015, 1016, 1019, 1020, 1029, 1031, 1046, 1048, 1049, 1054, 1056, 1058, 1061, 1063, 1070, 1075, 1076, 1228, 1426, 2017, 2019, 2020, 2025, 2026, 2031, 2032, 2034, 2039, 2040, 2043, 2046, 2048, 2051, 2052, 2053, 2054, 2060, 2061, 2065, 2066, 2068, 2070, 2073, 2080, 2085, 2088, 2090, 2091, 2093, 2095, 2098, 2102, 2106, 2107, 2110, 2111, 2116, 2117, 3088, 3089, 3094, 3098, 3102, 3103, 3104, 3105, 3112, 3214, 3430, 3643, 3721, 4342, 6401
brak spacji — po znaku interpunkcyjnym (oddzielającej go od tekstu lub granicy frazy), oddzielającej numer frazy od jej treści, rozdzielającej frazy): „*...,=że=*”, „*ziemskiej 1), <nazywamy>*”, „*(1miastu*”, „*sekundę2)*”, „*opodatkowania 2/(prowadzić...)*”	5, 18, 77, 166, 168, 172, 173, 188, 208, 216, 331, 370, 430, 437, 461, 707, 947, 952, 970, 996, 1005–08, 1009, 1011, 1034–35, 1053, 1057, 1068, 1168, 1240, 1261, 1273, 1276, 1442, 2022, 2033, 2036, 2042, 2047, 2071, 2081, 2089, 2120, 2123, 2124, 2151, 2353, 2426, 2519, 2611, 2719, 2766, 2769, 2848, 2874, 3071, 3099, 3100, 3110, 3267, 3273, 3286, 3288, 3290, 3293, 3294, 3296, 3296, 3349, 3353, 3372, 3418, 3478, 3528, 3648, 3690, 3698, 3789, 3799, 3806, 3809, 3823, 3870, 3976, 3995, 4083, 4119, 4204, 4212, 4299, 4343, 4843, 5405, 5432, 5607, 5625, 5703, 5783, 5819, 6044, 6115

Rodzaj błędu i przykład	Lista wystąpień
przecinki należące do frazy znajdujące się poza jej obrębem: „[*że jeszcze nie czas na ich poruszanie*]”	2236, 3433, 6015
nadmiarowa spacja (przed znakiem interpunkcyjnym, po łączniku, między znakiem granicy frazy a jej numerem lub tekstem): „*ciała*) .”, „*afro- azjatyckiego*”, „(*1 żadnego związku 1*)”, „(*mnie*”	18, 243, 354, 425, 669, 892, 1175, 1181, 1796, 1855, 1898, 2117, 2245, 2284, 2285, 2429, 2544, 2650, 3018, 3285, 3385, 3438, 3445, 3510, 3575, 3596, 3598, 3598, 3623, 3629, 3629, 3635, 3649, 3652, 3676, 3678, 3728, 3729, 3765, 3765, 3771, 3783, 3794, 3801, 3803, 3884, 3919, 3924, 3933, 3987, 3989, 3998, 4010, 4034, 4069, 4080, 4099, 4100, 4108, 4117, 4118, 4124, 4129, 4135, 4558, 4560, 4792, 4962, 5007, 5217, 5229, 5239, 5268, 5293, 5305, 5346, 5609, 5972, 5994, 6030, 6031, 6278, 6279, 6304, 6305, 6354, 6368, 6376, 6423, 6467, 6531, 6542, 6554, 6575, 6583, 6591, 6598, 6635, 6638, 6661, 6681, 6686, 6688, 6699, 6711
błędnie umieszczona spacja: „<*Wydaje się*>, [*iż*”	171, 553, 1168, 3341, 3429, 3468, 3510, 3699, 3716, 3727, 3786, 3907, 3924, 4067, 4100, 4180, 4288, 4310, 4335, 6023, 6102, 6104, 6167, 6447

C.2 Błędne oznaczenia elementów frazowych

Większość błędów w strukturze fraz wykryłem podczas tworzenia gramatyki korpusowej omówionej w rozdziale 4.4, s. 43.

Rodzaj błędu i przykład	Lista wystąpień
znak nie należący do wypowiednika oznaczony jako fraza: „(,)”, „[...]”	1085, 1088–89, 1091–92, 1124, 1125, 1143–44, 1170–71, 1936–37, 2441, 2625, 2760, 2763, 2871, 2893, 2902, 2906, 3215, 3218, 3220, 3436, 3442, 3746, 3763, 3813, 3895, 3980–81, 4173, 4390–91, 5407–08, 5278, 5627, 5968
błędne oznaczenie granic frazy (odwrócony lub błędny nawias, brak nawiasu, nawias nadmiarowy): „>*nie ma*>”, „(*gazetę*>”	457, 515, 526, 644, 1478, 1626, 2366, 2691, 2949, 3028, 3438, 3891, 3955, 3988, 4287, 4555, 4570, 4757, 4784, 4843, 4853, 5521, 5552, 5632, 5677, 5716, 5718, 6030, 6045, 6048, 6308, 6111, 6299–6300, 6530
brak lub niepełne oznaczenie nieciągłości frazy: „<*byście...*><*chcieli*>”	2630, 2772, 2833, 3019, 3077, 3078, 3417, 3709, 5447

Rodzaj błędu i przykład	Lista wystąpień
błędna numeracja frazy (brakujące oznaczenie numeru wymaganego, nadmiarowy numer frazy pojedynczej, niezgodność numerycznych oznaczeń początku i końca frazy, nieciągłość numeracji fraz, błędne oznaczenie frazy nieciągłej): „(*2 miast lub innych mieszkań ludzkich*)”, „(*1 od którego 2*)”, „(*1' przecięte 1*)”	171, 1217, 1549, 2794, 3071, 3437, 3575, 3652, 3730, 3786, 3792, 3907, 4489, 5369, 5768, 6308
błędna klasyfikacja fraz: „<*całkiem cicho*>”	86, 1219, 2037, 2286, 2833, 3004, 3417, 5131, 5355, 5394, 5558, 5602, 6621, 6665

C.3 Błędy w opisie parametrów próbek

Rodzaj błędu i przykład	Lista wystąpień
błędna klasyfikacja wypowiednika (`Z` zamiast `E`)	86, 1264
błędne oznaczenie schematu (`V.1` bez kategoryzacji frazy wymaganej)	2102

Rodzaj błędu i przykład	Lista wystąpień
niezgodność wartości parametru `OPIS` ze strukturą zadaną w tekście wypowiednika:	161, 171, 212, 280, 281, 282, 283, 427, 489, 495, 534, 554, 635, 672, 694, 803, 804, 805, 849, 851, 872, 899, 913, 1004, 1025, 1033, 1103, 1104, 1106, 1108, 1110, 1112, 1149, 1167, 1171, 1173, 1176, 1183, 1193, 1230, 1239, 1261, 1290, 1294, 1303, 1391, 1474, 1509, 1659, 1660, 1667, 1698, 1727, 1729, 1783, 1806, 1856, 1859, 1887, 1892, 1903, 1954, 1955, 1958, 1960, 1962, 1973, 2033, 2046, 2124, 2206, 2247, 2262, 2367, 2373, 2381, 2395, 2487, 2496, 2509, 2533, 2534, 2594, 2607, 2649, 2694, 2782, 2783, 2788, 2796, 2801, 2804, 2821, 2838, 3067, 3068, 3118, 3212, 3223, 3226, 3242, 3361, 3363, 3394, 3395, 3434, 3527, 3530, 3546, 3569, 3571, 3624, 3665, 4042, 4074, 4077, 4125, 4131, 4199, 4200, 4221, 4222, 4290, 4326, 4353, 4455, 4588, 4620, 4659, 4667, 4724, 4726, 4781, 4796, 4797, 4824, 4842, 4843, 4846, 4847, 4883, 4895, 4919, 4976, 4989, 5016, 5102, 5129, 5131, 5210, 5211, 5261, 5262, 5315, 5334, 5335, 5380, 5388, 5456, 5519, 5533, 5598, 5605, 5640, 5691, 5798, 5828, 5987, 6040, 6041, 6060, 6067, 6089, 6090, 6150, 6172, 6212, 6286, 6328, 6362, 6380, 6502, 6529, 6692, 6700

C.4 Niezgodność tekstu w wypowiednikach zależnych

Rodzaj błędu i przykład	Lista wystąpień
zamieniona kolejność wyrazów: „*ma się*” i „*się ma*”	2009
wypowiednik urwany: „*<Myślałem>* (*zechce pan*”	51, 6130, 6132
brakujący fragment tekstu: „*praojcze, dziewięćkroć pra*” i „*praojcze, po dziewięćkroć pra*”	101–102

C.5 Usterki „morfologiczne”

Rodzaj błędu i przykład	Lista wystąpień
literówki: „*cztelnikowi*”, „*blogosławić*”, „*luudzi*”	6, 8–9, 31, 76, 129, 158, 218, 310, 326, 352, 363, 365, 366, 367, 368, 371, 375–376, 391, 392, 393, 395, 396, 412, 414, 460, 553, 586, 590, 594, 647–649, 655, 657, 688, 728–729, 764, 772, 774, 776, 784, 793, 815–816, 897, 939, 963, 1001, 1162, 1217, 1219, 1274, 1283, 1284, 1291, 1371, 1439, 1481, 1500, 1508, 1521, 1534, 1579, 1592, 1626, 1633, 1636, 1637, 1648, 1717, 1765, 1781–82, 1785, 1790, 1858, 2020, 2051, 2095, 2124, 2128, 2150, 2196, 2212–13, 2220, 2230, 2297, 2351, 2376–77, 2395, 2484, 2486, 2479, 2512–13, 2515, 2567, 2572, 2573, 2579, 2630, 2668–69, 2786, 2824, 2843, 2852, 2881, 3116–17, 3190, 3280, 3348, 3354, 3377, 3399, 3416, 3457, 3473, 3520, 3528, 3529, 3532, 3573, 3585, 3598, 3628, 3652, 3660, 3664, 3665, 3666, 3680, 3684, 3686, 3687, 3688, 3695, 3699, 3711, 3729, 3731, 3731, 3736, 3737, 3741, 3760–61, 3782–83, 3802, 3812, 3832, 3882–83, 3886, 3907, 3925–26, 3931, 3948, 3979, 3981, 3983, 3987–88, 4017, 4079, 4149, 4230, 4239, 4254, 4277, 4286, 4305, 4306, 4307, 4377, 4389–91, 4409, 4412, 4447, 4449, 4482, 4487, 4492, 4838–39, 5042, 5158, 5276, 5481, 5486, 5509, 5567, 5581, 5591, 5612–14, 5661–64, 5718, 5723, 5725–26, 5730, 5742–44, 5747, 5770–71, 5780–81, 5802, 5948, 5955, 5975, 5996–97, 5999, 6025, 6033, 6043–44, 6127, 6133, 6149, 6181, 6184, 6201, 6220–22, 6244–45, 6290, 6292–94, 6303–07, 6317–18, 6326, 6332, 6362, 6364, 6369, 6372, 6375, 6382, 6401, 6443, 6449, 6455–57, 6466, 6469–70, 6501, 6553, 6606–07, 6648, 6649, 6659, 6686
użycie małej litery w miejscu wielkiej: „*życie Gospodarcze*”	6106
sklejone wyrazy: „*roślinęwodną*”	593, 715, 1889, 3364, 3416, 3430, 3483, 3701, 3960, 3974, 4204

C.6 Usterki „składniowe”

W bieżącym rozdziale zebrałem informację o poprawkach o charakterze ściśle składniowym, tj. umożliwiających akceptację wypowiednika przez gramatykę Świdzińskiego. Zmiany te nie wykraczają jednak poza poprawki interpunkcyjne — jak zakładam, dodane przecinki czy znaki zapytania były zgodne z intencją autora wypowiedzi, natomiast zostały przeoczone w procesie obróbki danych.

W kilku przypadkach próbki zawierające brakujące/nadmiarowe znaki zostały skorygowane po porównaniu ich z wypowiednikami zależnymi.

Rodzaj błędu i przykład	**Lista wystąpień**
błędne oznaczenie wyrazu nie należącego do wypowiednika składowego: „(*że*)”, „*ale*\”	712, 799, 968, 977, 1108, 1110, 1121, 1130, 1174, 1183, 1184, 1194, 1298, 1299, 1854, 1855, 1856, 1858, 1858, 1860, 1861, 2723, 2876, 2878, 2884, 2903, 2918, 2921, 2924, 2925, 2947, 2954, 2969, 2972, 2974, 2981, 2982, 2986, 2989, 2995, 2998, 3013, 3022, 3032, 3037, 3042, 3044–45, 3056, 3061, 3064 3068, 3194, 3213, 3225–26, 3359, 3375, 3434, 3455, 3489, 3509, 3520, 3554, 3578, 3604, 3623, 3630, 3632, 3690, 3697, 3702, 3720, 3824, 3825, 3828, 3831, 3834, 3841, 3856, 3880, 3916, 3924, 3967, 3985, 4000, 4094, 4125, 4133, 4145, 4239, 4244, 4264, 4267, 4283, 4289, 4308, 4312, 4347, 4370, 4402, 4405, 4405, 4420, 4422, 4438, 4443, 4459, 4477, 4497, 5822, 5955
brak wymaganej interpunkcji lub interpunkcja inna niż wymagana: „<*wiedziałem*> (*że o tym wiesz*)”	24, 57, 81, 82, 385, 432, 577, 945, 970, 993, 1006, 1016, 1023, 1068, 1463, 2020, 2032, 2034, 2040, 2080, 2095, 2117, 2118, 2122, 2159, 2162, 2595–96, 2830, 3088, 3099, 3983, 4144–45, 4201, 4485, 5229, 5533, 5772
nadmiarowy znak przestankowy: „*panny służące, czy szwaczki*”, „*Wallen, i Zurychskie*”	547, 553, 1506, 3792

Dodatek D

Płyta CD „Świgra Live”

Świgra Live[1] to udostępnione na dołączonej płycie CD kompletne środowisko analizy składniowej dla systemu Microsoft Windows. Stanowi ono pierwszą próbę przeniesienia składników Świgry z systemu linuksowego pod Windows, dokonane przeze mnie za zgodą Marcina Wolińskiego, autora głównego komponentu płyty — analizatora składniowego Świgra.

Po włożeniu płyty do napędu środowisko zgłasza się wyświetlając stronę startową:

Rysunek D.1: Ekran wyświetlany po włożeniu płyty do napędu

[1]Nazwa nawiązuje do powszechnie używanego określenia *Live* na oznaczenie płyt CD z oprogramowaniem tematycznym, gotowych do uruchomienia po włożeniu do napędu, często bez potrzeby dodatkowej instalacji.

W skład środowiska wchodzą:

- zestaw stron WWW ze skrótową informacją o korpusie wypowiedników, gramatyce Świdzińskiego oraz analizatorze Świgra, a także instrukcją korzystania z mechanizmu analizy składniowej (jej uproszczoną wersję zamieszczam poniżej),
- rozszerzona wersja gramatyki formalnej języka polskiego będąca wynikiem pracy,
- używany do przetwarzania danych analizator składniowy Świgra autorstwa Marcina Wolińskiego,
- wersja korpusu wypowiedników w formacie oryginalnym,
- XML-owy korpus wypowiedników wraz z mechanizmem do jego przeglądania,
- dodatkowe narzędzia, wyniki prac i testów.

D.1 Instrukcja korzystania ze środowiska analizy składniowej

Uruchomienie środowiska analizy składniowej w celu pracy interakcyjnej jest możliwe bezpośrednio z płyty CD, z wykorzystaniem linku umieszczonego na zgłaszającej się automatycznie stronie głównej (*Uruchom środowisko analizy* u dołu okna). Po jego kliknięciu zostanie wyświetlone okne dialogowe z prośbą o potwierdzenie, w którym należy wybrać polecenie *Otwórz*. Po uruchomieniu interpretera Prologu z załadowanym automatycznie programem analizatora Świgra można już przystąpić do analizy wpisując w oknie predykat `analiza` z argumentem w postaci treści analizowanego wypowiedzenia, np.

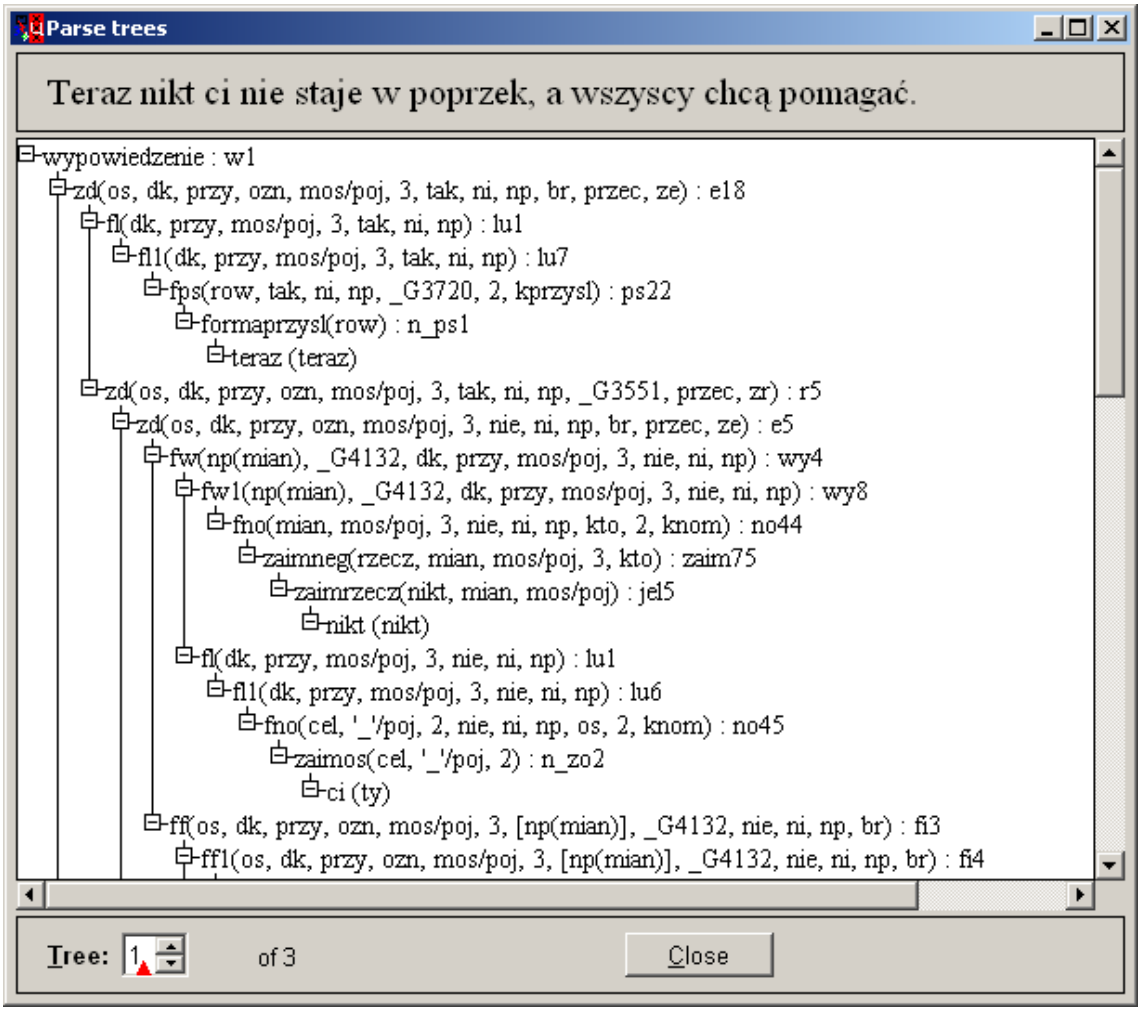
```
analiza('Teraz nikt ci nie staje w poprzek, a wszyscy chcą pomagać.').
```

Po naciśnięciu klawisza `Enter` rozpocznie się analiza składniowa, której wynikiem jest zestaw drzew rozbioru wyświetlany w osobnym oknie, jak na rys. D.2.

Wyświetlanie kolejnych drzew jest możliwe przy użyciu przycisków strzałek umieszczonych w lewym dolnym rogu okna.

Główne okno interpretera zawierać będzie informację łącznej liczbie drzew (linia zawierająca napis `trees`), liczbie kroków wywodu podczas analizy właściwej (`parse_inferences`) i przetwarzania jej wyników (`process_inferences`) oraz czasie analizy (odpowiednio `parse_cputime` i `process_cputime`):

```
parse_inferences: 456023
parse_cputime: 0.180259
edges: 436
trees: 3
process_inferences: 59055
process_cputime: 0.0901296
```



Rysunek D.2: Okno z wynikiem analizy składniowej w postaci drzew rozbioru

W przypadku braku drzew rozbioru dodatkowe okno pozostanie puste, a główne okno interpretera wyświetli liczbę drzew równą zeru i tekst `Porażka`.

Rozpoczęcie nowego procesu analizy następuje po wydaniu kolejnego polecenia z użyciem predykatu `analiza`.

D.2 Rozszerzona wersja gramatyki

Rozszerzona wersja gramatyki zamieszczona została w katalogu `swigra/swigra` i ma postać dwóch plików: `gfjp.dcg`, odpowiadającego zasadniczej części GFJP i `gfjp-elem.dcg`, zawierającego definicję jednostek funkcyjnych. Użycie gramatyki jako części programu analizy składniowej wymaga jej przetłumaczenia na postać klauzulową gotową do użycia przez Świgrę załączonym programem autorstwa Marcina Wolińskiego (`swigra/swigra/gengfjppl.bat`); wynikiem jego działania jest plik `gfjp.pl`, którego wygenerowana wersja została także umieszczona w katalogu.

D.3 Morfeusz i Świgra

W katalogu `swigra` zamieściłem analizator składniowy Świgra Marcina Wolińskiego gotowy do uruchomienia w środowisku Windows z wykorzystaniem udostępnianego na licencji LGPL[2] interpretera SWI-Prolog[3] Częścią edytora Świgra jest analizator morfologiczny Morfeusz[4], także autorstwa Marcina Wolińskiego, tu udostępniony w postaci biblioteki DLL (wersja dla Windows z 14 sierpnia 2005 roku).

D.4 Narzędzia do przetwarzania korpusu wypowiedników

Powstałe w wyniku opracowania korpusu wypowiedników narzędzia, które mogą także zostać wykorzystane do przetwarzania i masowej analizy składniowej innych korpusów zostały umieszczone w katalogu `narzedzia`. Większość z nich ma postać programów w języku Perl oraz odpowiadających im skryptów wsadowych interpretera poleceń Windows ilustrujących sposób wywołania. Bardziej złożone programy rozpoczyna komentarz opisujący funkcję narzędzia i jego parametry.

W podkatalogu `wypowiedniki` znalazły się narzędzia do przetwarzania korpusu wypowiedników w formacie opisanym w rozdziale 2.3 (s. 28):

- główny program do analizy składniowej wypowiedników z możliwością wyboru gramatyki, formatu wyniku, włączania modułu do sortowania drzew i ograniczania czasu analizy,
- program realizujący cząstkową analizę składniową wypowiedników z wykrywaniem reguły początkowej, której wyniki opisane zostały w rozdziale 8.1.5 (s. 81),
- program przenoszący strukturę składniową wypowiedników elementarnych na poziom wypowiedników złożonych,
- program tworzący słownik wymagań składniowych czasowników (patrz rozdział 4.2 na s. 42),
- program wypisujący grupy wypowiedników pochodzące od wspólnego wypowiednika złożonego,
- programy sczytujące elementy frazowe danego typu.

W podkatalogu `swigra` zamieściłem narzędzia do przetwarzania wyników analizy otrzymanych po uruchomieniu Świgry, tj.:

- konwerter wyników do formatu tekstowego, HTML-owego i XML-owego wraz z generatorem drzew w formacie PDF,
- program do zapisu pojedynczego drzewa w formacie wykorzystującym komponent `dtree` (patrz rozdział D.5 na s. 209) oraz program generujący postać drzew umieszczoną na płycie (z wykorzystaniem arkusza stylów XSL),

[2]Patrz `http://www.gnu.org/copyleft/lesser.html`.

[3]Patrz `http://www.swi-prolog.org`; warunki licencyjne udostępniania składników środowiska opisane są na stronie `http://www.swi-prolog.org/license.html`.

[4]Patrz `http://nlp.ipipan.waw.pl/~wolinski/morfeusz/morfeusz.html`; zasady licencji zawiera plik `http://nlp.ipipan.waw.pl/~wolinski/morfeusz/morfeusz-licencja.pdf`.

- program eliminujący frazy luźne, którego wyniki działania opisuję w rozdziale 13.3.2 (s. 164),
- programy zliczające i wypisujące drzewa izomorficzne (patrz rozdział 13.3.1 na s. 162),
- program do sortowania drzew rozbioru (patrz rozdział 12.3 na s. 148),
- narzędzia do masowego testu morfologicznego dowolnych zbiorów danych z wykorzystaniem analizatora Morfeusz.

D.5 XML-owy korpus wypowiedzeń

Korpus wypowiedników w formacie XML-owym (patrz rozdział 3, s. 31) ma postać bazy analiz składniowych, której rekordem jest pełne drzewo analizy danego wypowiednika z uwzględnieniem opisu grup (jak pamiętamy, w korpusie Świdzińskiego składowe danego wypowiednika nadrzędnego opisane są w rekordach zależnych, w korpusie XML-owym informacja ta jest dostępna już w opisie wypowiednika złożonego).

Dane korpusu XML-owego znajdują się w katalogu `korpus`. Pliki z drzewami rozbioru mają nazwy postaci `czterocyfrowy_numer_wypowiednika.xml` (np. `0123.xml`).

Drzewa analizy zapisane w formacie XML przetwarzane są napisanym przeze mnie arkuszem stylów XSLT[5] (`drzewo.xsl` w katalogu `xsl`) do postaci akceptowanej przez zewnętrzny, służący do ich wizualizacji na stronie HTML-owej darmowy moduł języka Javascript `dtree` autorstwa Geira Landrö[6].

Oczywiście, zaprezentowana wizualizacja drzewa rozbioru jest jednym z wielu możliwych sposobów prezentacji korpusowej informacji składniowej. Użycie XML-a umożliwia zastosowanie innego sposobu prezentacji poprzez prostą podmianę użytego arkusza stylów XSLT.

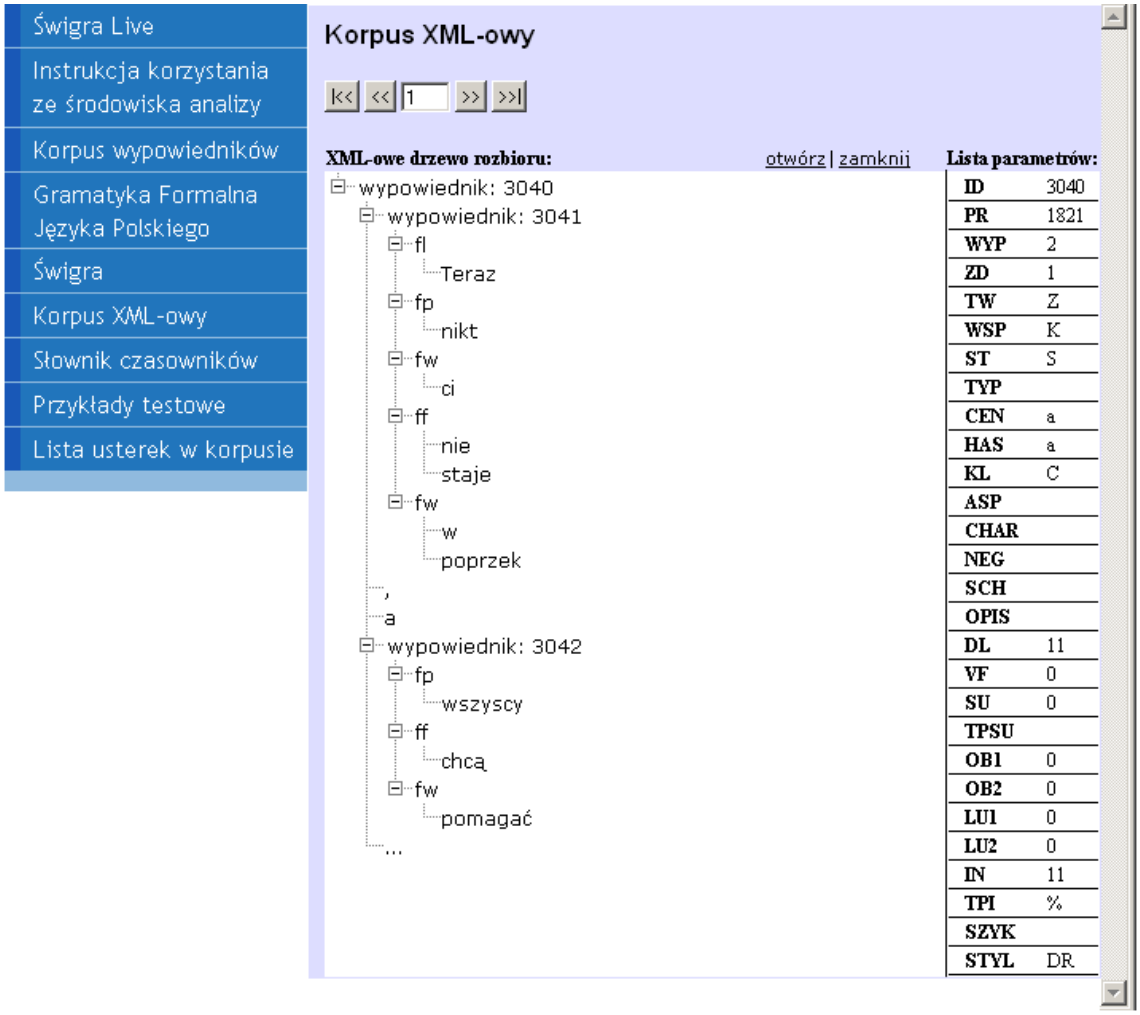
Użyty format danych (zestaw znaczników odpowiadających jednostkom analizy składniowej) został skonstruowany w sposób umożliwiający jego wykorzystanie do porównawczej prezentacji drzew rozbioru ręcznego i uzyskanych automatycznie, np. przy użyciu analizatora składniowego Świgra. Sukcesem zakończył się eksperyment polegający na zapisie drzewa automatycznej analizy składniowej w zaproponowanym wyżej formacie.

Postać XML-ową drzew otrzymywałem poprzez przekształcenie prologowej reprezentacji wyników Świgry napisanym przeze mnie programem w języku Perl[7]. Repertuar elementów XML-owych zawiera 57 jednostek, od reprezentującej cały analizowany tekst jednostki <`wypowiedzenie`> aż po jednostkę terminalną <`term`>. Przykład uzyskanego drzewa rozbioru prezentuje rys. D.4 na s. 211 (wariant wizualizacji bez

[5]XSLT jest opartym na XML-u językiem programowania umożliwiającym przetwarzanie dokumentów XML-owych, tu wykorzystanym do konwersji danych korpusu na format HTML wyświetlający drzewo rozbioru w przeglądarce internetowej. Zdecydowanie najlepsza na rynku pozycja o tym standardzie to [Kay, 2005].

[6]Patrz `http://www.destroydrop.com/javascripts/tree/`.

[7]Spis narzędzi zawiera rozdział D.4 na s. 208.



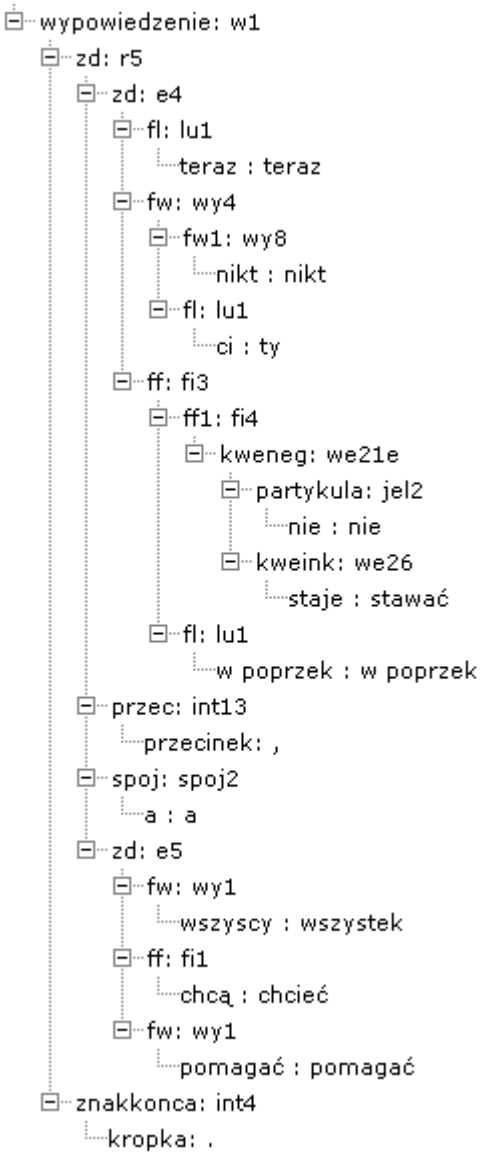
Rysunek D.3: Zrzut ekranu strony prezentującej korpus XML-owy

kompletu parametrów składniowych; dla ograniczenia wysokości drzewa liście zostały bezpośrednio powiązane z najwyższą jednostką nadrzędną występującą przed rozgałęzieniem).

Poziom drzewa odpowiada poziomowi analizy, tj. regule gramatyki. Węzłami drzewa są zatem, jak u Wolińskiego, stanowiące nagłówki użytych reguł jednostki nieterminalne gramatyki oraz jednostki terminalne — liście drzewa rozbioru. Zestaw jednostek terminalnych (i niektórych nieterminalnych, gdyż reguły najniższych poziomów nie są przez Świdzińskiego definiowane) także przejmuję od Wolińskiego, wykorzystując bezpośrednio wyniki analizatora Świgra. Ma to znaczenie dla „kształtu" liści drzew rozbioru, ponieważ Woliński wprowadza swoistą segmentację wypowiedzenia na jednostki poziomu niższego niż poszczególne słowa, wydzielając np. pseudoimiesłowy i cząstki aglutynacyjne[8]. W oryginalnym korpusie wypowiedników poza znakami interpunkcyjnymi jednostkami terminalnymi są zawsze słowa.

Zamieszczenie na płycie stworzonego w wyżej opisany sposób korpusu porównawczego zawierającego zarówno drzewa rozbioru ręcznego, jak też drzewa uzyskane

[8]Patrz [Woliński, 2004], s. 50–52.



Rysunek D.4: Drzewo analizy automatycznej

drogą analizy automatycznej nie wydawało się celowe ze względu na bardzo dużą niekiedy liczbę wyników uzyskiwanych z wykorzystaniem Świgry; użytkownik płyty może porównać efekt analizy wykorzystując udostępnione środowisko interpretera Prologu.

D.6 Pozostałe materiały

Katalog `tekst` zawiera pełny tekst niniejszej pracy w formacie PDF i źródłowym.

W katalogu `dodatki` zamieszczony został słownik wymagań czasownikowych (patrz rozdział 4.2, s. 42) w formacie tekstowym. Jest on dostępny w dwóch postaciach — oryginalnej (plik `sscp-orig.txt`), zgodnej z opisem z pracy [Świdziński, 1994b] oraz

wyekstrahowanej automatycznie z korpusu wypowiedników (w zapisie zbliżonym do oryginalnego — plik `sscp-ekstr.txt`). Te i wszystkie pozostałe pliki tekstowe na płycie zostały zapisane w kodowaniu Windows (`cp-1250`).

W katalogu tym znajdują się także pliki tekstowe `gfjp-a.txt` i `gfjp-b.txt` z testami analizatorów składniowych stworzonymi na potrzeby projektu *Zestaw testów do weryfikacji i oceny analizatorów języka polskiego*. Plik pierwszy zawiera 660 przykładów (515 poprawnych i 145 niepoprawnych) wyekstrahowanych z aneksu do książki, plik drugi, z przykładami z właściwej treści książki — 1376 zdań (1054 poprawnych, 296 niepoprawnych i 26 wątpliwych).

W podobnym formacie zapisane zostały stworzone na użytek tej pracy pliki `testy-licz.txt`, `testy-gno.txt` i `testy-gno2.txt` zawierające odpowiednio przykładowe zdania do testów konstrukcji liczebnikowych (patrz rozdział 10.1.3, s. 108) i konstrukcji z równorzędną grupą nominalną (w dwóch wersjach — pełnej i okrojonej, patrz rozdział 10.2.1, s. 109).

Ostatnim zbiorem w katalogu jest plik `oznajmienia.txt` z listą poddanych analizie oznajmień (patrz rozdział 12.5.3, s. 155) w postaci tabeli LaTeX-a.

W katalogu `swigra` zamieszczone zostały pliki analizatora składniowego Świgra w wersji minimalnej (m. in. bez narzędzi umożliwiających generowanie drzew TEX-owych). Dwa najważniejsze z nich to `gfjp.dcg` i `gfjp-elem.dcg` zawierające zmodyfikowaną w wyniku niniejszej pracy definicję jednostek gramatyki Świdzińskiego w formacie wykorzystywanym przez Świgrę; inne ciekawe zasoby to słownik wymagań czasownikowych w wersji dla Świgry (`slowczas.pl`) i plik zawierający zmodyfikowane analizy morfologiczne (`morfeusz.pl`).