MARCIN WOLIŃSKI

AUTOMATYCZNA ANALIZA SKŁADNIKOWA JĘZYKA POLSKIEGO

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

MARCIN WOLIŃSKI

AUTOMATYCZNA ANALIZA SKŁADNIKOWA JĘZYKA POLSKIEGO

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

Autor: Marcin Woliński
ORCID: 0000-0002-7498-1484
Instytut Podstaw Informatyki Polskiej Akademii Nauk
Jana Kazimierza 5, 01-248 Warszawa

Recenzenci wydawniczy: prof. dr hab. Zygmunt Saloni
prof. dr hab. Marek Świdziński

Redaktor prowadzący: Karolina Kozakowska

Korekta: Monika Szewczyk

Projekt okładki: Elżbieta Chojna

Skład i łamanie: Marcin Woliński

Publikacja dofinansowana przez
Instytut Podstaw Informatyki Polskiej Akademii Nauk

ISBN 978-83-235-3606-2 (druk) ISBN 978-83-235-3614-7 (pdf online)

Wydawnictwa Uniwersytetu Warszawskiego
00-497 Warszawa, ul. Nowy Świat 4
e-mail: wuw@uw.edu.pl
księgarnia internetowa http://www.wuw.pl

Wydanie 1, Warszawa 2019

Spis treści

Wstęp

Celem niniejszej pracy jest przedstawienie formalnego opisu składni obszernego podzbioru języka polskiego, który byłby przydatny w zadaniach przetwarzania języka naturalnego. Opis ten został zaimplementowany w postaci automatycznego analizatora składniowego Świgra 2, a jego weryfikację stanowi korpus składniowy Składnica.

Gramatyka formalna opisuje pewien zbiór zdań – język formalny. Zastosowana do opisu języka naturalnego może wskazywać, które zdania są poprawne, a które nie. Jednak celem stworzenia gramatyki dla języka naturalnego nie jest uzyskiwanie odpowiedzi binarnych. Dużo istotniejsze jest to, że gramatyka jawnie lub niejawnie przypisuje wypowiedzeniu pewną strukturę, która ma je reprezentować. Językoznawca w istocie myśli o konstrukcjach składniowych za pomocą tych struktur. W kontekście metod komputerowych reprezentacja struktury składniowej stanowi także dane wejściowe dla dalszych etapów przetwarzania, przede wszystkim do stworzenia reprezentacji semantycznej.

Ponieważ polszczyzna jest językiem fleksyjnym, analizę składniową wypowiedzeń trzeba poprzedzić analizą fleksyjną (do której konieczny jest opis fleksji, czyli odmiany wyrazów). Rozdzielenie opisu na etapy ułatwia pracę, a w wypadku implementacji komputerowej ma też dodatkowe uzasadnienie techniczne. Dla zapewnienia efektywności przetwarzania warto stosować możliwie najprostsze środki, jako że z rosnącą siłą formalizmów rośnie też złożoność obliczeniowa. Dlatego do opisu fleksji warto zastosować efektywne techniki związane z automatami skończonymi, podczas gdy do opisu składni potrzebne jest zastosowanie formalizmu o większej sile wyrazu.

Przedstawiony tu opis abstrahuje od semantyki, a więc uwzględnia tyle ze struktury języka, ile da się opisać poprzez interakcje cech formalnych, a nie znaczeń. Jego przedmiotem jest „gra kształtów", a nie „gra znaczeń".

Jest to opis języka ogólnego w wariancie pisanym, z naciskiem na jego staranną, redagowaną odmianę. Celem nie jest jednak formułowanie zaleceń poprawnościowych, w szczególności wyłapywanie wypowiedzeń niepoprawnych, lecz opisanie jak największej liczby konstrukcji faktycznie pojawiających się w tekstach. Z tego punktu widzenia opłaca się opisywać niektóre konstrukcje niepoprawne (dotyczy to w szczególności sposobu używania przecinków przez typowych użytkowników języka).

Niniejsza książka mieści się w nurcie prac nad opisem fleksyjnym i składniowym języka polskiego, w które autor jest zaangażowany od kilkunastu lat. Wcześniejszym etapem tych prac była implementacja gramatyki formalnej GFJP Marka Świdzińskiego (1992). Jej wynikiem był automatyczny analizator składniowy Świgra 1. Dzięki niemu udało się pokazać, że opis Świdzińskiego ma spory poziom spójności. Jednak, mimo że gramatyka ta jest bardzo rozbudowana, oparty na niej analizator akceptuje niewielki odsetek zdań polskich (około 30%).

Dlatego powstała koncepcja rozwinięcia tego opisu gramatycznego, aby osiągnąć większy odsetek zdań poprawnie analizowanych, i opracowania za jego pomocą korpusu oznakowanego informacją składniową, czyli tzw. banku drzew (ang. *treebank*). Przedmiotem pracy Woliński (2004) była możliwie wierna implementacja GFJP, natomiast tematem niniejszych rozważań jest przedstawienie nowej gramatyki wolnej od pewnych niedoskonałości tamtej. W sensie technicznym gramatyka ta jest w całości napisana na nowo. Za kształt poszczególnych reguł i zastosowane rozwiązania techniczne odpowiada autor niniejszej pracy. Co więcej, niektóre zasady opisu zostały wyraźnie zmienione w stosunku do GFJP. W ten sposób narodziła się Świgra 2[1].

Opracowany automatyczny analizator składniowy Świgra 2 jest wystarczająco sprawny, aby było możliwe przetwarzanie dziesiątek tysięcy wypowiedzeń. Pozwoliło to zbudować korpus składniowy Składnica. Świgra 2 była używana także w innych pracach z dziedziny inżynierii lingwistycznej. Łukasz Dębowski stosował ją w pracach dotyczących automatycznej ekstrakcji schematów walencyjnych czasowników (Dębowski i Woliński 2007). Elżbieta Hajnicz wykorzystywała analizę składniową do hipotetyzowania ram semantycznych (Hajnicz 2011). Gramatyka Świgra 2 stała się też podstawą gramatyki POLFIE (Patejuk 2015).

Określenie „analiza składnikowa języka polskiego" użyte w tytule tej pracy jest bardzo ogólne. Jednak każde przedsięwzięcie mające na celu opis języka naturalnego jest w jakiś sposób ograniczone. Przedstawiony tu opis można postrzegać jako pewien etap na drodze do celu: z pewnością z czasem pojawi się kolejny opis, reprezentujący wyższy poziom dokładności. Warto też podkreślić, że różne partie przedstawianego materiału są przemyślane w różnym stopniu. Opis fleksyjny jest wynikiem długiej ewolucji i można go traktować jako w znacznym stopniu zweryfikowany zarówno z punktu widzenia leksykografa (por. Saloni *et al.* 2015), jak i gramatyka. Niektóre elementy opisu składni

[1] W powstałych na wcześniejszym etapie rozwoju nowej gramatyki wspólnych pracach z Markiem Świdzińskim posługiwaliśmy się nazwą GFJP2. Nowa gramatyka formalna, stanowiąca część analizatora Świgra 2, jest jednak na tyle różna od GFJP, że nazwa ta jest myląca. Nowa gramatyka zasługuje na nową nazwę. Niestety nazwa Świgra 2 zdążyła zacząć funkcjonować w szerszym obiegu i na jej zmianę jest za późno. W związku z tym na przedstawiony tu opis będę używał określenia *gramatyka Świgra 2*. Z technicznego punktu widzenia gramatykę tę można utożsamić ze zbiorem reguł, których używa analizator Świgra 2.

przedstawione w rozdziale 2 są zupełnie nowe i stanowią dopiero pierwsze przybliżenie opisu formalnego.

Zakładanym odbiorcą książki jest informatyk zainteresowany technikami przetwarzania języka naturalnego. Jej autor nie jest z wykształcenia językoznawcą, dlatego dość szczegółowo wprowadza zastosowany system pojęć z zakresu fleksji i składni oraz ilustruje tekst wieloma przykładami językowymi, które powinny pomóc czytelnikowi w intuicyjnym uchwyceniu znaczenia poszczególnych pojęć. Osoby, które odbiorą to jako nadmiar oczywistych przykładów, proszone są o wyrozumiałość. Być może książka zainteresuje także językoznawców tym, że pokazuje, jak informatyk widzi pojęcia językoznawcze oraz jaki poziom ścisłości jest konieczny, aby opis „działał" jako program komputerowy.

Sposób sformułowania reguł gramatycznych zastosowany w tej pracy można nazwać inżynierskim, nie jest on głównym jej przedmiotem. Ważniejszą sprawą jest pokazanie proponowanego opisu gramatycznego. Książka stanowi przez to dokumentację struktur zawartych w korpusie Składnica, które można traktować jako niezależne od programu komputerowego, który je wygenerował.

Struktura książki

Rozdział 1 przedstawia przyjęte zasady powierzchniowego dystrybucyjnego opisu fleksji języka polskiego. Opis ten wywodzi się z koncepcji Zygmunta Saloniego, zwłaszcza jego klasyfikacji leksemów polskich. Istotnym aspektem niniejszej pracy jest pokazanie, że opis ten może być dobrą podstawą opisu składni. Do tego rozdziału mogą sięgnąć użytkownicy analizatora fleksyjnego Morfeusz 2 SGJP, aby poznać szczegóły zastosowanego systemu znakowania i stojące za nimi motywacje.

Tematem rozdziału 2 są struktury składniowe (drzewa) przypisywane wypowiedzeniom polskim przez omawianą tu gramatykę. Rozdział zawiera systematyzację opisanych konstrukcji składniowych i pokazuje, co zostało objęte opisem. Mogą doń sięgnąć osoby zainteresowane dalszym przetwarzaniem struktur składniowych generowanych przez analizator Świgra 2, w szczególności przetwarzaniem danych z korpusu składniowego Składnica.

W rozdziale 3 przedstawiono wykorzystywany w analizie automatycznej słownik walencyjny Walenty. O ile poprzedni rozdział dotyczy głównie systematycznych własności składniowych (przysługujących leksemom należącym do dużych klas wyrazów, np. klas gramatycznych), to słownik walencyjny notuje własności składniowe uwarunkowane leksykalnie, a więc charakterystyczne dla poszczególnych leksemów.

Rozdział 4 poświęcony jest implementacji gramatyki. Przedstawiono w nim istotne rozszerzenie formalizmu Definite Clause Grammar (DCG), a następnie omówiono sposób użycia go do realizacji gramatyki. Zaprezentowano też

kilka mechanizmów analizatora Świgra 2, przede wszystkim mechanizm realizowania wymagań składniowych.

Celem rozdziału 5 jest umieszczenie przedstawionego tu opisu na tle innych opisów składniowych języka polskiego. Porównanie dotyczy zarówno warstwy językoznawczej, jak i technicznych aspektów sposobu wyrażenia poszczególnych opisów.

W rozdziale 6 zaprezentowano korpus składniowy Składnica, stanowiący przykład wdrożenia przedstawionego tu opisu do zanalizowania pewnego zbioru tekstów. Korpus składniowy można traktować jako dokumentację adekwatności opisu, ponieważ zawarte w nim struktury składniowe zostały wygenerowane automatycznie za pomocą analizatora Świgra 2, a następnie ujednoznacznione i zweryfikowane przez ekspertów.

Tematem rozdziału 7 są techniki statystycznego ujednoznaczniania analiz składniowych. Program komputerowy, trenowany na danych korpusu składniowego, ma za zadanie wykonać ujednoznacznienie podobnie, jak to robili eksperci budujący korpus. Uzupełnienie analizatora regułowego o taki moduł pozwala zbliżyć się do ideału, czyli sytuacji, w której program komputerowy wskazuje dla danego wypowiedzenia dokładnie jedną strukturę składniową.

Konwencje notacyjne

Przytaczane przykłady wypowiedzeń i ich fragmenty (w szczególności wykładniki tekstowe form wyrazowych) składane są kursywą, np.:

(1) *Książka ukazała się w odpowiednim momencie, w okresie dyskusji nad nowymi programami.* [Skł.]

(2) **Książka ukazała się odpowiednim momentem.*

Gwiazdka, jak w przykładzie (2), sygnalizuje niepoprawność gramatyczną przytoczonego wypowiedzenia. Przy numerowanych przykładach w nawiasach kwadratowych podawane jest źródło:

[Skł.] – wypowiedzenie z korpusu Składnica (por. p. 6.1),

[NKJP1M] – wypowiedzenie z ręcznie znakowanego podkorpusu Narodowego Korpusu Języka Polskiego (NKJP, nkjp.pl) o wielkości jednego miliona segmentów,

[NJKP300] – wypowiedzenie ze zrównoważonego wariantu NKJP (300 milionów segmentów),

[NKJP1800] – wypowiedzenie z pełnego NKJP (1 800 milionów segmentów),

[Walenty] – przykład ilustrujący schemat walencyjny cytowany za słownikiem Walenty (zob. rozdz. 3).

Brak oznaczenia sygnalizuje przykład własny (skonstruowany).

Identyfikatory leksemów podawane są kapitalikami (CZYTAĆ, WARSZAWA). Definicje stosowanych pojęć z dziedziny fleksji są przytoczone w początkowych punktach rozdziału 1, a składni – rozdziału 2.

Podziękowania

Za zainteresowanie mnie problemami przetwarzania języka naturalnego jestem wdzięczny prof. Januszowi S. Bieniowi, który był animatorem wielu prac wspomnianych w tej książce. Profesorowi Zygmuntowi Saloniemu jestem niezmiernie zobowiązany za to, że pokazał mi, że możliwe jest rygorystyczne podejście do opisu fleksji, jak również nauczył mnie, że wszystkie pojęcia, które próbujemy przypasować do rzeczywistości językowej, to tylko modele, a nie część tej rzeczywistości. Profesorowi Markowi Świdzińskiemu zawdzięczam podróż w krainę składni. Niniejsza książka nie powstałaby bez bardzo wielu godzin dyskusji z profesorem Świdzińskim, a przede wszystkim bez Jego gramatyki jako punktu odniesienia. Dziękuję również wszystkim członkom Zespołu Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki za wsparcie mnie w wysiłku nad tą pracą.

Szczególne podziękowania należą się czytelnikom jej fragmentów: Markowi Świdzińskiemu, Łukaszowi Dębowskiemu, Elżbiecie Hajnicz, Witoldowi Kierasiowi, Alinie Wróblewskiej i wreszcie Tomaszowi Obrębskiemu, który jako pierwszy przeczytał całość, a jego wnikliwe uwagi pozwoliły usunąć wiele niedociągnięć tekstu.

1

Analiza fleksyjna

Polszczyzna należy do języków o bogatej odmianie wyrazów[1]. Słownik języka polskiego notujący 200 000 haseł można uznać za obszerny. Hasłom takiego słownika odpowiada jednak kilka milionów różnych realizacji pojawiających się w tekstach. Co więcej, polski jest językiem fleksyjnym, co oznacza, że końcówki fleksyjne, stanowiące podstawowy środek tworzenia form odmiany, kumulują kilka funkcji gramatycznych. Sprawia to, że tworzenie form jest dużo mniej regularne niż w innych językach syntetycznych (językach aglutynacyjnych), gdzie każdy element doklejany do tworzonej formy pełni jedną funkcję gramatyczną. Dlatego pierwszą przeszkodą, którą trzeba pokonać przy komputerowej analizie tekstu polskiego, jest wykonanie analizy fleksyjnej. Jej celem jest powiązanie pewnych ciągów znaków tekstu wejściowego z abstrakcyjnymi jednostkami językowymi opisywanymi w słownikach języka polskiego. W największym uproszczeniu chodzi o rozpoznanie, że na przykład pojawiające się w tekście ciągi *psem* i *psach* odnoszą się do tej samej jednostki językowej – leksemu rzeczownikowego PIES, a ciąg *psim* jest tekstową realizacją innej jednostki – przymiotnika PSI. Wymaga to określenia zasad wyróżniania poszczególnych jednostek i klas jednostek.

Jednocześnie zadaniem opisu fleksyjnego jest ustalenie takiego repertuaru kategorii gramatycznych i ich wartości, aby przypisując je poszczególnym wyrazom występującym w tekście, można było odpowiednio precyzyjnie opisać ich funkcję składniową (ich uwarunkowania składniowe). Tak więc na przykład uznaje się, że formy rzeczownikowe są klasyfikowane m.in. ze względu na przypadek gramatyczny, aby dało się scharakteryzować różnicę gramatyczną między formą *psem* (narzędnik) i *psu* (celownik), przekładającą się na różne dopuszczalne konteksty użycia tych form. Dobranie odpowiednich kategorii gramatycznych i zbiorów ich wartości pozwala skatalogować wszystkie możliwe realizacje tekstowe form, a więc zestawić słownik fleksyjny.

Przedstawione tu podejście do fleksji jest powierzchowne. Jego celem jest tylko opracowanie klasyfikacji form, która nie musi objaśniać mechanizmów ich tworzenia (wyróżniania morfemów, opisu ich interakcji, uwarunkowań ortograficznych itd.).

---

[1] Określeń „wyraz" i „konstrukcja" będę używać nieformalnie, gdy bardziej chodzi o odwołanie do intuicji niż rygorystycznie zdefiniowanego pojęcia.

W tym rozdziale zostanie przedstawiona stosowana w niniejszej pracy koncepcja opisu fleksyjnego polszczyzny, która została zaimplementowana w analizatorze i generatorze fleksyjnym Morfeusz 2 (por. p. 1.3). Była ona dopracowywana stopniowo. Jej zrąb powstał na potrzeby znakowania Korpusu IPI PAN (Woliński i Przepiórkowski 2001; Woliński 2003; Przepiórkowski 2003b), a dalsza ewolucja była związana z analizatorem Morfeusz 2 (Woliński 2014). Koncepcja ta oparta jest przede wszystkim na pracach zespołu skupionego wokół Zygmunta Saloniego, których zwieńczeniem jest opracowanie *Słownika gramatycznego języka polskiego* (Saloni *et al.* 2007b, 2012, 2015, dalej: SGJP)[2].

Opracowanie systemu znakowania fleksyjnego, który byłby użyteczny w zastosowaniach komputerowych, wymaga podjęcia szeregu decyzji dotyczących stopnia szczegółowości opisu i poziomu wierności językoznawczej. Opis stworzony 10 lat temu dawał preferencję językoznawczej elegancji (np. kwestia rodzaju gramatycznego i sposób opisu form czasu przeszłego, zob. dalej). Obecnie prezentowana wersja zawiera kompromisy zwiększające jego komputerową użyteczność. Postanowiono nie wprowadzać zbyt szeroko rozróżnień, które dotyczą tylko niewielkich grup form lub tylko niewielu leksemów (konsekwencje tego założenia widać na przykład w opisie deprecjatywności, liczebników zbiorowych, form poprzyimkowych zaimków osobowych).

Przedstawiana koncepcja powstała w celu wykorzystania co najmniej na dwa sposoby: jako system znakowania korpusu, który byłby użyteczny dla jego użytkowników poprzez zapewnienie odpowiedniej szczegółowości wyszukiwania form fleksyjnych, oraz jako dane używane w dalszym przetwarzaniu, przede wszystkim w analizie składniowej.

1.1. Podstawowe pojęcia

Jednym z celów analizy fleksyjnej jest pogrupowanie wyrazów występujących w tekście w abstrakcyjne jednostki języka. Jednostki te nazywa się *leksemami* (wyrazami słownikowymi, por. Saloni 1974a). Pojęcie to nie jest absolutne, lecz związane z poziomem szczegółowości opisu przyjętego w danym słowniku. W słowniku notującym znaczenia za osobne leksemy mogą być uważane jednostki różniące się znaczeniem. W SGJP przyjęto, że kryterium wyróżnienia leksemów są tylko różnice własności fleksyjnych, a pomocniczo – własności składniowych. Leksem jest jednostką abstrakcyjną, tak więc leksemy nie występują w tekstach; są wynikiem interpretacji jako elementy wykoncypowanego systemu językowego. Z każdym leksemem jest związany identyfikator, nazywany *lematem* (por. p. 1.9).

[2] Prezentacja w tym rozdziale wykorzystuje wcześniejsze prace autora, w szczególności artykuł Woliński (2003). Zawiera jednak niepublikowane wcześniej rozstrzygnięcia wprowadzone ostatnio, w szczególności dotyczące reprezentacji rodzaju gramatycznego, zob. p. 1.6.3.

Zapis tekstów przyjęty dla języka polskiego pozwala w naturalny sposób wyróżnić *słowa*: najdłuższe ciągi znaków niezawierające odstępów ani znaków interpunkcyjnych (co nie znaczy, że żadnych znaków nieliterowych). Leksemy najczęściej przejawiają się w tekstach w postaci słów, na przykład słowo *chomikiem* można uznać za realizację leksemu CHOMIK. Jednak niekiedy przedmiotem interpretacji fleksyjnej powinny być fragmenty tak wyróżnionych słów. Wyrazistym przykładem jest słowo *gdybyście* występujące w wypowiedzeniu:

(1) *Wiedzielibyście, gdybyście słuchali.*

Próba interpretacji tego słowa w całości prowadziłaby zapewne do uznania go za spójnik podrzędny, jako że wprowadza ono zdanie podrzędne *gdybyście słuchali*. Jednocześnie widać uwarunkowanie osobowo-liczbowe: słowo *gdybyście* może wystąpić razem ze *słuchali*, ale nie ze *słuchał* ani *słuchaliśmy*. Fakty te sugerują wprowadzenie bytu bardzo nietypowego: spójnika odmiennego przez osobę i liczbę (por. Świdziński 1981). Jednak prostszym rozwiązaniem jest uznanie, że osobnej interpretacji powinny podlegać fragmenty tego słowa *gdyby* i *ście*. Element podlegający interpretacji fleksyjnej będzie dalej nazywany *segmentem* (w terminologii angielskiej w takim znaczeniu jest zwykle używany wyraz *token*). Tak więc ciąg znaków *chomikiem* jest jednocześnie słowem i segmentem, a ciąg *gdybyście* jest słowem składającym się z segmentów *gdyby* i *ście* (zagadnienie wyróżniania segmentów rozwinięto w p. 1.5).

Segmenty są niezinterpretowanymi ciągami znaków, leksemy są abstrakcyjnymi jednostkami języka. Bytem, który wiąże te dwa poziomy, są formy fleksyjne. *Forma fleksyjna* to interpretacja segmentu przypisująca go do konkretnego leksemu i opisująca jego własności gramatyczne poprzez określenie wartości pewnych kategorii gramatycznych (por. p. 1.6). Na przykład segment *chomikiem* może zostać zinterpretowany jako forma fleksyjna rzeczownikowego leksemu CHOMIK o wartości narzędnika kategorii przypadka i pojedynczej wartości liczby.

Inaczej mówiąc, segmenty są tekstowymi wykładnikami (realizacjami) form fleksyjnych (stosując skrót myślowy, są wykładnikami leksemów).

Technicznie formę fleksyjną można uznać za trójkę uporządkowaną ⟨*segment, lemat, znacznik fleksyjny*⟩, gdzie *znacznik fleksyjny* (ang. *tag*) stanowi zwarty zapis cech gramatycznych form (zob. p. 1.10). Zbiór wszystkich form fleksyjnych danego leksemu nazywa się jego *paradygmatem (fleksyjnym)*.

Analiza fleksyjna[3] polega na wskazaniu dla danego segmentu wszystkich form wszystkich leksemów, których może on być wykładnikiem. W procesie

[3] W środowisku komputerowego przetwarzania języka bywa też równoważnie używany termin *analiza morfologiczna*. Jednak w językoznawstwie termin ten obejmuje badanie budowy wewnętrznej wyrazów, w szczególności określanie ich elementarnych składników – morfemów, co nie ma miejsca w przedstawionym tu opisie. Ponadto niektórzy badacze do analizy morfologicznej (w odróżnieniu od fleksyjnej) włączają też opis derywacji.

We wcześniejszych pracach, np. Woliński i Przepiórkowski 2001, stosowane jest określenie *analiza morfosyntaktyczna* dla podkreślenia, że część podawanych cech ma charakter

Tabela 1.1. Przykład wyników analizy fleksyjnej tekstu *Mam próbkę analizy fleksyjnej.* wykonanej przez program Morfeusz 2 SGJP

		segment	lemat	znacznik fleksyjny
0	1	*Mam*	MAMA MAMIĆ MIEĆ	subst:pl:gen:f impt:sg:sec:imperf fin:sg:pri:imperf
1	2	*próbkę*	PRÓBKA	subst:sg:acc:f
2	3	*analizy*	ANALIZA	subst:sg:gen:f subst:pl:nom.acc.voc:f
3	4	*fleksyjnej*	FLEKSYJNY	adj:sg:gen.dat.loc:f:pos
4	5	.	.	interp

tym nie uwzględnia się kontekstu, w którym wystąpił dany segment, wyniki są więc często niejednoznaczne.

Ujednoznacznianiem fleksyjnym nazywa się określanie na podstawie kontekstu, jaką formę fleksyjną realizuje dane wystąpienie segmentu. Następujące po sobie analizę i ujednoznacznianie fleksyjne nazywa się *tagowaniem* (ang. *tagging*).

Celem *hasłowania* (*lematyzacji*) jest wskazanie dla danego segmentu leksemu, którego formy jest on wykładnikiem. Jest to więc tagowanie ograniczone tylko do części informacji o formach – do lematów. Przybliżone hasłowanie polegające na odcięciu od słów części zmieniającej się przy odmianie bywa nazywane *stemowaniem* (ang. *stemming*). Metoda ta ma sens w odniesieniu do języków o ograniczonej fleksji, ale dla języka polskiego daje wyniki wysoce niezadowalające.

Operacją odwrotną do analizy fleksyjnej jest *synteza fleksyjna* – utworzenie wszystkich form odmiany leksemu na podstawie jego lematu.

W celu zilustrowania wymienionych pojęć w tabeli 1.1 przedstawiono przykład wyników analizy fleksyjnej. Każdy wiersz tabeli zawiera jedną formę fleksyjną, kreski oddzielają grupy interpretacji dla poszczególnych segmentów. Segmentowi *mam* zostały przypisane trzy interpretacje: jako forma liczby mnogiej rzeczownika MAMA, jako forma trybu rozkazującego czasownika MAMIĆ i wreszcie jako forma czasu teraźniejszego czasownika MIEĆ. Segment *analizy* został jednoznacznie przypisany do lematu ANALIZA, może on jednak być interpretowany zarówno jako forma liczby pojedynczej, jak i mnogiej – w różnych przypadkach.

---

składniowy, a nie czysto fleksyjny. Dotyczy to na przykład podawania rodzaju dla rzeczowników, skoro nie odmieniają się one przez rodzaj. Określenie to jest mylące, bo sugeruje analizę, która obejmuje morfologię i składnię.

Dlatego w niniejszej pracy postanowiłem pozostać przy odrobinę za wąskim terminie *analiza fleksyjna*.

Wielość interpretacji dla jednego segmentu jest w języku polskim zjawiskiem bardzo częstym, warto więc wprowadzić dwa pojęcia opisujące takie sytuacje (por. Świdziński *et al.* 2002). *Homonimia* to równość wykładników form należących do różnych leksemów. *Synkretyzm* to równość wykładników różnych form należących do tego samego leksemu[4]. W przytoczonym przykładzie segment *Mam* może być wykładnikiem homonimicznych form trzech różnych leksemów. Synkretyczne są różne formy leksemu ANALIZA o wykładniku *analizy*.

Stosowane znaczniki fleksyjne są pozycyjne. Pierwsza pozycja określa klasę gramatyczną (część mowy), następne reprezentują wartości kategorii gramatycznych przysługujących danej klasie. Na przykład znacznik subst oznacza rzeczownik, a po nim następują wartości liczby, przypadka i rodzaju. Oznaczenia są w większości skrótami łacińskich nazw wartości. Konstrukcja przyjętych kategorii gramatycznych i dopuszczalne wartości poszczególnych kategorii zostały omówione w dalszych podrozdziałach.

1.2. SGJP

Prezentowany tu system analizy fleksyjnej stanowi przystosowaną do potrzeb analizy komputerowej postać opisu fleksji w *Słowniku gramatycznym języka polskiego* (Saloni *et al.* 2007b, 2012, 2015).

Ideę opracowania słownika pokazującego odmianę polskich wyrazów i cechy gramatyczne ich form powziął Zygmunt Saloni pod wpływem analogicznego słownika języka rosyjskiego (Zalizniak 1977). Realizacja tego zamierzenia trwała około 30 lat, w czasie których Saloni wraz z nieformalnym zespołem, obejmującym pracowników kilku placówek naukowych, zajmował się opracowaniem poszczególnych części materiału.

Prace rozpoczęły się od analizy opisów gramatycznych największego dostępnego wówczas słownika języka polskiego (Doroszewski 1958–1969). Informacja zawarta w tych opisach pozwalała tylko na przybliżone ustalenie sposobu odmiany, była jednak ważnym punktem wyjścia. W końcu lat 70. prace były prowadzone na papierowych fiszkach (ok. 130 000 sztuk, por. Saloni *et al.* 2007a). Użytecznym materiałem był także indeks *a tergo* do słownika Doroszewskiego (Grzegorczykowa i Puzynina 1973). Indeks ten został później zdigitalizowany przez Roberta Wołosza.

Istotnym osiągnięciem było opracowanie przez Włodzimierza Gruszczyńskiego opisu odmiany (deklinacji) polskich rzeczowników pospolitych (Gruszczyński 1989).

---

[4] Niektórzy badacze obejmują pojęciem homonimii oba wymienione zjawiska, wyróżniając homonimię międzyparadygmatyczną (=homonimia) i wewnątrzparadygmatyczną (=synkretyzm).

Pierwszym efektem prac grupy obejmującym materiał fleksyjny jako całość był przygotowany do publikacji przez Saloniego schematyczny indeks form fleksyjnych zainicjowany przez Jana Tokarskiego (Tokarski 1993). W dziele tym zakończenia form odmiany zostały powiązane z zakończeniami form podstawowych. System fleksyjny potraktowano jako potencję, a więc jako opis mechanizmów, a nie odmiany konkretnych leksemów. Indeks wykorzystano do konstrukcji kilku programów komputerowych (Szafran 1993; Wołosz 2005; Woliński 2006b). Indeks był opracowywany w postaci elektronicznej – pliku tekstowego z rygorystycznie zadanymi kolumnami reprezentującymi poszczególne informacje[5].

Saloni podjął prace nad opisem odmiany czasowników (koniugacji) na podstawie wcześniejszych prac Tokarskiego, który był również autorem systemu wzorców koniugacyjnych używanego w słowniku Doroszewskiego (Tokarski 1951). Doprowadziły one do wydania precyzyjnego słownika odmiany czasowników (Saloni 2001, 2007). Na tym etapie z grupą zaczął współpracować Marcin Woliński, który stopniowo przejął zadanie formowania komputerowego modelu opracowywanych danych. Tak więc *Czasownik polski* od pierwszego wydania był publikowany na podstawie relacyjnej bazy danych (Saloni i Woliński 2004, 2003). Dane te zostały także wykorzystane od razu do komputerowej analizy fleksyjnej.

Na tym etapie realne stało się myślenie o wydaniu słownika fleksyjnego obejmującego cały materiał języka polskiego (Gruszczyński i Saloni 2006; Gruszczyński 2001; Saloni i Woliński 2005). W ramach podjętych prac opis wszystkich części mowy został stopniowo doprowadzony do stopnia precyzji reprezentowanego przez *Czasownik polski*. Dane opisujące poszczególne części mowy zostały w szczególności sprowadzone do wspólnego modelu komputerowego (Woliński 2009).

Odmiana poszczególnych leksemów jest w SGJP opisywana poprzez przypisanie wzoru fleksyjnego. Podstawowa zasada tworzenia wzorów fleksyjnych polega na odcięciu od wszystkich form odmiany wspólnej części początkowej i potraktowaniu dalszych części zmiennych jako wzoru fleksyjnego. Wzory tworzone według takiej mechanicznej zasady są dość liczne, obecna wersja słownika operuje 1116 wzorami.

Początkowo planowane było wydanie słownika zarówno w postaci papierowej, jak i elektronicznej, jednak po dyskusji uznano, że celowe będzie ograniczenie projektu do wersji elektronicznej. Tak więc dwa pierwsze wydania (Saloni *et al.* 2007b, 2012) miały postać programu komputerowego dystrybuowanego na płycie CD i towarzyszącej książki ze wstępem teoretycznym (Saloni *et al.* 2007a). Trzecie wydanie przyjęło postać aplikacji internetowej http://sgjp.pl (Saloni *et al.* 2015; Woliński i Kieraś 2016). Na potrzeby tego wydania wzory fleksyjne poddano rewizji i opracowano ich nową klasyfikację (Saloni 2016).

[5] Plik ten jest dostępny do pobrania pod adresem http://sgjp.pl/siat/.

W trzecim wydaniu SGJP osiągnął wielkość ponad 330 000 leksemów[6] odpowiadających prawie 4 300 000 wykładników tekstowych. Można przyjąć, że słownik obejmuje cały materiał leksykalny słownika Doroszewskiego rozszerzony o formy ekscerpowane ze współczesnych korpusów i innych słowników. Wypada jednak zaznaczyć, że słownik nie zawiera informacji frekwencyjnej i nie był uzupełniany na podstawie frekwencji wyrazów w korpusie. Część siatki haseł stanowią regularnie derywowane leksemy pochodne, niektóre z nich mniej lub bardziej potencjalne, np. w słowniku notowane są rzeczownikowe nazwy cech typu BIAŁOŚĆ (skoro jest BIAŁY) i WYŻLOŚĆ (skoro jest WYŻLI).

SGJP jest źródłem danych fleksyjnych dla *Wielkiego słownika języka polskiego* (Żmigrodzki 2013–2018), a także jest często przywoływany przez *Wikisłownik* (https://pl.wiktionary.org/).

1.3. Morfeusz SGJP

Pierwsza wersja analizatora fleksyjnego Morfeusz została opracowana około roku 2001. Początkowo dane programu stanowił indeks Tokarskiego (1993, SIaT) skonfrontowany z listą haseł słownika Doroszewskiego (Doroszewski 1958–1969). Opis czasowników pochodził z *Czasownika polskiego* (Saloni 2001). Ta wersja programu została retrospektywnie nazwana *Morfeuszem SIaT* (Woliński 2006b).

Następnie te przybliżone dane były zastępowane bardziej precyzyjnymi danymi SGJP – wynik został udostępniony jako *Morfeusz SGJP*. W kolejnym etapie program zaimplementowano od nowa w celu zwiększenia funkcjonalności (Woliński 2014) – powstała wersja *Morfeusz 2*. Do modułu analizy dodano moduł syntezy. Informacje dostarczane przez program wzbogacono o kwalifikatory ze słownika i prostą klasyfikację nazw własnych. Ważnym nowym elementem jest możliwość zadawania reguł segmentacji tekstu, co umożliwiło skonstruowanie analizatora tekstów historycznych uwzględniającego zmieniające się reguły pisowni łącznej i rozdzielnej (por. p. 1.5).

Podjęto także prace nad połączeniem SGJP z danymi tworzonego społecznościowo słownika fleksyjnego SJP.pl. W ich wyniku powstał słownik Polimorf (Woliński *et al.* 2012). Obecnie program jest dystrybuowany z oboma wariantami słownika (tabela 1.2).

Morfeusz stanowi wyłącznie interfejs do danych słownikowych – nie zawiera modułu interpretującego słowa spoza słownika ani mechanizmu ujednoznaczniającego wyniki. Ten ostatni fakt jest istotny, ponieważ w języku

[6] Wypada zaznaczyć, że leksemy w SGJP są wyodrębnione inaczej niż hasła słownika Doroszewskiego i inaczej niż leksemy analizatora Morfeusz (por. tabela 1.2 i p. 1.7.1). W szczególności w SGJP osobne hasła stanowią odsłowniki i imiesłowy przymiotnikowe, które w słowniku Doroszewskiego i w analizatorze Morfeusz stanowią część leksemu czasownikowego.

Tabela 1.2. Słowniki dystrybuowane z programem *Morfeusz 2* (dane liczbowe dla wersji z końca 2017 roku)

słownik	leksemy	wykładniki
SGJP	264 166	4 037 250
Polimorf	315 055	3 844 535

polskim bardzo częstym zjawiskiem jest homonimia i synkretyzm. Dokładne określenie liczbowe ich poziomu istotnie zależy od przyjętego sposobu znakowania tekstu. Jeżeli przyjąć opisywany tu system zaimplementowany w analizatorze Morfeusz, okazuje się, że 34,5% segmentów w tekście może być interpretowanych jako forma więcej niż jednego leksemu, natomiast 68,5% segmentów ujawnia jakąkolwiek formę niejednoznaczności, w tym dotyczącą wartości przypisanych kategorii gramatycznych (obliczenia wykonano na zrównoważonym 300-milionowym NKJP).

Warto zwrócić uwagę, że w niektórych pracach bardziej użyteczne są wyniki niejednoznaczne niż interpretacje ujednoznacznione statystycznie. Analizę fleksyjną można stosunkowo łatwo wykonać automatycznie z dużą dokładnością. Jest to jedynie kwestia zgromadzenia odpowiednio bogatego słownika fleksyjnego. Ujednoznacznianie interpretacji konkretnych wystąpień wyrazów wymaga odwoływania się do zjawisk składniowych i semantycznych, przez co może być wykonane algorytmicznie tylko w przybliżeniu. Najlepsze statystyczne programy ujednoznaczniające (tagery) dla języka polskiego osiągają skuteczność ok. 93% (Kobyliński i Kieraś 2016). To oznacza, że, przyjmując niezależność ujednoznacznienia na poszczególnych pozycjach tekstu, można oszacować prawdopodobieństwo poprawnego zinterpretowania całego zdania zawierającego 15 wyrazów na mniej niż ½. Niestety wykluczenie poprawnej interpretacji choćby jednego wyrazu może spowodować błędny wynik algorytmu interpretującego zdanie jako całość, np. podczas analizy składniowej. Jednak w automatycznej analizie składniowej wielość interpretacji fleksyjnych nie musi przeszkadzać – to właśnie wykonanie analizy składniowej wypowiedzenia może dać ujednoznacznienie na poziomie fleksyjnym.

Można oczekiwać, że prawdopodobieństwo błędu ujednoznacznienia jest większe dla form rzadkich. Ale to te właśnie formy mogą być interesujące dla użytkownika korpusu. Dlatego przy konstrukcji Narodowego Korpusu Języka Polskiego zdecydowano, że dostępne dla użytkownika muszą być zarówno wszystkie interpretacje fleksyjne przypisane bezkontekstowo, jak i wybrana spośród nich interpretacja właściwa ze względu na kontekst. W konsekwencji wyszukanie interpretacji nietypowych może wymagać przejrzenia większej liczby potencjalnych kontekstów, ale nie jest już niemożliwe.

Morfeusz wewnętrznie używa minimalnych automatów skończonych do zwartej reprezentacji słowników (Woliński 2006b). Program jest udostępniany w wersji gotowej do użycia na główne platformy sprzętowe (Linux, OS X, Win-

dows). Ma on postać biblioteki dynamicznej (API C++), jako że podstawowym odbiorcą ma być programista wbudowujący analizator w tworzone narzędzia przetwarzania tekstu. Udostępniane są także moduły pozwalające na użycie Morfeusza w programach w Pythonie, Perlu, Javie, SWI Prologu oraz interfejs graficzny dla mniej technicznie zaawansowanych użytkowników.

Przyjęty system znakowania tekstu (ang. *tagset*) jest głównym przedmiotem rozważań w pozostałej części tego rozdziału (por. także Woliński 2003; Przepiórkowski i Woliński 2003a,b,c; Przepiórkowski 2009; Woliński 2001). System ten został pierwotnie opracowany na potrzeby znakowania Korpusu IPI PAN (Przepiórkowski *et al.* 2003; Przepiórkowski 2004a; Woliński i Przepiórkowski 2001), a następnie jego wariant został użyty do znakowania Narodowego Korpusu Języka Polskiego (NKJP, Przepiórkowski *et al.* 2012). System znaczników Morfeusza nie był nigdy dokładnie tożsamy z tagsetem NKJP. Przedstawiona w tej pracy najnowsza wersja zmniejsza rozbieżności między tymi systemami.

Istotną kwestią dla użyteczności analizatora fleksyjnego jest licencja, na jakiej jest on udostępniony. Analizator taki działa poprzez kopiowanie do swoich wyników fragmentów używanego słownika fleksyjnego. Dlatego wynik analizy staje się utworem zależnym w stosunku do użytego słownika fleksyjnego, w związku z czym ograniczenia licencyjne słownika zaczynają dotyczyć wyników analizatora[7]. W związku z powyższym zdecydowano, żeby program *Morfeusz 2* i towarzyszące mu słowniki fleksyjne SGJP i Polimorf były dystrybuowane na bardzo liberalnej licencji BSD, która dopuszcza dowolne zastosowania, żądając jedynie zachowania informacji o autorstwie.

Morfeusz jest dość powszechnie używany przez polskie środowisko naukowe. Program został użyty do znakowania Korpusu IPI PAN (Przepiórkowski 2004a) i Narodowego Korpusu Języka Polskiego (Przepiórkowski *et al.* 2012). Na jego bazie opracowano liczne programy ujednoznaczniające analizy (tagery): tager HMM (Dębowski 2004), TaKIPI (Piasecki 2007), PANTERA (Acedański 2010), WMBT (Radziszewski i Śniatowski 2011), Concraft-pl (Waszczuk 2012), WCRFT (Radziszewski 2013), PoliTa (Kobyliński 2014), Toygger (Krasnowska-Kieraś 2017). Morfeusz został zintegrowany z systemami tworzenia płytkich parserów SproUT (Piskorski *et al.* 2004) i Spejd (Przepiórkowski 2008), a także z narzędziem opisu odmiany jednostek wieloczłonowych Multiflex (Savary 2005) i Toposław (Marciniak *et al.* 2011). Wreszcie Morfeusz jest składnikiem opisywanego tu parsera Świgra 2 i parsera POLFIE (Patejuk 2015).

1.4. Dystrybucyjne podejście do fleksji

Niniejsza praca dotyczy przetwarzania języka w jego odmianie pisanej. Obiektywnie istniejącym przedmiotem badań są więc teksty stanowiące cią-

[7] Taką opinię uzyskano od prawnika specjalizującego się w prawie autorskim.

gi znaków pisarskich (a w komputerze obecnie najczęściej reprezentowane jako ciągi znaków Unicode). Wszystkie pojęcia nałożone na tę reprezentację są już interpretacją, modelem badanego zjawiska. Zależą one od przyjętego podejścia badawczego. Powstaje pytanie, czy da się podać systematyczną zasadę tworzenia takiego modelu w odniesieniu do fleksji.

Podejściem adekwatnym do komputerowego opisu języka wydaje się dystrybucjonizm, którego przedstawicielami są w szczególności Zygmunt Saloni i Marek Świdziński; ich poglądy miały największy wpływ na ukształtowanie prezentowanego tu opisu.

Jak pisze Saloni (1974a, s. 9): „Z punktu widzenia gramatycznego najważniejszą cechą form fleksyjnych jest ich ściśle określona przydatność jako pewnego rodzaju składników tekstu, czyli innymi słowy – zdolność użycia ich w zdaniach w określonych pozycjach". Idea opisu polega więc na ścisłym odwołaniu do przedmiotu obserwacji – tekstów – i sklasyfikowaniu występujących w nich segmentów poprzez zebranie możliwych kontekstów ich występowania. Jeżeli możliwe konteksty się różnią, uznaje się, że segmentom trzeba przypisać różne charakterystyki gramatyczne. Konstruowana jest w ten sposób relacja równoważności dystrybucyjnej.

Podejście dystrybucyjne rozumiane mechanicznie, jako zamiar przejrzenia wszystkich występujących w tekstach kontekstów segmentu, było do niedawna nierealne. W epoce ogromnych korpusów tekstowych i technik *big data*, nie tylko stało się ono możliwe, ale zyskało dużą popularność w maszynowym uczeniu, zwłaszcza w wektorowej semantyce dystrybucyjnej, np. *word2vec* (Mikolov *et al.* 2013). W niniejszej pracy przedstawiono jednak bardziej tradycyjne podejście do modelowania języka.

Oczywiście równoważność musi być rozpatrywana na odpowiednim poziomie abstrakcji. W przeciwnym razie za dystrybucyjnie równoważne zostałyby uznane jedynie dokładne synonimy, które jednak praktycznie nie występują ze względu na ekonomię ekspresji językowej – w języku „nie opłaca się" mieć elementów dokładnie zastępowalnych. Tak więc nawet bliskie synonimy różnią się choćby nacechowaniem emocjonalnym, a to sprawia, że ich użycia w tekstach będą różne.

Drugim problemem jest konieczność generalizacji. Dostępnym przedmiotem badań może być zbiór tekstów, jednak celem badania jest zbudowanie modelu języka rozumianego jako potencja tworzenia tekstów. Ograniczenie badania do korpusu tekstów, nawet bardzo dużego, skutkowałoby przekłamaniami wynikającymi z przypadkowej nieobecności w nim pewnych konstrukcji należących do języka. Tak więc rozważając kwestie dystrybucji wyrazów, trzeba w pewnym stopniu odwoływać się do wyczucia badacza co do możliwości lub niemożliwości wygenerowania pewnych konstrukcji, które przez korpus nie zostały poświadczone.

Konstrukcję równoważności dystrybucyjnej można sobie wyobrazić jako przeglądanie możliwych kontekstów (wypowiedzeń z wyjętym jednym segmentem) i wyszukiwanie takich ich grup, których nie można rozróżnić po-

przez wskazanie segmentu pasującego tylko do niektórych kontekstów w grupie. Na przykład można wyróżnić klasę reprezentowaną przez poniższe konteksty, do których pasuje długa lista segmentów typu *psem*, *kotem*, *mężem*, ...

(2) Przejęła się małym [ ].
(3) Nie jestem [ ] sąsiadki.
(4) Wychodzi na spacer z mężem i [ ].
(5) Przejęła się miaucząym [ ].

Warto podkreślić, że klasy abstrakcji relacji równoważności dystrybucyjnej są klasami form fleksyjnych, a nie klasami segmentów. W szczególności relacja ta nie stanowi podziału zbioru segmentów należących do języka, istnieją bowiem segmenty, które mogą się pojawić w kontekstach należących do różnych klas. Na przykład segment *psa* może wystąpić w obu poniżej wymienionych kontekstach, które jednak należą do różnych klas, bo istnieją segmenty, które je różnicują (np. segmenty *dziewczyny* i *dziewczynę*):

(6) *Nie ma* [ ].
(7) *Widzę* [ ].

Trzeba więc przyjąć, że istnieją (co najmniej) dwie formy o wykładniku *psa*, należące do klas reprezentowanych przez konteksty (6) i (7). Dalsza część opisu polega w zasadzie na sklasyfikowaniu (nazwaniu) tak wyróżnionych form fleksyjnych za pomocą wartości kategorii gramatycznych.

Użyteczną dla opisu gramatycznego relację równoważności dystrybucyjnej form uzyskuje się poprzez abstrahowanie od znaczenia. Odpowiada to mniej więcej dopuszczeniu wszystkich wypowiedzeń, które mogłyby mieć sens w „możliwych światach", a więc również we śnie lub w baśni, a odrzucenie wypowiedzeń gramatycznie niepoprawnych, w których jakieś wyrazy stoją w kontekstach dla nich niewłaściwych. Tak więc zdanie (8) zostanie uznane za gramatyczne, mimo że jest w nim problem semantyczny: pojęcie abstrakcyjne nie nadaje się na wykonawcę czynności czytania. Można jednak sobie wyobrazić, że ktoś użyje tego zdania, przypisując mu odpowiednio metaforyczne znaczenie. Dlatego jest ono równie dobrym kontekstem dla formy abstrakcyjnego rzeczownika SPRAWIEDLIWOŚĆ, jak dla konkretnego rzeczownika DZIEWCZYNA. Zdanie (9) nie należy do języka, bowiem kontekst po wyrazie *przez* wymaga form typu *most*, *rzeczkę*, które należą do innej klasy dystrybucyjnej niż *mostem*.

(8) *Sprawiedliwość nie czyta gazet.*
(9) **Szła przez mostem.*

Drugim etapem dystrybucyjnego opisu fleksji jest wyróżnienie leksemów. Jego podstawą nie jest dystrybucja, ale „identyczność lub regularne zróżnicowanie" odniesienia do obiektów pozajęzykowych (Saloni 1974a,b). Na tej zasadzie można na przykład postulować istnienie leksemu PIES o wykładnikach *pies*, *psa*, *psach*, *psami*, *psem*, *psie*, *psom*, *psu*, *psy*, *psów*, ponieważ wszystkie te segmenty w tekście odnoszą się do konkretnego egzemplarza (lub konkretnych

egzemplarzy) czworonoga pewnego gatunku lub służą wyrażeniu sądu o tym gatunku. Podjęcie decyzji, które wykładniki zostaną zaliczone do danego leksemu, wymaga analizy regularności ich występowania, a więc istnienia serii zróżnicowanych analogicznie. Tak więc wykładnik *psem* uznaje się za przynależny do leksemu PIES, bo analogiczne zależności można wskazać dla segmentów *kotem*, *kaszalotem*, *berkiem* itd. Można by także rozważać zaliczenie segmentu *psim* do tego samego leksemu, tutaj jednak opozycja nie jest tak samo regularna: co prawda dla KOT można wskazać analogiczną formę *kocim*, ale dla leksemów ABSURD, BUREK, KUDŁACZEK i innych takiej formy wskazać nie sposób. Niekiedy decyzja o zaliczeniu danych form do tego samego leksemu jest przedmiotem arbitralnej oceny. Na przykład formy typu *bielszy* istnieją tylko dla ok. 1000 z 30 000 przymiotników. Powstaje więc pytanie, czy to zjawisko jest wystarczająco regularne, by uznać je za fleksyjne, a więc zaliczyć te formy do leksemu przymiotnikowego. Poszczególni badacze różnie odpowiadają na to pytanie.

Następnym krokiem opisu jest wprowadzenie kategorii fleksyjnych, pozwalających odróżnić od siebie wszystkie formy, z których składają się poszczególne leksemy. Repertuar i zbiór wartości poszczególnych kategorii jest motywowany intuicją badacza, a po części tradycją. Celem jest tu optymalne wychwycenie regularności w materiale językowym.

Na przykład, aby odróżnić od siebie formy typu *pies* i typu *psy*, wprowadza się kategorię gramatyczną liczby o dwóch wartościach: pojedyncza i mnoga. Intuicją prowadzącą do wprowadzenia tej kategorii jest fakt, że forma *pies* odnosi się do jednego obiektu pozajęzykowego, a forma *psy* – do grupy obiektów. Jednak w istocie kategorie gramatyczne nie modelują odniesienia wyrazów do rzeczywistości, tylko klasyfikują typy kontekstów, w szczególności charakteryzują inne formy wchodzące w związki składniowe z daną formą. Wartości kategorii gramatycznych można definiować przez wskazanie wybranych kontekstów (nazywa się je kontekstami testowymi), do których pasują jedynie formy charakteryzujące się daną wartością danej kategorii. Tak więc typowo występuje zbieżność kategorii liczby z odniesieniem pozajęzykowym, ale w wypadku[8] słów takich jak *skrzypce* używane są w ich kontekście formy liczby mnogiej przymiotników i czasowników, mimo że opis odnosi się do jednego przedmiotu:

(10) *W futerale znajdowały się zdezelowane skrzypce.*
(11) *Kiedyś na festynie usłyszałem, jak człowiek we fraku grał na dużych skrzypcach.*

Podobnie konwencjonalna jest kategoria rodzaju gramatycznego (por. p. 1.6.3).

[8] W niniejszej pracy, zgodnie z konwencją zaproponowaną kiedyś przez Saloniego, formy leksemu PRZYPADEK są używane wyłącznie jako użycia terminu gramatycznego, a w kontekstach takich jak w tym zdaniu stosowane są formy leksemu WYPADEK.

Najistotniejsze kategorie gramatyczne stosowane w tej pracy zostaną omówione w punkcie 1.6.

Ostatnim elementem opisu jest klasyfikacja leksemów do klas gramatycznych, czyli części mowy. W przyjętym tu systemie (por. p. 1.7) podstawą klasyfikacji leksemów odmiennych jest zestaw kategorii gramatycznych, przez które się one odmieniają. Tak więc na przykład przyjęto, że rzeczowniki to leksemy odmienne przez liczbę i przypadek, ale nie rodzaj, a liczebniki to leksemy odmienne przez przypadek i rodzaj, ale nie liczbę. Leksemy nieodmienne są klasyfikowane ze względu na własności składniowe.

Nazwy klas są dobrane zgodnie z tradycją, tak więc na przykład za rzeczowniki uznaje się „jednowyrazowe nazwy przedmiotów konkretnych" i wyrazy dzielące z nimi własności gramatyczne (Saloni 2005, s. 29). Jednak już ścisłe rozumienie liczebników znacząco rozchodzi się z intuicją wyrazów służących do liczenia, na przykład leksem *jeden* ze względu na swoją odmienność musi zostać uznany za przymiotnik (por. p. 1.7.6).

1.5. Segmentacja tekstu

Wykonanie analizy fleksyjnej wymaga najpierw wyróżnienia w tekście segmentów, a więc odcinków tekstu (ciągów znaków), które podlegają interpretacji fleksyjnej.

W pierwszym przybliżeniu interpretacji podlegają słowa rozumiane jako maksymalne sekwencje znaków pisarskich niezawierające odstępu ani znaków interpunkcyjnych. Nie jest jednak prawdą, że segment może składać się jedynie z liter. W prezentowanym opisie za stanowiący część segmentu uznano apostrof używany w odmianie wyrazów obcych, np. *Chomsky'ego*. Tak samo potraktowano dywiz (łącznik) używany w odmianie skrótowców (*PRL-u*, *PiS-em*). Dywiz został jednak uznany za osobny segment w formacjach przymiotnikowych takich jak *biało-czerwony* (por. p. 1.7.5). Za osobny segment uznano także kropkę używaną obowiązkowo w niektórych skrótach (zob. p. 1.7.8).

W pewnych kontekstach interpretacji wydają się podlegać odcinki dłuższe niż słowa. Tradycyjnie jako fleksyjne traktuje się na przykład tzw. formy analityczne (a więc złożone z wielu słów) czasu przyszłego (np. *będzie czytał*) i trybu rozkazującego (np. *niech czyta*). Za segmenty przekraczające granice słów uważane są też wyrażenia typu *po polsku*. Jednak dopuszczenie segmentów dłuższych niż słowa raczej komplikuje interpretację, a nie ją upraszcza. Problemów interpretacyjnych mogą nastręczać na przykład następujące wypowiedzenia:

(12) *Będzie pisał lub czytał.*
(13) *Mówił po polsku i angielsku.*

W wypowiedzeniu (12) trzeba by albo uznać, że słowo *Będzie* jest częścią zarówno segmentu *Będzie pisał*, jak i *Będzie czytał* (a więc że segmenty nie muszą

być rozłączne); albo też, że dopiero cały ciąg *Będzie pisał lub czytał* jest jednym segmentem. Oba te rozwiązania spowodowałoby spiętrzenie problemów zarówno technicznych, jak i interpretacyjnych.

W prezentowanej tu koncepcji przyjęto (por. Przepiórkowski i Woliński 2003c), że segmenty muszą być rozłączne (każdy znak tekstu musi być częścią tylko jednego segmentu) i ciągłe (pomiędzy znakami należącymi do jednego segmentu nie mogą występować znaki do niego nienależące). W konsekwencji segmenty nie mogą też zawierać znaków odstępu.

Oznacza to, że każde ze słów w przykładzie (12) jest interpretowane jako wykładnik osobnej formy fleksyjnej. Podobnie swoistą interpretację otrzymują na przykład segmenty *polsku* (ta forma została włączona do paradygmatu leksemu POLSKI) i *indziej* występujące w *gdzie indziej* (ta interpretacja wymaga stworzenia specjalnej jednostki słownika, leksemu nieodmiennego INDZIEJ scharakteryzowanego jako *człon frazeologizmu*).

Warto nadmienić, że takie rozstrzygnięcie nie oznacza uznania wszelkich zależności między jednostkami oddzielonymi odstępem za składniowe. Stanowi raczej wyraz tego, że w komputerowym przetwarzaniu tekstu ze względów praktycznych dobrze jest uniknąć komplikacji związanych z jednostkami nieciągłymi na wstępnym etapie przetwarzania. Zjawisko „form analitycznych" wymaga uwzględnienia na dalszych etapach przetwarzania. W przedstawianym dalej opisie składniowym do reprezentacji tego rodzaju zjawisk wykorzystano formalizm poziomu składniowego (służy do tego pojęcie form składniowych, p. 2.3). Jednak można sobie wyobrazić wydzielenie tego etapu przetwarzania i zrealizowanie go za pomocą mechanizmów prostszych, a przez to bardziej efektywnych.

Są też sytuacje, gdy jednolity ciąg znaków alfanumerycznych jest interpretowany jako więcej niż jeden segment. W punkcie 1.7.2 (strona 45) przedstawiono zasady rozbijania słów reprezentujących czas przeszły i tryb warunkowy czasowników na formę pseudoimiesłowu, aglutynantu i partykuły warunkowej. Jako osobne segmenty są traktowane: aglutynacyjna postać *-ń* zaimka osobowego ON, segmenty *-że*, *-ż* i *-ć* stanowiące wykładniki leksemów wzmacniających ŻE i Ć oraz segment *-li* stanowiący wykładnik leksemu pytającego LI (por. Saloni *et al.* 2012, p. 1.4, s. 21).

Na więcej tego rodzaju zjawisk można natrafić przy przetwarzaniu tekstów historycznych. Na przykład w polszczyźnie siedemnastowiecznej partykuła *nie* bywa pisana łącznie z formami czasownikowymi, a przyimki – łącznie z następującym po nich rzeczownikiem (Kieraś *et al.* 2017).

Istnieją argumenty za tym, żeby uznać znaki interpunkcyjne za elementy składniowo istotne w języku polskim, w szczególności pełniące rolę spójników współrzędnych (Świdziński 1992). Dlatego zostały one również uwzględnione w opisie fleksyjnym jako osobne segmenty. Z technicznego punktu widzenia segmenty te są interpretowane jako wykładniki „leksemów" o nazwach równokształtnych z analizowanym segmentem i scharakteryzowane znacznikiem interp.

1.6. Najważniejsze kategorie gramatyczne

W tym punkcie przedstawiono kategorie gramatyczne istotne dla konstrukcji podziału leksemów na klasy gramatyczne. Pewne dodatkowe kategorie użyteczne w opisie odmiany leksemów poszczególnych klas będą wymieniane przy nich.

Przy ustalaniu zbioru wartości dopuszczalnych dla danej kategorii gramatycznej (fleksyjnej) przestrzegana jest zasada odwoływania się do maksymalnego zróżnicowania (por. Mańczak 1956, s. 118). Tak więc to, że w wypadku leksemu OKNO forma używana w kontekście wołaczowym jest identyczna z formą używaną w kontekście mianownikowym nie może być podstawą do stwierdzenia, że leksem ten ma mniej wartości kategorii przypadka. Skoro istnieją leksemy (np. ŁAWKA), w których dystrybucja form różni się, to rozróżnienie to musi stosować się do wszystkich leksemów rzeczownikowych, a w istocie do wszystkich leksemów, którym przysługuje wartość przypadka. Jeżeli istnieją powody wyróżnienia na przykład siedmiu wartości kategorii przypadka, to w odniesieniu do wszystkich form nieobojętnych na kategorię przypadka stosuje się tych siedem wartości. Ta prosta zasada ma szczególnie daleko idące konsekwencje, jeśli chodzi o kategorię rodzaju gramatycznego.

1.6.1. Kategoria liczby

Kategoria liczby jest przede wszystkim kategorią fleksyjną form rzeczownikowych. W typowym wypadku odróżnia ona formy odnoszące się do jednego obiektu pozajęzykowego (*pies*, *kanarkiem*, *urzędniczce*) od form odnoszących się do grupy obiektów (*psy*, *kanarkami*, *urzędniczkom*). Formom pierwszego typu przypisuje się wartość liczby pojedynczej (sg), drugiego – mnogiej (pl). Analogiczne podziały można też dostrzec w zbiorach form uzgadniających się z rzeczownikami, a więc form przymiotnikowych i czasownikowych.

Prawidłowość ta jest zaburzona dla tzw. rzeczowników *plurale tantum*, które nie posiadają form liczby pojedynczej. W ich wypadku odniesieniem pozajęzykowym bywa pojedynczy obiekt (SKRZYPCE), niepodzielna masa (POMYJE), para osób (WUJOSTWO). Wszystkie formy rzeczowników tej klasy wchodzą jednak w związki gramatyczne z takimi formami, jak mnogie formy rzeczowników dwuliczbowych. Tak więc ich wartość kategorii gramatycznej liczby uznaje się za mnogą.

1.6.2. Kategoria przypadka

Bardzo przejrzysty wywód pojęcia przypadka w języku polskim można znaleźć w pracy Saloniego (2005), na niej oparty jest więc opis w tym podrozdziale. Punktem wyjścia jest obserwacja zróżnicowania w tekstach form odnoszących się do tego samego przedmiotu materialnego (a więc form prototypowych rzeczowników). W zależności od kontekstu składniowego ten sam obiekt zostanie

raz nazwany *pies*, a w innym miejscu *psa* (rozważa się przy tym formy mające tę samą wartość kategorii liczby).

Wartości kategorii przypadka wprowadza się, dzieląc możliwe konteksty składniowe, w których może wystąpić forma rzeczownika. Chodzi o znalezienie podziału o minimalnej liczbie klas, charakteryzującego się tym, że dowolny kontekst z danej klasy powoduje pojawienie się tej samej formy rzeczownika.

Należy przy tym zwrócić uwagę, że tożsamość form to coś więcej niż tylko tożsamość segmentów. W odmianie rzeczowników w sposób systematyczny te same segmenty pełnią różną funkcję – czyli są synkretyczne. Tak więc postuluje się dwie różne formy rzeczownika reprezentowane w tekście przez segment *psa*, ponieważ do kontekstów *Nie ma* *psa* i *Widzę* *psa* pasują różne wykładniki innych rzeczowników. Na przykład w wypadku leksemu KROWA są to segmenty *krowy* i *krowę* odpowiednio[9].

W skrajnym wypadku leksemów powierzchniowo nieodmiennych jak GNU lub KAKADU ten sam wykładnik może zostać zinterpretowany jako dowolna forma rzeczownika, ponieważ wykładnik ten może wystąpić we wszystkich kontekstach właściwych dla form rzeczowników.

Przyjmuje się, że dla języka polskiego wystarcza podział na 7 klas. Sześć z nich, wyróżnionych na podstawie łączliwości z czasownikami i przyimkami, Saloni ilustruje następującymi kontekstami (które stanowią reprezentantów odpowiednich klas):

(14) *To* *pies/ręka*. mianownik nom
(15) *Nie ma* *psa/ręki*. dopełniacz gen
(16) *Przyglądam się* *psu/ręce*. celownik dat
(17) *Widzę* *psa/rękę*. biernik acc
(18) *Interesuję się* *psem/ręką*. narzędnik inst
(19) *Opowiadam o* *psie/ręce*. miejscownik loc

Dzięki wyborowi bardzo neutralnych semantycznie czasowników konteksty te mogą służyć jako konteksty testowe przy badaniu przypadka konkretnej formy.

Istnieją wykładniki części leksemów rzeczownikowych, które nie pasują do żadnego z wymienionych kontekstów testowych. Są to formy używane, żeby zwrócić się bezpośrednio do odbiorcy wypowiedzi, np. *chłopcze, kobieto, Jolu, dziadziu, sędzio*. Są one traktowane tradycyjnie jako reprezentujące siódmą wartość przypadka – wołacz voc, choć są wyróżniane na podstawie innych kryteriów, jako że nie są wymagane przez żaden kontekst czasownikowy ani przyimkowy. Wchodzą one jednak w związki z przymiotnikami (*Drogi chłopcze!*).

Jak pisze Saloni (Saloni *et al.* 2012, s. 34), lepiej uzasadnione teoretycznie byłoby uznanie wołacza za wartość dodatkowej kategorii fleksyjnej rzeczownika.

[9] Aby opisać prawidłowości układu synkretyzmów form rzeczownikowych, trzeba by odwołać się do kategorii rodzaju wprowadzonej dalej. W tym miejscu potrzebna jest jednak jedynie obserwacja zróżnicowań wykładników.

W takim opisie formy wołaczowe łączyłyby się z formami mianownikowymi podrzędnych przymiotników.

W przyjętym tu opisie, za SGJP, uznano, że wołacz jest jednym z przypadków, a w związku z tym również pewne formy przymiotników są wołaczowe (są one zawsze synkretyczne z formą mianownika).

1.6.3. Kategoria rodzaju

Kategoria rodzaju w języku polskim jest konstruowana jako podział zbioru leksemów rzeczownikowych (Mańczak 1956; Saloni 1976). Zasadniczym kryterium podziału jest łączliwość form tych leksemów z formami czasowników, przymiotników i liczebników, a więc klas, których leksemy są odmienne przez rodzaj. Konstrukcja opiera się więc na założeniu, że rzeczowniki nie odmieniają się przez rodzaj, wszystkim formom danego rzeczownika ma przysługiwać ta sama wartość, natomiast odmieniają się przez niego przymiotniki, czasowniki i liczebniki.

Zrąb klasyfikacji rodzajowej rzeczowników wyznacza się poprzez obserwację ich łączliwości z formami przymiotnikowymi. Największe zróżnicowanie obserwuje się dla form biernika liczby pojedynczej i mnogiej. Prowadzi ono do wyróżnienia pięciu klas rodzajowych (Mańczak 1956):

	m1	m2	m3	n	f
acc. sing.	*tego*		*ten*	*to*	*tę*
acc. pl.	*tych*	*te*			

Tabela zawiera formy przymiotnika TEN, z którymi łączą się formy rzeczowników poszczególnych rodzajów. W każdej kolumnie stoi inna para form przymiotnikowych. Każda para determinuje jedną klasę rodzajową rzeczowników. Klasy te noszą konwencjonalne nazwy: m1 – męskoosobowy (np. CHŁOPIEC), m2 – męskozwierzęcy (np. KOT), m3 – męskorzeczowy (np. STÓŁ), n – nijaki (np. DRZEWO), f – żeński (np. DZIEWCZYNA). Nazwy odnoszą się do typowych reprezentantów klas, relacjonują jednak cechę czysto gramatyczną, rzeczownikami męskozwierzęcymi są więc także na przykład nazwy tańców (POLONEZ, MAZUR), mogą one też nazywać istoty żeńskie (BABSZTYL).

Analiza połączeń z formami czasownikowymi nie wnosi drobniejszego zróżnicowania.

Saloni (1976) rozwija ten system poprzez analizę związków z formami liczebników. Zauważa najpierw, że wśród rzeczowników nijakich można wyróżnić takie, których formy łączą się z formami tzw. liczebników głównych (np. *siedmioma oknami*), i takie, które łączą się z formami tzw. liczebników zbiorowych (np. *siedmiorgiem dzieci*). Staje się to podstawą rozdzielenia rodzaju nijakiego na n1 i n2:

	m1	m2	m3	n1	n2	f
acc. sing.	*tego*		*ten*	*to*		*tę*
acc. pl.	*tych*	*te*				
acc. pl.				*pięcioro*	*pięć*	

W ramach tej koncepcji nie dzieli się liczebników na główne i zbiorowe, oba te zbiory form współistnieją w obrębie paradygmatu liczebnika. Forma *dwoje*, przynależna do paradygmatu leksemu DWA odróżnia się od formy *dwa* wartością rodzaju.

Następnie Saloni zajmuje się grupą rzeczowników *plurale tantum*, a więc nie posiadających form liczby pojedynczej, jak WUJOSTWO, DRZWI, SPODNIE. Jak pisze Saloni: „Zbadanie ich łączliwości z formami przymiotnikowymi i czasownikowymi każe już wyodrębnić je w osobną grupę rodzajową: nie dopuszczają one bowiem w ogóle połączeń z formami fleksyjnymi liczby pojedynczej". Analiza prowadzi do wyróżnienia trzech klas (trzech rodzajów przymnogich) dla rzeczowników *plurale tantum*. Klasa p1 (np. PAŃSTWO, WUJOSTWO) wyróżnia się łączliwością z formą *tych* (a więc z takimi formami przymiotników i czasowników jak rzeczowniki m1). Pozostałe rzeczowniki *plurale tantum* (np. DRZWI, SKRZYPCE) łączą się z formą *te* (jak rzeczowniki dwuliczbowe z wyjątkiem m1, nazywane skrótowo niemęskoosobowymi). Klasy p1 i p2 dopuszczają połączenie z formami liczebników (co do zasady zbiorowymi) w odróżnieniu od klasy p3, dla której bezpośrednie połączenie jest niemożliwe (np. dla rzeczowników SPODNIE, POMYJE).

	m1	m2	m3	n1	n2	f	p1	p2	p3
acc. sing.	*tego*		*ten*	*to*		*tę*	⊥		
acc. pl.	*tych*	*te*					*tych*	*te*	
acc. pl.	*pięciu*	*pięć*		*pięcioro*	*pięć*		*pięcioro*		⊥

W SGJP przyjęto dziewięć wartości kategorii rodzaju zaproponowanych przez Saloniego (Saloni *et al.* 2012). Wadą tego systemu w zastosowaniach komputerowych jest jego ogromna szczegółowość. Ze względu na uzgodnienia wszystkie dziewięć wartości rodzaju może przysługiwać formom przymiotników i czasowników, mimo że w ich paradygmatach tak szczegółowych rozróżnień nie można się dopatrzeć.

Również w analizatorze Morfeusz stosowany był pełny system Saloniego. Nie został on jednak wdrożony przy znakowaniu Narodowego Korpusu Języka Polskiego (Przepiórkowski 2009), ponieważ tak duża liczba rodzajów powodowałaby trudności zarówno dla językoznawców znakujących korpus, jak i dla algorytmów maszynowego uczenia zastosowanych do zagadnienia ujednoznaczniania fleksyjnego. Ustalenie bowiem rodzaju danej formy przymiotnikowej lub czasownikowej wymagałoby częściej analizy kontekstu, w którym być może występuje uzgadniający się rzeczownik. Przy braku rzeczownika pojawia się nieusuwalna niejednoznaczność form. Dlatego przy znakowaniu NKJP przyjęto użycie tylko jednego rodzaju nijakiego oraz pominięcie rodzajów przy-

mnogich (które zostały zinterpretowane jako m1 lub n), czyli powrócono do systemu Mańczaka.

Łączliwość rzeczowników *plurale tantum*

Rozróżnienie rodzajów p2 i p3 wydaje się najsłabiej zarysowanym w systemie rodzajowym Saloniego. Według Andrzejczuk (2011) liczenie desygnatów leksemów tych rodzajów idzie użytkownikom polszczyzny z trudem, a użycie z nimi liczebników zbiorowych nie wydaje się powszechnie akceptowane. Powstaje więc wątpliwość, czy na możliwości bezpośredniego połączenia formy rzeczownikowej z liczebnikiem można oprzeć rozróżnienie rodzajowe. Autorka uważa, że nie. Rozchwianie decyzyjne widać zresztą w samym SGJP: ok. 400 leksemów ma przypisany rodzaj alternatywny p3/p2 (na przykład dla nazw geograficznych typu Ustrzyki, Bełdany, Andy). Opis taki sygnalizuje niepewność autorów co do możliwości użycia liczebnika zbiorowego. Skoro dotyczy ona aż ¼ leksemów p3, rozróżnienie wydaje się wątpliwe.

System rodzajów Saloniego można też skrytykować na gruncie teoretycznym. Mianowicie, żeby być wiernym przedstawionym założeniom konstrukcji kategorii rodzaju, należałoby uznawać, że dane dwa leksemy rzeczownikowe różnią się rodzajem tylko wówczas, gdy istnieją formy tych leksemów o tych samych wartościach liczby i przypadka oraz forma jakiegoś leksemu wchodzącego w uzgodnienie rodzajowe z tymi pierwszymi, wykazująca łączliwość z tylko jedną z dwóch form rzeczownikowych (Woliński 2001). Zastrzeżenie o istnieniu form o tych samych wartościach liczby i przypadka jest istotne, gwarantuje ono bowiem, że kategoria rodzaju faktycznie jest rozpatrywana osobno od innych. Tymczasem rodzaje przymnogie zgodnie z przytoczonym uprzednio cytatem zostały wprowadzone na podstawie braku ich łączliwości z formami liczby pojedynczej. Jednak z tego faktu zdaje już sprawę kategoria liczby.

Ponieważ formy rzeczowników *plurale tantum* nie łączą się z formami liczby pojedynczej przymiotników, to nie wiadomo, do której z klas m2, m3, n1, n2, f miałyby należeć te z nich, które łączą się z formą *te*. (Nie ma problemu z łączliwymi z formą *tych* – one zachowują się jednoznacznie jak m1). Decyzja o przypisaniu ich do jednej z wcześniej wyróżnionych klas byłaby niesprzeczna, ale arbitralna. Argumentem za wyborem dokonanym w NKJP może być obserwacja łączliwości z liczebnikami: z wymienionych klas tylko rzeczownikom nijakim zdarza się łączyć z liczebnikami zbiorowymi (por. Woliński 2001; Przepiórkowski 2003a).

Warto zauważyć[10], że pominięcie w NKJP rodzajów przymnogich powoduje utrudnienie dla użytkowników korpusu: nie istnieje sposób zadania pytania o wystąpienia form rzeczowników *plurale tantum*, których dystrybucja w tekście może być ciekawym tematem badań korpusowych.

[10] Na problemy te zwrócił moją uwagę Witold Kieraś, za co niniejszym dziękuję.

Ponadto zadanie w wyszukiwarce korpusowej pytania o rzeczowniki nijakie daje w wynikach wystąpienia m.in. rzeczowników NOŻYCE, ŁAKOCIE, GORCE i KRĘTOROGIE, które raczej nie należą do oczekiwanych wyników (gdyby próbować dotworzyć do wymienionych leksemów formy liczby pojedynczej na podstawie przypisanych im wzorców fleksyjnych, żaden z nich nie okazałby się rzeczownikiem morfologicznie nijakim). Jednak ten zarzut dotyczy również samej klasy p2: zróżnicowanie morfologiczne jest tak samo duże w obrębie klasy p2 (np. NOŻYCE i ŁAKOCIE), jak w stosunku do dwuliczbowych rzeczowników nijakich. Klasyfikacja jest bowiem oparta na łączliwości, a nie na kształcie wypełnień klatek paradygmatu.

Należy więc przyznać, że brak form liczby pojedynczej rzeczownika jest cechą wartą uwzględnienia w znakowaniu korpusu. Nie oznacza to jednak, że musi on być wyrażony zróżnicowaniem rodzajowym.

Łączliwość z formami liczebnikowymi

Kwestia łączliwości rzeczowników nijakich z formami liczebnikowymi straciła we współczesnym języku ostrość. Jest raptem kilka leksemów, które wykazują łączliwość jedynie z formami zbiorowymi (DZIECKO, OKO, UCHO[11]). Inne leksemy przypisane w SGJP do rodzaju n1 spotyka się w tekstach (zwłaszcza mniej starannych) z liczebnikami głównymi, co ilustrują poniższe przykłady:

(20) *Malajowie [w gromadzie Plejady] widzą sześć kurcząt, którym towarzyszy niewidoczna matka.* [NKJP300, Gazeta Wyborcza]

(21) *Złożywszy arcykapłanowi w darze trzy cielęta, trzy korce mąki i garniec wina, pomodliła się gorąco i pełna radości wróciła do Ramataim.* [NKJP300, Zenon Kosidowski, *Opowieści biblijne*]

(22) *Przed dwoma laty dorosłe ptaki miały z reguły po cztery, a nawet pięć piskląt.* [NKJP300, Dziennik Bałtycki]

(23) *Dochowali się dwóch córek i dwóch synów oraz siedmiu wnucząt.* [NKJP300, Gazeta Pomorska]

Według tradycyjnej oceny poprawnościowej wypowiedzenia te są niepoprawne (Jadacka 2005), jednak są na tyle częste, że prawdopodobnie mamy do czynienia ze zmianą zachodzącą w systemie językowym. Z praktycznego punktu widzenia sensowną strategią wydaje się akceptowanie na potrzeby automatycznego przetwarzania tekstów zdań zawierających takie połączenia.

W tekstach można znaleźć też użycia liczebników zbiorowych z rzeczownikami innych rodzajów niż n1 i przymnogie. Przede wszystkim liczebników zbiorowych używa się z rzeczownikami m1, jeżeli fraza odnosi się do zbiorowo-

[11] Dla ostatnich dwóch istnieją leksemy o homonimicznym lemacie, które wykazują łączliwość z formami głównymi: *człowiek ma dwoje oczu*, ale *dwa oka sieci były rozerwane*.

ści osób obu płci (przykłady (24) i (25)) – takie użycia są zgodne z zaleceniami poprawnościowymi.

(24) *W przedpokoju czekało kilkoro pacjentów, ale doktora jeszcze nie było.* [NKJP300]

(25) *We trzy wyprowadziły czternaścioro pięciolatków na piknik nad rzeką.* [NKJP300]

Zdarzają się jednak także użycia form zbiorowych na przykład z rzeczownikami żeńskimi. Niektóre takie użycia mają charakter ustabilizowanych zbitek (zwłaszcza z liczebnikiem OBOJE, por. (26)). Jednak forma zbiorowa wydaje się być także używana jako środek stylistyczny (28).

(26) *Jesteś wewnętrznie zdolny do prawdziwej, platonicznej przyjaźni z osobami obojga płci bez romantycznego zaangażowania się.* [NKJP300]

(27) *Kiedyś miałem dwoje szczęk.* [NKJP300, Gałczyński]

(28) *Że bratem moim jesteś od dziś, przyjmę z ręki twojej te siedmioro owiec na znak, że twoją była ta studnia.* [NKJP300]

Karpowicz (1994) na podstawie badań ankietowych (o niewielkiej skali) proponuje zmiany w zaleceniach normatywnych dotyczących liczebników zbiorowych. Obserwuje mianowicie, że w opinii uczestników ankiety zbiorowe „formy uznane za poprawne rażą swoją sztucznością" (por. przykład (30)). Formy liczebników głównych bywają wbrew normie używane do zbiorów różnopłciowych, zwłaszcza jeśli różnopłciowość można wywnioskować z kontekstu. Złożone liczebniki główne są używane w kontekstach tradycyjnie wymagających liczebników zbiorowych, tak więc zdanie (29) zostaje określone przez ankietowanych jako „zupełnie zrozumiałe".

(29) *Przyleciał samolot z trzystu siedemdziesięcioma pięcioma pasażerami.*

(30) *Przyleciał samolot z trzystu siedemdziesięciorgiem pięciorgiem pasażerów.*

Podsumowując: kwestia form głównych i zbiorowych liczebników daje ostre preferencje łączliwości tylko dla bardzo niedużej grupy leksemów. Dlatego rozciąganie jej na wszystkie formy przymiotników i czasowników należy uznać za zbyt duży koszt w zastosowaniach komputerowych. Co ciekawe, zbliżone wnioski można znaleźć w pracy Saloniego (1974b, s. 100).

Przyjęte rozwiązanie

W prezentowanym tu opisie, wdrożonym w najnowszej wersji analizatora Morfeusz 2, przyjęto rozwiązanie pośrednie, zachowujące większość informacji z modelu Saloniego. Zbiór wartości kategorii rodzaju został ograniczony do pięciu: m1, m2, m3, n, f. Jednak w znacznikach opisujących rzeczowniki rodzaju n pojawiła się nowa pozycja o wartościach ncol i col opisująca łączliwość odpowiednio z liczebnikami głównymi i zbiorowymi. Zrezygnowano

z rozróżnienia między rodzajami p2 i p3. Rzeczownikom *plurale tantum* przypisano rodzaj jak w NKJP, jednak otrzymują one dodatkowy atrybut pt. W ten sposób na potrzeby uzgodnień rodzajowych rzeczowniki *plurale tantum* dzielą się na te, które łączą się z męskoosobowymi (m1) formami czasowników i przymiotników, oraz łączące się z innymi formami, zaklasyfikowane jako nijakie. Tak wprowadzone oznaczenia można odczytywać jako hierarchiczne symbole rodzajów: n:col dla Saloniowskiego n1, n:ncol dla n2, m1:pt dla p1, n:pt dla p2 i p3.

Przedstawiając to inaczej: kategoria rodzaju zostaje wyróżniona na podstawie łączliwości rzeczowników z formami przymiotnikowymi, a dodatkowe rozróżnienia postulowane przez Saloniego (z pominięciem p2/p3) zostają wyrażone za pomocą nowej kategorii, którą można nazwać *przyrodzajem*, przysługującej rzeczownikom i liczebnikom. Jej wartości to ncol, col i pt, przy czym oznaczenie pt należy traktować jako uszczegółowienie wartości col pomagające w wyszukiwaniu korpusowym rzeczowników *plurale tantum*. Z punktu widzenia składniowego pt jest równoważne col. Ten zbiór wartości może się wydawać niespójny, jednak wszystkie one odpowiadają rozróżnieniom rodzajowym w systemie Saloniego.

Jako wartości przyrodzaju w znacznikach form liczebnikowych nijakich dla odróżnienia form typu *dwa* od form typu *dwoje* stosuje się wyłącznie symbole ncol i col. Tak więc należy uznać, że wartość pt formy rzeczownikowej uzgadnia się z wartością col formy liczebnikowej. Symboli tych nie stosuje się dla form leksemów niewykazujących tego zróżnicowania (np. STO, PARĘ, DUŻO). Łączliwość rzeczowników m1 z liczebnikami zbiorowymi (*dwoje państwa*, *dwoje studentów*) musi być opisana w sposób specjalny na poziomie gramatycznym, nie wprowadzono bowiem szeregów form m1:col równokształtnych z n:col.

Przewaga tego systemu nad poprzednimi polega na tym, że dodatkowe oznaczenia są stosowane tylko tam, gdzie odpowiadają za łączliwość. Opatrywane nimi są jedynie formy rzeczowników i liczebników. Połączenia rzeczowników z przymiotnikami i czasownikami są opisywane prostszym pięcioklasowym systemem rodzajów.

1.6.4. Kategoria osoby

Kategoria osoby zdaje sprawę ze zróżnicowania fleksyjnego czasowników reprezentowanego na przykład przez formy *czytam/czytasz/czyta*, *przeczytam/przeczytasz/przeczytała*. Formy należące do tych trzech zbiorów, opisywanych jako osoba pierwsza pri, druga sec i trzecia ter, wykazują łączliwość z formami mianownikowymi pochodzącymi z rozłącznych zbiorów leksemów. Formy pierwszoosobowe łączą się wyłącznie z formami *ja/my*, drugoosobowe – *ty/wy*, podczas gdy trzecioosobowe wykazują szeroką łączliwość z mianownikami rzeczowników.

1.7. Klasy gramatyczne

Stosowany w niniejszej pracy podział leksemów na klasy gramatyczne opiera się na artykule Saloniego (1974a,b). Punktem wyjścia rozważań Saloniego jest stwierdzenie, że stosowany tradycyjnie w gramatykach polskich podział na tzw. części mowy nie spełnia warunków poprawnego podziału logicznego. Jest bowiem wynikiem zastosowania niespójnego zestawu kryteriów mieszających uwarunkowania fleksyjne, syntaktyczne i semantyczne. W szczególności tradycyjnie wyróżniana klasa zaimków ma charakter semantyczny i krzyżuje się z innymi wprowadzanymi rozróżnieniami (stąd klasy zaimków rzeczownych, przymiotnych, liczebnych, przysłownych).

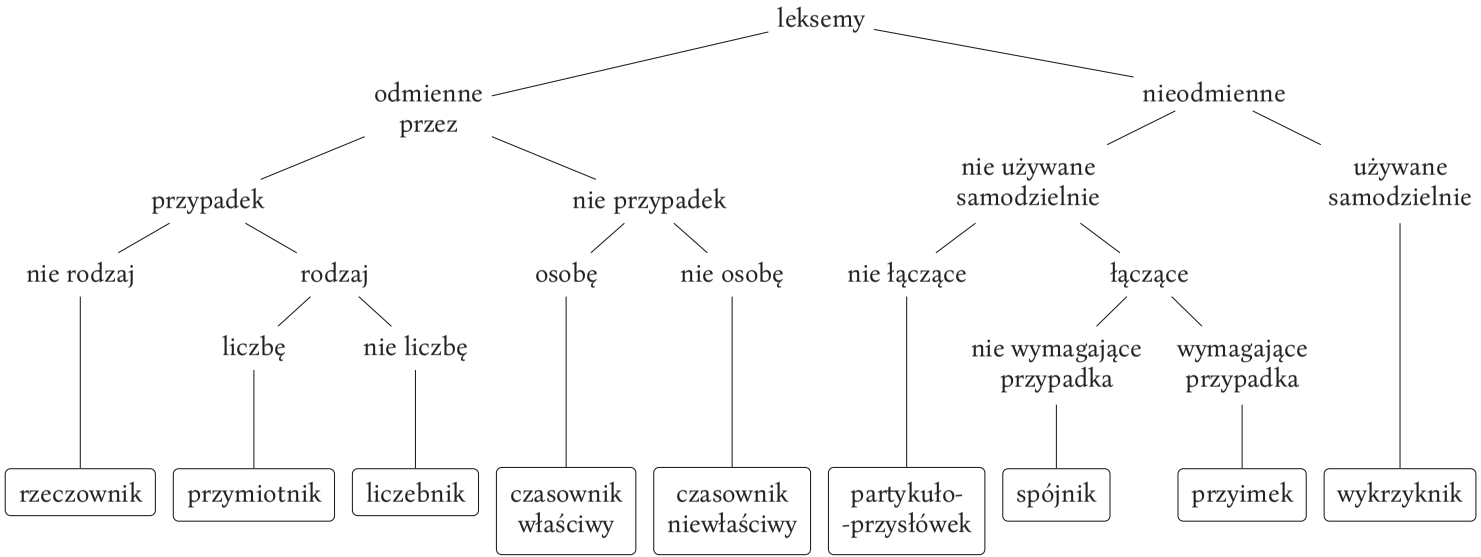
Saloni zauważa (za Tokarskim), że w procesie nauki szkolnej niektóre wyrazy pozostają poza wyodrębnionymi klasami. Taki stan jest jednak nie do przyjęcia, gdy chce się ująć całość materiału językowego w słowniku i powiązanej z nim gramatyce. Tym bardziej dotyczy to opisu tworzonego na potrzeby metod przetwarzania komputerowego.

Klasyfikacja leksemów polskich przedstawiona przez Saloniego (1974a,b) opiera się w pierwszej kolejności na kryteriach fleksyjnych (a więc odmienności leksemów poszczególnych klas przez poszczególne kategorie gramatyczne), a następnie – składniowych (przede wszystkim w odniesieniu do leksemów nieodmiennych). Przy tym autor deklaruje, że w wypadku sprzeczności między tymi kryteriami przyznaje prymat zasadzie fleksyjnej. Uzasadnieniem dla takiego uszeregowania kryteriów jest to, że fleksja odgrywa decydującą rolę w wyodrębnianiu leksemów, które mają być klasyfikowane (Saloni i Świdziński 2001, s. 95). Ponadto opis składniowy powinien być poprzedzony opisem fleksyjnym.

Klasyfikacja zaproponowana przez Saloniego (1974a,b) i powtórzona w Saloni i Świdziński (2001, rozdz. IV) została przedstawiona na rysunku 1.1. W podziale tym są dwie klasy wyróżnione negatywnie. Pierwszą jest klasa czasowników niewłaściwych wyróżniona jako leksemy odmienne, które nie odmieniają się ze względu na żadne kategorie wymienone na diagramie. Leksemy te odmieniają się mianowicie przez kategorię trybu i czasu. Niektóre z nich, jak TRZEBA, mają odmianę czysto analityczną (*trzeba by*, *trzeba było*, *będzie trzeba*), a więc syntetyczny wykładnik form tego leksemu jest tylko jeden, ale leksem zostaje umieszczony w grupie odmiennych (por. p. 1.7.2).

Drugą klasą wyróżnioną negatywnie są partykuło-przysłówki. W SGJP przyjęto podział tej klasy na partykuły i przysłówki nieodmienne nieodprzymiotnikowe na podstawie własności składniowych i tak też jest w prezentowanym tu opisie (por. p. 1.7.7).

W klasyfikacji z pracy Saloni i Świdziński (2001) konstrukcje typu *na schwał* i *po omacku* zostają uznane za jednolite formy fleksyjne leksemów NA SCHWAŁ i PO OMACKU. Ponieważ założono, że interpretowane fragmenty tekstu nie mogą zawierać odstępu, te konstrukcje muszą zostać zinterpretowane inaczej, mianowicie jako związki frazeologiczne, których jeden składnik jest



Rysunek 1.1. Klasyfikacja leksemów polskich według Saloniego (1974a,b)

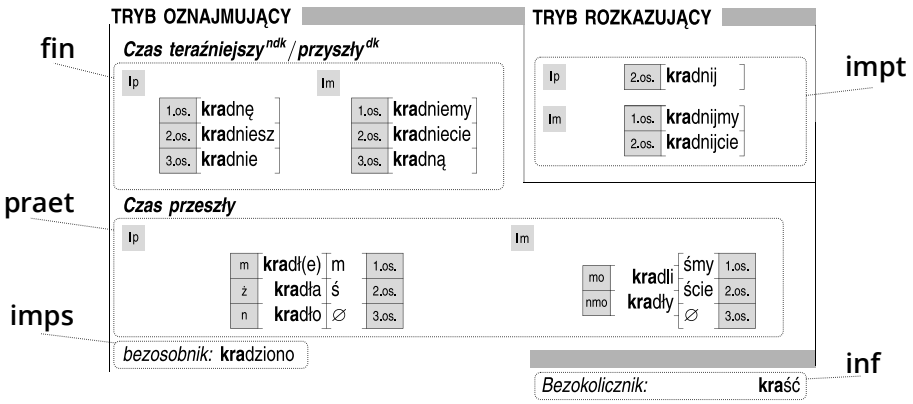
elementem niesamodzielnym ograniczonym do konkretnego kontekstu. Taka też interpretacja została przyjęta w SGJP.

Dokładniejsza charakterystyka poszczególnych klas gramatycznych zostanie przedstawiona w kolejnych punktach tego podrozdziału.

1.7.1. Leksemy i fleksemy

Saloni (1974a, s. 8) deklaruje, że leksemy są wyróżniane tak, aby uzyskać regularność opozycji w obrębie paradygmatu. Jednak dla pewnych części mowy niektóre z form obecnych w paradygmacie są zneutralizowane ze względu na niektóre kategorie. O ile na przykład kategorię rodzaju uznaje się za fleksyjną dla czasowników (widać to na przykład w formach czasu przeszłego *kradł*, *kradła*, *kradło*), to formy czasu teraźniejszego są zneutralizowane ze względu na rodzaj (*kradnie*), choć, jak zaznacza Saloni, są używane w kontekstach, w których wyrażenie kategorii rodzaju jest możliwe. Formami czasownika zneutralizowanymi ze względu na wszystkie kategorie są bezokolicznik, bezosobnik i imiesłowy przysłówkowe, które wypada zaliczyć do leksemu czasownikowego ze względu na ich regularność wobec pozostałych form (są one obecne w paradygmatach wszystkich lub nieomal wszystkich czasowników). Na rysunku 1.2 przedstawiono odpowiedni fragment paradygmatu czasownika według pracy Saloni (2007).

W przedstawionej tu koncepcji stosuje się dwustopniowe grupowanie form. Leksemy zostają wyróżnione według tych zasad, co w klasyfikacji Saloniego i w SGJP. Drugi poziom wprowadza podział leksemów na części, w obrębie których formy są zróżnicowane ze względu na te same kategorie fleksyjne. Odpowiednią jednostkę zaproponował Bień (1991), nazywając ją *fleksemem*. Bień przyjmuje, że fleksem jest zbiorem form fleksyjnych różniących się tylko wartościami kategorii liczby, przypadka, rodzaju i osoby. W niniejszym opisie,



Rysunek 1.2. Fragment paradygmatu odmiany czasownika KRAŚĆ (za: Saloni 2007) i jego podział na fleksemy

aby zmniejszyć liczbę różnych typów fleksemów, dopuszcza się odmienność w obrębie fleksemów również ze względu na inne kategorie. Cechą wyróżniającą fleksemów pozostaje jednolite (lub niemal jednolite) zróżnicowanie ze względu na właściwe im kategorie gramatyczne.

Tak więc leksemy traktowane są jako zbiory fleksemów, które z kolei są zbiorami form fleksyjnych. Leksem nieodmienny składa się zawsze z jednego fleksemu. Typ fleksemu determinuje jednoznacznie typ leksemu, do którego on należy, dlatego to typy fleksemów są używane jako pierwszy element charakterystyki form fleksyjnych podawanej w formie znaczników fleksyjnych.

Na rysunku 1.2 zaznaczono niektóre fleksemy czasownikowe i podano ich oznaczenia. Są to: fleksem nieprzeszły fin, przeszlik praet, rozkaźnik impt, bezosobnik imps, bezokolicznik inf.

Podział wszystkich typów leksemów na fleksemy i stosowane oznaczenia klas fleksemów można znaleźć w tabeli 1.3 na stronie 42. W tabeli 1.4 zestawiono kategorie gramatyczne istotne dla opisu fleksemów poszczególnych klas.

1.7.2. Czasowniki

Według klasyfikacji Saloniego czasowniki to leksemy odmienne, ale nie przez przypadek. Tak rozumiane czasowniki dzielą się na dwie podklasy: właściwe, których formy są zróżnicowane ze względu na osobę (np. (31)), oraz niewłaściwe, niewykazujące zróżnicowań osobowych (np. (32), (33)). Obu podklasom przypisuje się odmianę przez tryb i czas.

(31) – *Chodź*$_{\text{sec}}$, *pokażę*$_{\text{pri}}$ *ci, co* *jest*$_{\text{ter}}$ *do sprzątania.* [Skł.]

(32) *Trzeba* *tylko zgłosić działalność gospodarczą i zarejestrować się w kuratorium oświaty.* [Skł.]

(33) *Po chwili* *słychać* *było stamtąd dziewczęcy płacz.* [Skł.]

Czasowniki właściwe odmieniają się przez tryb, czas, osobę, liczbę i rodzaj. Formami osobowymi nazywa się takie ich formy, które dopuszczają obecność mianownikowego podmiotu nominalnego, do którego forma czasownika dostosowuje wartość osoby, liczby i rodzaju. Formy osobowe należą do szerszej klasy form finitywnych, czyli konstytuujących zdanie. Oprócz form osobowych klasa ta obejmuje bezosobnik (34) i odpowiednie formy czasowników niewłaściwych (zob. dalej).

(34) *Na trotyl* *natrafiono* *w miejscu, które wskazał radiesteta, zatrudniony przez rodzinę zaginionego.* [Skł.]

Formy niefinitywne to bezokolicznik i imiesłowy przysłówkowe (współczesny i uprzedni), np.:

(35) *Wycieczka musiała* *zmienić* *trasę marszu i zgubiła się.* [Skł.]

(36) *Dziergała je błyskawicznie,* *czytając* *przy tym trudne książki lub* *oglądając* *filmy.* [Skł.]

(37) *Cofała się także artyleria, wystrzelawszy wszystkie ładunki.* [Skł.]

Trzeba zwrócić uwagę na to, że bezokolicznik, będący formą niefinitywną (wskazuje na to jego łacińska nazwa *infinitivus*) bywa używany nietypowo – w funkcji finitywnej – jako centrum zdania (por. Świdziński 1992, s. 28) jak w zdaniach:

(38) *Podprawić przecierem, zielem angielskim, solą, pieprzem i czosnkiem.* [Skł.]

(39) *Czym zapełnić dziurę w budżecie?* [Skł.]

Odmiana przez czas i tryb

W ujęciu SGJP (Saloni *et al.* 2012, s. 93 i nast.) formom osobowym czasownika przysługują trzy wartości trybu: oznajmujący, rozkazujący i warunkowy. Zróżnicowanie czasowe form jest różne dla poszczególnych trybów. W trybie oznajmującym wyróżnia się czas przeszły, teraźniejszy, przyszły oraz, we współczesnym języku praktycznie nie używany, czas zaprzeszły. Ten ostatni ma postać czysto analityczną, tworzy się go, dodając do form czasu przeszłego posiłkową formę czasownika BYĆ, np. *przeczytał był*.

Tryb rozkazujący nie wykazuje zróżnicowań czasowych. Obejmuje on 3 formy osobowo-liczbowe, np. *weź, weźmy, weźcie*. Inne kombinacje osobowo-liczbowe odpowiadają konstrukcji z partykułą NIECH, np. *niech wezmę, niech weźmie, niech weźmiemy*.

Tryb warunkowy ma dwa warianty. Pierwszy, podstawowy, ma postać:

(40) *Chwaliłbyś ją.*

(41) *Byś ją chwalił.*

Rzadziej używany drugi wariant trybu warunkowego ma postać analityczną, złożoną z formy czasu przeszłego zestawionej z posiłkową formą warunkową czasownika BYĆ:

(42) *Byłbyś jej nie chwalił, gdybyś wiedział więcej.*

Ten drugi wariant trybu warunkowego bywa nazywany wariantem nierzeczywistym lub czasem przeszłym trybu warunkowego (Saloni 2007, s. 22).

Formę bezosobnika (np. *chwalono*) można przypisać do czasu przeszłego trybu oznajmującego. Możliwa jest też analityczna postać trybu warunkowego (*chwalono by*).

W prezentowanym systemie podlegają interpretacji wyłącznie formy syntetyczne. Przy takim ograniczeniu zróżnicowanie czasowe form jest widoczne wyłącznie w obrębie trybu oznajmującego. W związku z tym zdecydowano się nie reprezentować jawnie czasu ani trybu, wszystkie te rozróżnienia wyrażając przez przyporządkowanie do fleksemów.

Formy osobowe typowego czasownika właściwego podzielono na następujące fleksemy: fleksem nieprzeszły fin, którego interpretacja gramatyczna

Tabela 1.3. Klasy leksemów i ich rozbicie na fleksemy. W wypadku leksemów złożonych z tylko jednego fleksemu wspólna nazwa leksemu i fleksemu zajmuje dwie kolumny tabeli.

leksem	fleksem	ozn.
czasownik	forma nieprzeszła	fin
	forma przyszła czasownika BYĆ	bedzie
	aglutynant czasownika BYĆ	aglt
	pseudoimiesłów[1]	praet
	przeszlik[2]	praet
	forma trybu warunkowego[2]	cond
	rozkaźnik	impt
	bezosobnik	imps
	bezokolicznik	inf
	imiesłów przysłówkowy współczesny	pcon
	imiesłów przysłówkowy uprzedni	pant
	odsłownik	ger
	imiesłów przymiotnikowy czynny	pact
	imiesłów przymiotnikowy bierny	ppas
czasownik typu WINIEN (forma teraźniejsza)		winien
predykatyw		pred
rzeczownik	rzeczownik	subst
	forma deprecjatywna	depr
przymiotnik	przymiotnik	adj
	przymiotnik przyprzymiotnikowy	adja
	forma poprzyimkowa	adjp
przysłówek		adv
liczebnik		num
zaimek nietrzecioosobowy		ppron12
zaimek trzecioosobowy		ppron3
zaimek SIEBIE		siebie
przyimek		prep
spójnik współrzędny		conj
spójnik podrzędny		comp
wykrzyknik		interj
partykuła		part
człon frazeologizmu		frag

[1] Czas przeszły interpretowany jako konstrukcja.
[2] Czas przeszły interpretowany jako jednolita forma.

Tabela 1.4. Kategorie przysługujące poszczególnym klasom fleksemów. Symbol ⊕ oznacza, że dla danej klasy dana kategoria jest fleksyjna. Symbol ⊙ oznacza, że pewna wartość ustalona kategorii przysługuje wszystkim formom danego fleksemu.

ozn.	liczba	przypadek	rodzaj	osoba	stopień	aspekt	zanegowanie	akcentowość	poprzyimkowość	akomodacyjność	przyrodzaj	aglutynacyjność	wokaliczność
fin	⊕			⊕		⊙							
bedzie	⊕			⊕		⊙							
aglt	⊕			⊕		⊙							⊕
praet[1]	⊕		⊕			⊙						⊕	
praet[2]	⊕		⊕	⊕		⊙							
cond[2]	⊕		⊕	⊕		⊙							
impt	⊕			⊕		⊙							
imps						⊙							
inf						⊙							
pcon						⊙							
pant						⊙							
ger	⊕	⊕	⊙			⊙	⊕						
pact	⊕	⊕	⊕			⊙	⊕						
ppas	⊕	⊕	⊕			⊙	⊕						
winien	⊕		⊕			⊙							
pred													
subst	⊕	⊕	⊙								⊙		
depr	⊙	⊕	⊙										
adj	⊕	⊕	⊕		⊕								
adja													
adjp		⊕											
adv					⊕								
num	⊙	⊕	⊕							⊕	⊕		
ppron12	⊙	⊕	⊕	⊙				⊕					
ppron3	⊕	⊕	⊕	⊙				⊕	⊕				
siebie		⊕											
prep		⊙											⊕
conj													
comp													
interj													
part													
frag													

zostanie omówiona w ustępie poświęconym kategorii aspektu; przeszlik praet, reprezentujący czas przeszły trybu oznajmującego; rozkaźnik impt i, w jednym z wariantów znakowania, fleksem warunkowy cond. Do klasy form finitywnych należy oprócz wymienionych jeszcze bezosobnik imps.

Kategoria aspektu

Kategoria ta klasyfikuje leksemy czasownikowe na dokonane perf i niedokonane imperf. Od wartości aspektu zależy zasób form leksemu czasownikowego. W trybie oznajmującym czasowniki niedokonane tworzą syntetyczne formy czasu teraźniejszego (*czyta*), a ich czas przyszły ma postać analityczną (*będzie czytać/czytał/czytała/czytało*); czasowniki dokonane tworzą syntetyczne formy czasu przyszłego (*przeczyta*), czas teraźniejszy dla nich nie istnieje. Czasowniki obu aspektów różnią się też zasobem tworzonych form imiesłowowych. Ilustruje to tabela 1.5 (wzorowana na Saloni *et al.* 2012, s. 88).

Tabela 1.5. Zasób form czasownikowych w zależności od aspektu

	aspekt	
	niedokonany	dokonany
	tryb oznajmujący	
czas teraźniejszy	*czyta*	–
czas przyszły	*będzie czytać* *będzie czytał(a/o)*	*przeczyta*
	imiesłowy przysłówkowe	
współczesny	*czytając*	–
uprzedni	–	*przeczytawszy*
	imiesłowy przymiotnikowe	
czynny	*czytający*	–
bierny	*czytany*	*przeczytany*

W prezentowanym opisie przyjęto oznaczanie wymienionych w tabeli fleksyjnie paralelnych syntetycznych form trybu oznajmującego *czyta/przeczyta* wspólnym znacznikiem oznaczającym formę nieprzeszłą fin (por. Saloni *et al.* 2012, s. 86, przypis 2). Forma ta jest interpretowana jako forma czasu teraźniejszego dla czasowników niedokonanych i forma czasu przyszłego dla czasowników dokonanych.

Z tego schematu wyłamuje się tylko jeden czasownik – BYĆ. Jako jedyny czasownik niedokonany ma on oprócz formy czasu teraźniejszego (*jestem, jesteś, ..., jesteście, są*) również nieanalityczne formy czasu przyszłego (*będę, będziesz, ..., będziecie, będą*). Dla tych ostatnich wprowadzono specjalny fleksem przyszłych form finitywnych czasownika BYĆ oznaczony bedzie.

Warto podkreślić, że aspekt nie jest kategorią fleksyjną czasowników. Czasowniki niedokonane i dokonane często daje się połączyć w naturalne pary czasowników różniących się aspektem (CZYTAĆ – PRZECZYTAĆ), zależność

ta nie ma jednak regularności właściwej fleksji. Zdarza się bowiem, że jakiś czasownik nie ma odpowiednika (np. WYZIERAĆ), jak też, że dla jednego czasownika niedokonanego można wskazać wielu kandydatów na odpowiednik dokonany, którzy wnoszą różne naddatki znaczeniowe (CIOSAĆ – WYCIOSAĆ, NACIOSAĆ, UCIOSAĆ).

Wpływ składniowy aspektu dotyczy nielicznych czasowników wymagających frazy bezokolicznikowej, jest jednak wyrazisty. Na przykład czasownik ZACZYNAĆ dopuszcza przy sobie jedynie bezokoliczniki czasowników niedokonanych:

(43) *Jan zaczyna czytać książkę.*
(44) **Jan zaczyna przeczytać książkę.*

Formy czasu przeszłego

Czas przeszły w języku polskim może być realizowany w dwóch wariantach szyku:

(45) *Ja to czytałem.*
(46) *Ja m to czytał.*

Nieciągły czas przeszły w zdaniu (46) brzmi archaicznie, jednak ta sama konstrukcja użyta po spójniku ŻE daje zdanie nienacechowane:

(47) *Nie wiedziałem, że to czytaliście.*
(48) *Nie wiedziałem, że ście to czytali.*

Co więcej, istnieją konteksty (np. po spójniku GDYBY), w których wariant nieciągły jest jedynym możliwym:

(49) **Przyszedłbym, gdyby to czytaliście.*
(50) *Przyszedłbym, gdyby ście to czytali.*

W prezentowanej tu koncepcji przyjęto, że słowa takie jak *gdybyście* składają się z dwóch segmentów (*gdyby* i *ście*). Pierwszy z nich w oczywisty sposób jest interpretowany jako spójnik, dobrej analogii dla interpretacji drugiego segmentu dostarcza zaś następujące wypowiedzenie:

(51) *Świniam.*

W wypowiedzeniu tym występują wykładniki dwóch form fleksyjnych: formy leksemu rzeczownikowego ŚWINIA i skróconej (aglutynacyjnej) formy czasu teraźniejszego leksemu czasownikowego BYĆ (Saloni i Świdziński 2001; Tokarski 1973).

Na tej podstawie przyjęto, że w zdaniu (47) czas przeszły jest wyrażony za pomocą posiłkowej formy aglutynacyjnej *ście* i formy *czytali*, którą Saloni (2001) nazywa *pseudoimiesłowem*. W konstrukcji tej aglutynant niesie charakterystykę osobowo-liczbową, a pseudoimiesłów rodzajowo-liczbową.

Dla konsekwencji przyjęto, że czas przeszły czasowników zawsze jest analityczny i również w zdaniu (48) ma postać konstrukcji złożonej z formy pseudoimiesłowu i formy aglutynantu (por. Saloni 1974b, s. 95). Zgodnie z tymi założeniami słowo *widziałem* jest interpretowane jako:

0	1	*widział*	WIDZIEĆ	praet:sg:m1.m2.m3:imperf
1	2	*em*	BYĆ	aglt:sg:pri:imperf:wok

W interpretacji tej możliwe jest operowanie prostszym systemem znaczników: przy pseudoimiesłowie nie trzeba notować wartości osoby, przy aglutynancie zaś – rodzaju.

Opis składniowy bazujący na przedstawionym tu znakowaniu musi zdawać sprawę z tego, że w funkcji czasu przeszłego czasownika może wystąpić sam pseudoimiesłów (wtedy konstrukcja ma wartość ter kategorii osoby) lub konstrukcja złożona z pseudoimiesłowu i formy aglutynacyjnej (wtedy wartość osoby konstrukcji jest równa wartości osoby aglutynantu). Warto zwrócić uwagę, że pseudoimiesłów oprócz tworzenia czasu przeszłego, może także wystąpić jako składowa czasu przyszłego (*będę widziała*) i trybu warunkowego (*widziała·by·m*, zob. dalej). Jest to dodatkowym uzasadnieniem dla nieopatrywania pseudoimiesłowu wartością czasu.

System w takiej postaci został wdrożony przy znakowaniu Narodowego Korpusu Języka Polskiego[12]; niestety nie można powiedzieć, że został powszechnie zaakceptowany. Jakkolwiek taka interpretacja form czasu przeszłego jest językoznawczo uzasadniona, to dzielenie pewnych słów, które mogłyby być interpretowane w całości, utrudnia przetwarzanie w systemach niezawierających komponentu interpretującego jednostki wieloczłonowe. Dlatego w wersji 2 programu Morfeusz zapewniono możliwość wyboru. Na życzenie użytkownika syntetyczne formy czasu przeszłego mogą być interpretowane w całości. W takim wariancie fleksem praet obejmuje nie tylko pseudoimiesłów, ale całe formy przeszlików, ich opis zawiera więc osobę, rodzaj i liczbę:

widziałem	praet:sg:m1.m2.m3:pri:imperf
widziałeś	praet:sg:m1.m2.m3:sec:imperf
widział	praet:sg:m1.m2.m3:ter:imperf
widziałam	praet:sg:f:pri:imperf
…	…

Również w tym wariancie znakowania analityczny czas przeszły jest interpretowany jako złożony z formy aglt i formy praet (która ma jednak przypisaną wartość 3 osoby). Różnice w atrybutach znaczników stosowanych w obu wariantach uwzględniono w tabeli 1.4.

[12] Trzeba przyznać, że nawet w tym zastosowaniu spowodował pewne kłopoty: w wyniku takiego znakowania liczba segmentów w korpusie zaczęła się znacząco różnić od liczby słów, więc podawanie wielkości korpusu wymagało zawsze tłumaczenia sposobu liczenia.

Kategoria aglutynacyjności

W wypadku nielicznych czasowników, gdy pseudoimiesłów występuje z aglutynatem, następuje alternacja (fonologiczna lub ortograficzna), z której trzeba zdać sprawę w znakowaniu (np. GNIEŚĆ: *gniótł*, ale *gniotł·em*). Te formy pseudoimiesłowu, w których ujawnia się to zróżnicowanie, są opatrywane wartością kategorii *aglutynacyjności*: wartość agl dla postaci występującej z aglutynantem (*gniotł*) i nagl – dla postaci bez aglutynantu. Kategoria ta nie jest stosowana w wariancie znakowania przeszlików jako jednolitych form.

Kategoria wokaliczności

Formy aglutynacyjne czasownika BYĆ mają dwie postaci różniące się obecnością inicjalnej litery *e*, np. *·ście* i *·eście*. Postaci te są odróżniane wartością kategorii *wokaliczności*: forma wokaliczna wok aglutynantu jest używana po formach kończących się spółgłoską (np. *czytał·em*, *gdzież·eś*), forma niewokaliczna nwok jest wariantem posamogłoskowym (*czytała·m*, *gdyby·ś*).

Ta sama kategoria służy także do opisu analogicznego zjawiska dotyczącego form niektórych przyimków (np. *z/ze*, *przed/przede*), w tym wypadku jednak *e* pojawia się na końcu formy i uwikłanie dotyczy formy występującej po danej.

Formy trybu warunkowego

Konstrukcje wyrażające tryb warunkowy są interpretowane analogicznie do konstrukcji czasu przeszłego w trybie oznajmującym. Tak więc w domyślnym wariancie znakowania Morfeusz interpretuje tryb warunkowy jako wyrażony za pomocą trzech form: pseudoimiesłowu, partykuły BY i aglutynantu, niezależnie od tego, czy odpowiednie segmenty są ciągłe (*widział·by·m*), czy nie (*by·m widział*). Partykuła BY jest traktowana jako wyznacznik trybu warunkowego, nie ma zaś żadnego znacznika odpowiadającego temu trybowi. Znakowanie wygląda następująco:

0	1	*widział*	WIDZIEĆ	praet:sg:m1.m2.m3:imperf
1	2	*by*	BY	part
2	3	*m*	BYĆ	aglt:sg:pri:imperf:nwok

Taki sposób znakowania pozwala interpretować nie tylko nieciągły wariant szyku, ale i nietypowe konstrukcje jak np.:

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

Pierwsze słowo jest interpretowane jako cztery segmenty: *potrzebował·że·by·ś*; jednym z nich jest wykładnik formy leksemu wzmacniającego ŻE.

Biorąc pod uwagę, że w trybie warunkowym aglutynant musi następować bezpośrednio po cząstce *by*, być może byłoby bardziej naturalne wprowadzić

aglutynant warunkowy o wykładnikach *bym*, *byś*, *by*, *byśmy*, *byście*. Pozwoliłoby to zmniejszyć liczbę segmentów, na które dzielone są słowa, ale wymagałoby dodania kolejnego typu fleksemów.

W wariancie analizy, w którym ciągłe realizacje czasu przeszłego są interpretowane jako jednolite formy, podobnie dzieje się z odpowiednimi realizacjami trybu warunkowego. Wymagało to wprowadzenia w Morfeuszu 2 fleksemu warunkowego cond:

widziałbym	cond:sg:m1.m2.m3:pri:imperf
widziałbyś	cond:sg:m1.m2.m3:sec:imperf
widziałby	cond:sg:m1.m2.m3:ter:imperf
widziałbym	cond:sg:f:pri:imperf
…	…

Fleksemy czasownikowe jednoelementowe

W obrębie leksemu czasownikowego jest kilka form niewrażliwych na żadne kategorie fleksyjne, reprezentowanych jako jednoelementowe fleksemy. Dotychczas wymieniono jeden z nich – *bezosobnik* imps, czyli formę bezosobową typu *widziano*, *myto*.

Pozostałe fleksemy jednoelementowe to: *bezokolicznik* inf (np. *widzieć*, *zobaczyć*), *imiesłów przysłówkowy współczesny* pcon (*widząc*) i *imiesłów przysłówkowy uprzedni* pant (*zobaczywszy*).

Imiesłowy przymiotnikowe i odsłowniki

Do szeroko rozumianego leksemu czasownikowego zaliczany bywa także imiesłów przymiotnikowy czynny (*widzący*), imiesłów przymiotnikowy bierny (*widziany*, *zobaczony*), a przez niektórych także odsłownik (rzeczownik odsłowny, czyli odczasownikowy, zwany też gerundium, *widzenie*, *zobaczenie*). Saloni i Świdziński (2001, s. 100) postulują niezaliczanie imiesłowów przymiotnikowych i odsłowników do leksemu czasownikowego, aby wyraźnie oddzielić koniugację od deklinacji. Ponadto piszą, że zaliczenie imiesłowu biernego do leksemu czasownikowego wymagałoby wprowadzenia kategorii strony. Również w SGJP są one uznawane za osobne leksemy stanowiące regularne derywaty odczasownikowe.

Jednak gerundia i imiesłowy dziedziczą w sposób systematyczny wymagania walencyjne czasowników (por. p. 3.1.10). Dlatego w opisie ukierunkowanym składniowo wygodnie jest zaliczyć je do leksemu czasownikowego – wówczas jedno hasło słownika walencyjnego będzie mogło opisywać własności wszystkich tych form. Ponadto pojęcie fleksemu pozwala uniknąć wprowadzania kategorii strony. Dlatego w przyjętym w tej pracy opisie fleksji do leksemu czasownikowego należą trzy fleksemy deklinacyjne: imiesłów przymiotnikowy czynny pact, bierny ppas i odsłownik ger.

Kolejnym przymiotnikowym derywatem od czasownika jest tzw. imiesłów przeszły przymiotnikowy (Saloni *et al.* 2012, s. 108). Większość jego form jest tworzona poprzez dodanie do pseudoimiesłowu końcówki przypadkowo-rodzajowo-liczbowej, np. *poszarzał – poszarzały, poszarzała, poszarzałe, poszarzałego, poszarzałej* itd. Derywat ten jest tworzony od nielicznych czasowników, wyłącznie dokonanych i od czasownika BYĆ. Ze względu na niską regularność został uznany za osobny leksem przymiotnikowy niepowiązany z czasownikiem.

Rzeczownikowe derywaty, dla których zależność od czasownika jest mniej wyrazista niż dla odsłowników, są opisywane jako samodzielne rzeczowniki. Tak więc PROSZENIE jest odsłownikiem hasłowanym do PROSIĆ, natomiast PROŚBA jest osobnym rzeczownikiem. Wymagania walencyjne takich derywatów nie są regularne względem czasownika i wymagają osobnego opisania. Na przykład rzeczownik PROŚBA ma wymaganie *o coś*, ale nie *kogoś*, w odróżnieniu od czasownika PROSIĆ i gerundium PROSZENIE (por. Saloni i Świdziński 2001, s. 186).

Kategoria zanegowania

Kategoria fleksyjna *zanegowania* przysługuje odsłownikom i imiesłowom przymiotnikowym czynnym i biernym. Wprowadzenie jej było konieczne w związku z zasadą ortograficzną łącznej pisowni cząstki *nie·* z formami tych wyrazów. Skoro na przykład formy *celebrowanie* i *niecelebrowanie* należą do tego samego leksemu, konieczna jest cecha odróżniająca. Wartościami zanegowania są aff – brak cząstki *nie·* i neg – jej obecność.

Kategoria zanegowania nie dotyczy przymiotników ani przysłówków, ponieważ na przykład ŁADNY i NIEŁADNY są interpretowane jako osobne leksemy. Nie ma też ona związku z negacją zdaniową (por. p. 3.1.4).

Strona bierna

W SGJP (Saloni *et al.* 2012, s. 109) i przedstawionym tu opisie nie uznaje się istnienia kategorii fleksyjnej strony. W języku polskim (w odróżnieniu od łaciny) nie ma syntetycznych wykładników strony biernej. Strona bierna jest zawsze konstrukcją tworzoną z różnymi czasownikami nadrzędnymi (co najmniej BYĆ, ZOSTAĆ, BYWAĆ) i z użyciem różnych ich form, np. niefinitywnych: *będąc ciągle chwalonym, zostawszy pochwalonym przez ojca, bycie chwalonym, lubi być chwalonym*.

Strona bierna została więc uznana za regularną konstrukcję składniową tworzoną z udziałem imiesłowu przymiotnikowego biernego. To, czy dla danego czasownika jest ona możliwa, jest notowane w słowniku walencyjnym (por. p. 3.1.3).

Czasowniki odmienne nietypowo

Jest w polszczyźnie kilka leksemów (Saloni *et al.* 2012, s. 113), które nie przypominają struktury typowego leksemu czasownikowego, ale pełnią w zdaniu tę samą funkcję co formy finitywne czasowników. Należą do tej grupy leksemy POWINIEN, WINIEN, RAD i, być może, kilka innych. Ich odmiana jest nietypowa: mają one mianowicie czas teraźniejszy w trybie oznajmującym o postaci *powinienem*, *powinieneś*, *powinien*, *powinnam*, *powinnaś* itd. Jest on wyrażany w sposób podobny do czasu przeszłego innych czasowników. Dlatego słowa te są analogicznie interpretowane w analizatorze Morfeusz. W celu ich interpretacji wprowadzono osobny fleksem i leksem oznaczony winien. Słowo *powinienem* jest interpretowane jako forma *powinien* fleksemu winien i aglutynant *em*.

Leksemy te nie mają form nieosobowych, w szczególności bezokolicznika, dlatego są również nietypowo lematyzowane. W postaci konstrukcji można wyrazić czas przeszły (*powinienem był*, *powinnaś była*) i tryb warunkowy (*powinien bym*, *powinna byś*).

Czasowniki niewłaściwe

Cechą definicyjną czasowników niewłaściwych jest nieodmienność przez osobę. Pociąga ona za sobą niemożliwość użycia ich form jako centrum zdania z uzgadniającą się co do liczby, osoby i rodzaju mianownikową formą rzeczownika (podmiotem-mianownikiem).

Ze względu na sposób odmiany dzielą się one na dwie grupy. Do pierwszej (Saloni *et al.* 2012, s. 117) należą czasowniki posiadające tylko jedną formę syntetyczną, czyli powierzchniowo nieodmienne jak TRZEBA. Zostały one opisane przez zaliczenie do fleksemu nieodmiennego pred (predykatyw).

Druga grupa (Saloni *et al.* 2012, s. 115) obejmuje czasowniki posiadające trzy formy syntetyczne, jak BRAKOWAĆ (*brakuje*, *brakowało*, *brakować*)[13]. Oto przykłady ich użycia:

(53) *Jednocześnie brakuje im czasu na zabiegi pielęgnacyjne.* [Skł.]

(54) *Lekarze zdecydowali się na operację, ale brakowało dawcy.* [Skł.]

(55) *Tego ołtarza to mi tak będzie brakowało...* [NKJP300]

(56) *Jeszcze tego nieszczęścia brakowałoby mi na stare lata.* [NKJP300]

(57) *Cóż by takim damom mogło brakować oprócz smaku pieczonych w ognisku kartofli.* [NKJP300]

Ponieważ czasowniki z tej grupy bywają homonimiczne z czasownikami właściwymi[14], ich formy otrzymują takie same znaczniki, jak formy odpowiednich

[13] Trzeba wyróżnić cztery formy syntetyczne tych czasowników, jeżeli tryb warunkowy jest traktowany jako jednolita forma (np. *brakowałoby*).

[14] Na przykład właściwy czasownik BRAKOWAĆ oznacza czynność wykonywaną przez brakarza.

czasowników właściwych. Zakłada się tym samym, że informację o możliwości użyć niewłaściwych pozyskuje się ze słownika walencyjnego (są one możliwe, gdy istnieje schemat składniowy niezawierający podmiotu nominalnego, por. p. 3.1.3).

Do form nietypowych należy *to* w wypowiedzeniach (58) i (59). Analogiczne zdanie w czasie przeszłym ma postać (61). W czasie teraźniejszym możliwe jest dodanie formy *jest* jak w zdaniu (60). Forma *to* w dwóch ostatnich zdaniach nie jest formą rzeczownikową, bo zawierają one oprócz niej po dwie wymagane frazy nominalne.

(58) *Sport* *to* *zdrowie.*

(59) *Dyskrecja, takt i skromność* *to* *podstawowe pojęcia bon tonu.* [Skł.]

(60) *PZN* *to jest* *zrzeszenie wszystkich podmiotów, które działają na terenie kraju.* [Skł.]

(61) *Pani uważa, że śmierć Dunina* *to było* *morderstwo...?* [NKJP300]

(62) *Bo talmudysta* *to nie był* *zawód.* [NKJP300]

(63) *Lato* *to nie* *czas dla słów.* [NKJP300]

W SGJP przyjęto, że omawiana jednostka jest specyficzną partykułą. W konsekwencji wypowiedzenia (58) i (59) nie zawierają finitywnej formy czasownika, co jest niewygodne dla implementacji komputerowej. Aby uniknąć tego problemu, w prezentowanym tu opisie przyjęto za NKJP (Przepiórkowski *et al.* 2012, s. 76), że *to* może być formą finitywną. W tym celu wprowadzono leksem TO sklasyfikowany jako predykatyw o analitycznych formach czasu teraźniejszego *to* i *to jest*, czasu przeszłego *to był/a/o*, przyszłego *to będzie* i trybu warunkowego *to by był/a/o*. Problematyczny dla takiego zaklasyfikowania leksemu TO jest fakt, że zachowuje się on jak czasownik właściwy, dopuszczając podmiot mianownikowy. O obecności podmiotu świadczą uzgodnienia rodzajowe w zdaniach (61) i (62) – podmiotem jest fraza nominalna mianownikowa stojąca po formie *to*. Nietypowa jest też negacja: partykuła NIE stoi w tym wypadku w postpozycji: *to nie* (przykłady (63) i (62)).

1.7.3. Rzeczowniki

Rzeczowniki, zgodnie z klasyfikacją Saloniego, są odmienne przez przypadek i mają ustaloną wartość kategorii rodzaju. Większość leksemów rzeczownikowych jest też odmienna przez liczbę. Wyjątek stanowią leksemy *plurale tantum* (por. p. 1.6.1), które traktowane są jako defektywne, posiadające jedynie formy liczby mnogiej. W SGJP są też notowane nieliczne leksemy rzeczownikowe nieposiadające form liczby mnogiej. Są to wyłącznie tzw. zaimki rzeczowne takie jak KTO, CO, KTOŚ, COŚ, KTÓŻ, NIKT. Nie zostały zaliczone do tej grupy rzeczowniki typu PIERZE, STUDENTERIA czy SZCZEROŚĆ, których da się użyć w formie mnogiej (Saloni *et al.* 2012, s. 33):

(64) *Daj mi spokój z tymi swoimi* *pierzami/studenteriami/szczerościami*.

Deprecjatywność

Dla rzeczowników rodzaju m1 można wskazać dwie formy mianownika i wołacza liczby mnogiej, różniące się łączliwością z formami czasownikowymi i przymiotnikowymi:

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

Według Saloniego (1988) takie pary form istnieją dla każdego rzeczownika rodzaju m1, choć czasami ich wykładniki są identyczne (np. KOLEJARZ), a czasami któraś z form jest tylko potencjalna (np. niedeprecjatywna forma rzeczownika CHAM).

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

(66) **Przyszli głupi profesory.*

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

Innym rozwiązaniem mogłoby więc być wyniesienie form deprecjatywnych do osobnych leksemów (defektywnych), którym przypisany byłby rodzaj m2. Taki opis byłby jednak w sprzeczności z seryjnością występowania form deprecjatywnych dla rzeczowników m1. Bień (1991) proponuje wprowadzenie w obrębie leksemu rzeczownikowego dwóch fleksemów różniących się rodzajem. Powoduje to pewien problem teoretyczny, oznacza bowiem, że w obrębie leksemu rzeczownikowego współwystępują formy różniące się rodzajem (choć w obrębie każdego fleksemu rodzaj jest ustalony).

Ze względów praktycznych to rozwiązanie zostało przyjęte w prezentowanym tu opisie. Tak więc oprócz podstawowego fleksemu rzeczownikowego subst wprowadzono fleksem rzeczownikowy deprecjatywny depr, występujący tylko dla rzeczowników m1. Formy fleksemu deprecjatywnego mają przypisaną wartość rodzaju m2, co zdaje sprawę z ich własności składniowych.

Uniforemność

Dla niektórych rzeczowników rodzaju żeńskiego istnieją dwa możliwe wykładniki dopełniacza liczby mnogiej, np. *kropli/kropel*, *głuszy/głusz*, *funkcji/funkcyj*. W SGJP uznano (Saloni *et al.* 2012, p. 3.1.5, s. 43), że zróżnicowanie to nie jest przejawem żadnej kategorii gramatycznej, a więc że z punktu widzenia grama-

tycznego warianty te są swobodne[15]. Jeden z wariantów jest zawsze synkretyczny z inną formą rzeczownika – jest on w SGJP oznaczany kwalifikatorem *hom.*. Drugi wariant, oznaczany *char.*, jest charakterystyczny dla tej formy rzeczownika. Warianty często różnią się nacechowaniem stylistycznym, w szczególności dla części rzeczowników forma charakterystyczna jest archaiczna (np. *funkcyj, armij*). Te wypadki są oznaczane dodatkowym kwalifikatorem *arch.* Wymienione kwalifikatory są częścią interpretacji generowanych przez program Morfeusz 2.

Rzeczowniki wielorodzajowe

Co prawda rodzaj gramatyczny rzeczownika uznaje się za ustalony dla danego leksemu, zdarzają się jednak wypadki rozchwiania rodzajowego. Na przykład następujące zdania:

(67) *Zjadł smaczny kotlet.*
(68) *Zjadł smacznego kotleta.*

wskazują, że biernik rzeczownika KOTLET może być równy mianownikowi lub dopełniaczowi (to drugie użycie jest potoczne). Wynikają z tego sprzeczne wskazówki co do rodzaju tego rzeczownika: m3 albo m2. Trzeba by więc uznać, że są to dwa rzeczowniki różnych rodzajów, które różnią się wyłącznie formą biernika liczby pojedynczej. Jest to jednak rozwiązanie bardzo niepraktyczne, bo dla wszystkich pozostałych form nie ma wskazówek gramatycznych, do którego z leksemów należy dane wystąpienie.

W SGJP w takich wypadkach wprowadzane jest tylko jedno hasło z informacją o dwóch możliwych rodzajach (Saloni *et al.* 2012, s. 39). Takie rozwiązanie zostało też przyjęte w Morfeuszu: leksem KOTLET zawiera formy należące do dwóch różnych podrodzajów męskich.

Warto zaznaczyć, że rozwiązanie to jest stosowane w wypadku rozchwiania gramatycznego leksemu, którego denotacja jest ustalona (w przykładzie niezależnie od rodzaju chodzi o ten sam obiekt – kotlet) i który jest opisany jako jedno hasło w SGJP. Jeżeli denotacje są różne, postulowane są osobne leksemy, np. SĘDZIA:S1 m1 – mężczyzna wyrokujący w sprawach spornych i SĘDZIA:S2 f – kobieta wyrokująca w sprawach spornych. Pary tego rodzaju nie są (niestety – z punktu widzenia przetwarzania komputerowego) notowane systematycznie w SGJP.

Nazwy własne

SGJP zawiera przybliżoną klasyfikację nazw własnych (Saloni *et al.* 2012, p. 3.3.2, s. 48), wyróżniającą takie klasy, jak: nazwy osób (ze wskazaniem imion,

[15] Ściślej rzecz biorąc, w pierwszym wydaniu słownika postulowana była kategoria gramatyczna uniforemności, jednak zrezygnowano z niej w następnych wydaniach.

nazwisk, pseudonimów, przydomków), nazwy geograficzne, nazwy firm itp. Klasyfikacja ta jest dostępna w wynikach Morfeusza 2. Co prawda operuje ona zbyt ogólnymi klasami, aby posłużyć jako podstawa konstrukcji reprezentacji semantycznej zdania, jednak niektóre jej elementy, jak wyróżnienie imion i nazwisk mogą być bardzo użyteczne przy przetwarzaniu składniowym.

Słownik nie zawiera nazw wieloczłonowych, więc w konsekwencji nazwa taka jak *Nowy Sącz* jest interpretowana jako złożona z dwóch form fleksyjnych: przymiotnika NOWY i rzeczownika SĄCZ, który jest notowany w słowniku i oznaczony jako nazwa geograficzna.

1.7.4. Zaimki osobowe

Klasyfikacja leksemów Saloniego nie wprowadza pojęcia zaimka. Zaimki są bowiem tradycyjnie wyróżniane na podstawie mieszanych kryteriów, głównie semantycznych, które są ortogonalne do cech uwzględnianych w tej klasyfikacji. Gdyby chcieć je wprowadzić, trzeba by wyróżnić podklasę zaimków w większości z klas wyróżnionych na podstawie klasyfikacji Saloniego. W opisie fleksyjnym nie ma potrzeby, aby tak uczynić[16].

W proponowanym tu opisie uznano jednak za konieczne wyróżnienie osobnej klasy zaimków osobowych. Obejmuje ona kilka leksemów, które nie pasują do klasy rzeczowników, ponieważ nie mają ustalonego rodzaju. Ponadto części z nich można przypisać wartość osoby różną od trzeciej, która jest typowa dla rzeczowników (rzeczownikom można przypisać wartość osoby w tym sensie, że mianownikowe formy rzeczowników łączą się jako podmioty z trzecioosobowymi formami czasowników).

Zaimki osobowe zostały opisane za pomocą dwóch klas leksemów (i fleksemów): *zaimków nietrzecioosobowych* ppron12 i *zaimków trzecioosobowych* ppron3. Pierwsza z klas zawiera cztery leksemy: JA, TY, MY i WY. Przyjęto, że fleksemy tej klasy są odmienne przez przypadek, rodzaj (mimo że zachodzi pełna neutralizacja tej kategorii) i akcentowość. Klasa zaimków trzecioosobowych zawiera tylko jeden leksem o postaci hasłowej ON, odmienny przez liczbę, przypadek, rodzaj, akcentowość i poprzyimkowość. Leksemy i fleksemy obu tych klas mają przypisaną słownikową wartość liczby i osoby.

Jeszcze jedną klasę specjalną, stworzoną w celu opisania wysoce nietypowego leksemu, stanowi zaimek SIEBIE. Leksem ten ma nieustalony rodzaj, co jest przeciwwskazaniem do uznania go za rzeczownik. Klasa ta obejmuje tylko jeden leksem zawierający jeden fleksem odmienny tylko przez przypadek i w dodatku defektywny (pozbawiony mianownika i wołacza). Należą do niego mianowicie formy o wykładnikach *siebie*, *sobie*, *sobą* (Saloni i Świdziński 2001,

[16] W analizatorze Świgra 2 stosowane jest za GFJP pojęcie zaimków do wyróżnienia leksemów o własnościach nietypowych, wymagających osobnego opisu na potrzeby składni. Na przykład spośród rzeczowników wydzielono te leksemy, które wnoszą do konstrukcji cechę pytajności. Jest to jednak zabieg o charakterze czysto technicznym i z tak rozumianych zaimków można by zrezygnować.

s. 90). W SGJP do leksemu tego zaliczono także dopełniaczową i biernikową formę o kształcie *się*. Segment *się* może też wg SGJP być wykładnikiem mianownika defektywnego rzeczownika nijakiego występującego w zdaniach typu

(69) *Nie bije się dzieci.*

W opisie przyjętym na potrzeby NKJP uznano, że ze względu na trudności interpretacyjne segment *się* jest traktowany zawsze jako wykładnik samodzielnego nieodmiennego leksemu SIĘ (partykułowego). Leksem SIEBIE został ograniczony do trzech form. Te rozstrzygnięcia przyjęto także w niniejszej pracy.

Kategoria akcentowości

Kategoria *akcentowości* została wprowadzona dla odróżnienia form zaimka osobowego występujących w zdaniu na pozycji akcentowanej akc (np. *jego*) od form występujących na pozycji nieakcentowanej nakc (np. *go*). (Podobnie jak kategoria poprzyimkowości akcentowość po raz pierwszy została zaproponowana przez Saloniego, 1981).

Kategoria poprzyimkowości

Kategoria poprzyimkowości została wprowadzona do opisu, aby zdać sprawę ze zróżnicowania form występującego wyłącznie dla trzecioosobowego zaimka osobowego ON. Mianowicie pewne jego formy, np. *niego* i *niemu*, nie pasują do żadnego z kontekstów testowych przypadka wymienionych na stronie 30. Z teoretycznego punktu widzenia powinno to prowadzić do zwiększenia liczby przypadków stosowanych w opisie, w tym wypadku do rozróżnienia dopełniacza wymaganego przez czasownik, w którego kontekście właściwa jest forma *jego/go*, od dopełniacza wymaganego przez przyimek, w którego kontekście właściwa jest forma *niego*. Analogiczne rozróżnienia konieczne byłyby również dla celownika i biernika. Rozwiązanie takie byłoby wyjątkowo niepraktyczne, skoro omawiane zróżnicowanie występuje wyłącznie dla kilku form (por. Saloni 2005, s. 44). Zamiast tego wprowadzono oznaczenie odpowiednich form jako nacechowanych ze względu na dodatkową kategorię *poprzyimkowości*. Wartości tej kategorii to: poprzyimkowa praep i niepoprzyimkowa (przyczasownikowa) npraep. Wynikające z tego ograniczenie dystrybucji form musi być opisane w gramatyce.

1.7.5. Przymiotniki

Za przymiotniki adj zostały uznane leksemy, które mają kategorie fleksyjne przypadka, liczby i rodzaju (Saloni *et al.* 2012, s. 65). Formy przymiotników polskich są wysoce synkretyczne: wykładników typowego przymiotnika jest jedenaście, przypisanie im interpretacji przypadkowo-liczbowo-rodzajowych powoduje wyróżnienie kilkudziesięciu form fleksyjnych.

W obrębie paradygmatu przymiotnika można wskazać formy neutralne ze względu na wymienione kategorie. Zostały one zaliczone do dodatkowych fleksemów w obrębie leksemu przymiotnikowego, omówionych w kolejnych podpunktach.

Formy występujące w złożeniach

Pierwszą z form neutralnych jest forma występująca jako nieostatni człon złożeń typu *biało-czerwony*. Ma ona zawsze taką samą postać niezależnie od charakterystyki przypadkowo-liczbowo-rodzajowej członu ostatniego. Złożenia takie są przy interpretacji rozbijane na wiele segmentów (każdy łącznik stanowi osobny segment). Przyjęto taką interpretację, ponieważ konstrukcja ta jest regularna zarówno składniowo, jak i semantycznie. Nie ma więc powodu do obciążania słowników wykorzystywanych w dalszych etapach przetwarzania leksemami typu POLSKO-NIEMIECKI, które miałyby bardzo niską frekwencję, a ich własności można by doskonale przewidzieć na podstawie własności składników. Ponadto konstrukcja ta może być rozbudowywana, np. *stosunków polsko-ukraińsko-białorusko-rosyjskich*.

Odpowiedni fleksem został nazwany *formą przyprzymiotnikową* adja. Forma ta często, ale nie zawsze jest równokształtna z przysłówkiem derywowanym od danego przymiotnika, np. *biało-czerwony* i *zrobiło się biało*, ale *poważno-komiczny* i *zrobiło się poważnie*.

Ślady odmiany krótkiej (rzeczownikowej) przymiotników

W odmianie przymiotników polskich zachowały się pewne ślady historycznych form odmiany rzeczownikowej, zwanej też krótką lub niezłożoną. Najbardziej wyrazistym przykładem jest istniejąca dla niektórych przymiotników specjalna forma występująca po przyimku PO, np. *po polsku*, *po ojcowsku* (por. Saloni *et al.* 2012, s. 69). Jest ona pozostałością po dawnej krótkiej formie męskiej i nijakiej celownika. Dla tych przymiotników nie jest ona równokształtna z żadną regularną formą przymiotnika, podczas gdy dla pozostałych przymiotników jest w tym kontekście używana forma celownika: *po macoszemu*, *po naszemu*.

Drugim śladem krótkiej odmiany przymiotników są formy występujące w konstrukcjach typu *od dawna*, *z daleka*, *za widna*, *bez mała*, *do późna*. Formy te są pozostałością krótkiej formy dopełniacza męsko-nijakiego. Są one zawsze synkretyczne ze współczesną formą żeńskiego mianownika, jednak kontekst, w którym występują – po przyimku wymagającym dopełniacza – jest niezgodny z taką interpretacją. Aby zdać z tego sprawę oraz zwiększyć zgodność znakowania z korpusami historycznymi, w których odmiana krótka musi być znakowana, zdecydowano o traktowaniu również tych form jako specjalnych.

W celu opisania wymienionych dwóch typów form wprowadzono fleksem poprzyimkowy przymiotnika adjp. Należące do niego formy są charakteryzo-

wane wartością przypadka: celownika adjp:dat dla form typu *(po) polsku* i dopełniacza adjp:gen dla form typu *(z) polska*.

Problemów interpretacyjnych nastręcza występująca dla niektórych przymiotników mianownikowa forma krótka typu *pełen, wesół* (obok *pełny, wesoły*) (por. Saloni *et al.* 2012, s. 69). Formy te, zależnie od leksemu, mogą lub nie mogą być używane jako określenie rzeczownika (atrybutywnie):

(70) *Przyniosła pełen/pełny kosz kwiatów.*
(71) *Widziałem *wesół/wesoły kabaret.*
(72) *Piotr był bardzo wesół.*
(73) *Kosz był pełen/pełny.*

Formy, które mogą być użyte przy rzeczowniku, są interpretowane jako wariantywny mianownik lub biernik przymiotnika (a więc otrzymują tę samą interpretację jak forma typowa). Jeżeli jednak połączenie takie jest niemożliwe, forma zostaje przypisana do specjalnego fleksemu nazwanego przymiotnikiem predykatywnym adjc.

Kategoria stopnia

Wbrew SGJP (Saloni *et al.* 2012, s. 70) w opisie przyjęto, że stopień stanowi kategorię fleksyjną przymiotników i przysłówków. Co prawda tylko wyraźna mniejszość przymiotników jest stopniowalna, jednak zebranie form stopnia równego pos, wyższego com i najwyższego sup we wspólny leksem optymalizuje słownik, pozwalając łącznie opisać ich własności (np. ewentualne wymagania walencyjne). Własności składniowe i semantyczne form stopnia wyższego i najwyższego są bowiem regularne względem form stopnia równego.

Warto zaznaczyć, że kategoria stopnia ma znaczenie składniowe, jako że istnieją konteksty dopuszczające tylko formy konkretnego stopnia. Na przykład konstrukcje porównawcze z NIŻ muszą zawierać formę stopnia wyższego, np. *wyższy niż Piotr*, ale nie **wysoki niż Piotr*. Dlatego formom przymiotników niestopniowalnych należy przypisać wartość stopnia równego. Tak więc przyjęto, że fleksem przymiotnikowy adj jest odmienny przez liczbę, przypadek, rodzaj i stopień. Przymiotniki syntetycznie niestopniowalne traktowane są jako defektywne, posiadające jedynie formy stopnia równego.

Przymiotniki z *nie*

Od prawie każdego przymiotnika można utworzyć kolejny leksem przymiotnikowy poprzez dodanie cząstki *nie* na początku (np. ŻELAZNY – NIEŻELAZNY). Leksemy te są w słowniku notowane osobno. Niekiedy zbiegają się one z przymiotnikami o innym znaczeniu (np. DZIELNY – NIEDZIELNY), ale w związku z nienotowaniem znaczeń są traktowane jako jeden leksem.

1.7.6. Liczebniki

Leksemy liczebnikowe odmieniają się przez przypadek i rodzaj, a nie odmieniają się przez liczbę. W SGJP leksem liczebnikowy obejmuje tradycyjnie rozumiane liczebniki główne (*dwa*) i zbiorowe (*dwoje*) (por. Saloni 1977). Te dwie grupy form są w SGJP rozróżniane wartością rodzaju (n1, p1 i p2 – zbiorowe, pozostałe – główne). W przyjętej tu koncepcji do ich rozróżnienia służy osobna kategoria, która w punkcie 1.6.3 została nazwana przyrodzajem. W paradygmatach liczebników formy główne otrzymują oznaczenie ncol, a formy zbiorowe – col. Jeżeli liczebnik nie wykazuje tego zróżnicowania (np. wszystkie formy liczebnika STO mogą wystąpić zarówno w kontekście wymagającym liczebnika głównego, jak i zbiorowego), to oznaczenia te nie są stosowane. Wprowadzone w ten sposób oznaczenia form liczebnikowych są więc bardziej precyzyjne niż fleksem zbiorowy numcol stosowany w znakowaniu NKJP (Przepiórkowski *et al.* 2012, s. 73).

Większość liczebników ma ustaloną mnogą wartość liczby, sygnalizującą łączliwość z mnogimi formami rzeczowników. Jednak niektóre liczebniki obowiązkowo (PÓŁ, PÓŁTORA) lub fakultatywnie (DUŻO, TROCHĘ) łączą się z formami rzeczownikowymi o wartości pojedynczej liczby. W wypadku łączliwości fakultatywnej uznaje się istnienie dwóch leksemów (Saloni *et al.* 2012, s. 74), które mogą mieć różną odmianę:

(74) *Ilu*$_{\text{pl}}$ *kobiet*$_{\text{pl}}$ *tu brakuje?*
(75) *Ile*$_{\text{sg}}$ *mąki*$_{\text{sg}}$ *tu brakuje?*

Tak więc forma dopełniacza liczebnika ILE łączącego się z formami rzeczownikowymi liczby mnogiej ma postać *ilu*, a liczebnika ILE łączącego się z liczbą pojedynczą – *ile*.

Dla liczebników wprowadzono także fleksem reprezentujący formę występującą w złożeniach numcomp. W większości wypadków nie pojawia się ona w wynikach analizy, bo słowa takie jak *czterodrzwiowemu* są traktowane jako formy leksemu CZTERODRZWIOWY. Forma numcomp może pojawić się samodzielnie przy analizie pierwszego członu konstrukcji takich jak *cztero- albo pięciodrzwiowy*.

Odmienne przez liczbę i przypadek leksemy TYSIĄC, MILION, MILIARD itp. są rzeczownikami. Widać więc, że konstrukcje wyrażające liczność (np. *szesnastoma tysiącami dwieście trzydziestoma dwoma*) w opisie Saloniego składają się z form liczebników i rzeczowników. Istnieją także powierzchniowo nieodmienne liczebniki TYSIĄC, TYSIĄCE, MILION itd., ale nie występują one w szeregach liczebnikowych. Następujące wypowiedzenia ilustrują typowy kontekst składniowy pozwalający odróżnić konstrukcje liczebnikowe (76)–(78) od rzeczownikowych (79) i (80):

(76) *Wokół niego* *dwanaście*$_{\text{num}}$ *gwiazd migotało na niebie.*
(77) *Wokół niego* *tysiąc*$_{\text{num}}$ *gwiazd migotało na niebie.*

(78) *Wokół niego* *tysiące*$_{\text{num}}$ *gwiazd migotało na niebie.*
(79) *Wokół niego* *tysiąc*$_{\text{subst}}$ *gwiazd migotał na niebie.*
(80) *Wokół niego* *tysiące*$_{\text{subst}}$ *gwiazd migotały na niebie.*

W wypadku użycia liczebnika czasownik występuje w rodzaju nijakim w liczbie pojedynczej niezależnie od własności gramatycznych towarzyszącego mu rzeczownika.

Według klasyfikacji Saloniego leksem JEDEN, jako odmienny przez przypadek, rodzaj i liczbę (*jeden, jedna, jedno, jedni, jedne, ...*) jest przymiotnikiem. Konieczne było jednak wprowadzenie także powierzchniowo nieodmiennego liczebnika JEDEN, który używany jest w pozycji jednostek w wieloczłonowych nazwach liczb większych od 20. W kontekście wymagającym dowolnego przypadka jego formy mają postać *jeden*:

(81) *Zabrakło* *jednego*$_{\text{adj:gen}}$ *wyrazu.*
(82) *Zabrakło dwudziestu* *jeden*$_{\text{num:gen}}$ *wyrazów.*
(83) *[...] przyznania mieszkania* *jednemu*$_{\text{adj:dat}}$ *poszkodowanemu.*
(84) *[...] przyznania mieszkania dwudziestu* *jeden*$_{\text{num:dat}}$ *poszkodowanym.*

Kategoria akomodacyjności

Kategoria *akomodacyjności*, zaproponowana w artykule Bienia i Saloniego (1982), przysługuje wyłącznie formom liczebnikowym. Kategoria ta odróżnia formy liczebnika wiążące się z formami rzeczownika o tej samej wartości przypadka (uzgadniające congr, np. *dwaj*) od wiążących formy o wartości przypadka równej dopełniaczowi (rządzące rec, np. *dwóch, dwu*)[17]:

(85) *Przyszli dwaj chłopcy.* (congr)
(86) *Przyszło dwóch chłopców.* (rec)
(87) *Przyszło dwu chłopców.* (rec)

Konsekwencje składniowe tej cechy liczebników omówiono w punkcie 2.8.3.

Odmiennie niż we wspomnianym artykule (Bień i Saloni 1982) wartość akomodacyjności przypisywana jest wszystkim formom liczebników.

1.7.7. Leksemy nieodmienne

Klasyfikacja leksemów nieodmiennych w SGJP operuje bardziej szczegółowymi klasami niż pierwotna klasyfikacja Saloniego. Rozróżnienie między klasami nieodmiennymi odbywa się na zasadzie różnic własności składniowych. Na potrzeby analizy komputerowej wykorzystano klasy wyróżnione w SGJP, niektóre podziały zostały jednak uznane za zbyt subtelne i powodujące zbyt

[17] Formy dopełniaczowe liczebników zostały uznane za uzgadniające, czyli wartość rec oznacza wymaganie przypadka różnego od własnego.

dużo homonimii. Zdecydowano się na tym poziomie na opis mniej szczegółowy. Na poziomie składniowym jest on wysubtelniany poprzez dalsze zróżnicowanie własności jednostek (np. klasyfikację cech składniowych partykuł omówiono w p. 2.8.4).

Wykrzykniki wydzielone w klasyfikacji Saloniego na podstawie możliwości samodzielnego ich użycia jako wypowiedzenia, zostały w SGJP podzielone na wykrzykniki (z podklasami) i dopowiedzenia, które różnią się tym, że stanowią reakcję na wypowiedź innego rozmówcy (np. *Tak!*). Obie klasy są tutaj traktowane jako *wykrzykniki* interj. Dobrym testem na przynależność do tej klasy jest niemożność wystąpienia jako określenie jakiejkolwiek innej jednostki.

Jako *przyimki* wyróżniono leksemy pełniące w wypowiedzeniu funkcję łączącą i wymagające określonego przypadka (owym nosicielem przypadka jest najczęściej forma rzeczownika, ale są też sytuacje, gdy wymagany jest przymiotnik/fraza przymiotnikowa, np. *przerabiać kotłownie z węglowych na ekologiczne*). Przyimki, np. *jako*, *niż*, mogą (nietradycyjnie) wymagać mianownika (Saloni *et al.* 2012, s. 130). Nieliczne przyimki występują postpozycyjnie względem rzeczownika, jako tzw. poimki. SGJP notuje tu jednostki TEMU (*godzinę temu*) i NAPRZECIW (*komuś naprzeciw*).

Zrezygnowano z wyróżnionej w SGJP klasy relatorów (względników, Saloni *et al.* 2012, s. 131). Klasa ta obejmuje 22 jednostki typu SKĄD, DLACZEGO, ODKĄD, które są klasyfikowane jako homonimiczny relator/przysłówek, oraz dwa „czyste" relatory: GDY i ILEKROĆ. Na potrzeby analizy komputerowej uznano, że nie warto wprowadzać homonimii, a owe dwie jednostki również można uznać za przysłówki o specyficznych cechach. Tak więc większość z nich może wprowadzać konstrukcję względną lub konstytuować pytanie, a dwie z nich nie mają zdolności tworzenia pytania.

Za *spójniki* uznano leksemy pełniące funkcję łączącą, ale nie wymagające określonego przypadka i nie stanowiące składnika jednego z łączonych wyrażeń. Spójniki zostały podzielone na osobne klasy spójników współrzędnych conj i podrzędnych comp na podstawie rozwiniętych opisów w SGJP (Saloni *et al.* 2012, s. 132).

W klasyfikacji Saloniego klasa partykuło-przysłówków została wyróżniona na zasadzie negatywnej: jako leksemy nieużywane samodzielnie i niełączące. Na potrzeby SGJP klasa ta została rozbita na podstawie zdolności do wchodzenia w relację składniową z rzeczownikami. Leksemy zdolne do łączenia się z rzeczownikami uznano za *partykuły*, niezdolne zaś – za *przysłówki*. Wydzielono także klasy operatorów trybu (NIECH, BY, BYLEBY, ...), modyfikatorów deklaratywności (NIE, CZY, CZYŻBY, ...), operatorów adnumeratywnych (NIESPEŁNA, PRZESZŁO) i operatorów adsubstantywnych (LADA, BYLE).

Saloni i Świdziński (2001) postulują traktowanie przysłówków (odprzymiotnikowych) jako form odpowiednich leksemów przymiotnikowych, a więc w szczególności przypisywanie im lematu przymiotnikowego. Koncepcja ta utrudnia jednak analogiczną interpretację wyrazów, które można uznać za

przysłówki ze względu na podobne własności składniowe, a które nie pochodzą od przymiotników. Dlatego przysłówki, podobnie jak w SGJP, są tu traktowane jako samodzielne leksemy (odmienne przez stopień).

Uznano za *przysłówki* adv leksemy zaklasyfikowane tak w SGJP, pozostałe zaś wymienione wcześniej grupy połączono w klasę *partykuł* part. W konsekwencji przysłówki są klasą dość jednorodną co do właściwości składniowych, natomiast partykuły wymagają w zasadzie opisu własności składniowych każdej z nich z osobna (por. p. 2.8.4). W klasie partykuł znalazła się m.in. partykuła trybu warunkowego BY oraz partykuła stosowana w zwrotnych użyciach czasowników SIĘ.

W SGJP są także notowane wybrane produktywne prefiksy i sufiksy (Saloni *et al.* 2012, s. 147). Te pierwsze są wykorzystywane przez analizator fleksyjny, co pozwala analizować formy takie jak *antyamerykańskiego* dzięki połączeniu prefiksu *anty-* z formą przymiotnika AMERYKAŃSKI. Analiza prezentowana jest tak, jakby w słowniku był leksem ANTYAMERYKAŃSKI.

1.7.8. Skróty

Dla rozważań w tym punkcie istotne jest rozróżnienie między skrótami i skrótowcami. Przy czytaniu tekstu na głos skróty rozwija się, zastępując je pełnymi słowami (*np.* → [na przykład]), a skrótowce realizuje się literowo (*PZPR-u* → [pezetpeeru]). Skrótowce mogą być odmienne (*PZPR, PZPR-u, PZPR-em, ..., PZPR-y, PZPR-ami, ...*) i zachowują się składniowo jak rzeczowniki. Dlatego też w znakowaniu nie są odróżniane od innych rzeczowników. Grupą wymagającą specjalnego potraktowania są natomiast skróty.

Skróty nie są oczywiście klasą gramatyczną, jednak ze względu na swoją nietypowość są traktowane w systemie znaczników jak osobna klasa gramatyczna. Skrótom zamiast lematu jest przypisywane ich rozwinięcie. Opis gramatyczny sprowadza się do znacznika brev.

Kropka stawiana po niektórych skrótach jest traktowana jako osobny segment. Dzieje się tak dlatego, że po niektórych skrótach (*itd., br.*) może następować koniec zdania i wtedy kropka spełnia dwie role jednocześnie: sygnału końca zdania i obowiązkowego elementu skrótu. Skróty, które wymagają obecności kropki, są oznaczane znacznikiem brev:pun, skróty, które są pisane bez kropki, mają oznaczenie brev:npun.

Swoistym rodzajem skrótu są zapisy liczbowe w postaci ciągu cyfr – one też przy czytaniu byłyby interpretowane poprzez podanie słownego rozwinięcia zapisu cyfrowego. Ciągi cyfr są opatrywane przez Morfeusza znacznikiem dig, w polu lematu jest umieszczany ten sam ciąg cyfr. Uznano też, że na tym etapie przetwarzania niecelowe byłyby próby dokładniejszego interpretowania zapisów dat, liczb dziesiętnych itp., a w związku z tym jako jeden segment ze znacznikiem dig zostanie zinterpretowany dowolny ciąg złożony z cyfr, przecinków i kropek.

1.8. Nieregularności fleksyjne

We współczesnym języku pewne polskie słowa mogą być używane tylko w jednym określonym kontekście. Na przykład słowo *wznak* występuje tylko w *na wznak*, *króćset* – w *do króćset*, a *Banja* – w *Banja Luka*. W SGJP wyrazy takie nazwano „uwikłanymi frazeologicznie" (Saloni *et al.* 2012, p. 7.12, s. 141)[18]. W prezentowanym tu opisie zdefiniowano dla nich osobną klasę oznaczaną znacznikiem frag. Segmenty takie nie są dalej charakteryzowane gramatycznie.

Zdarza się też, że pewne słowa są używane w kontekstach dla siebie nietypowych. Na przykład słowa stanowiące wykładniki przysłówków MIĘKKO i WPROST mogą być użyte w konstrukcji ze słowem *na*: *na miękko*, *na wprost*. Konstrukcje te pełnią taką funkcję składniową jak formy przysłówków. Tego rodzaju konstrukcje nie są wyróżniane w znakowaniu fleksyjnym, inaczej mówiąc, uznaje się je za wyrażenie złożone z formy przyimka i formy przysłówka, mimo że przyimki co do zasady nie łączą się z przysłówkami (Saloni *et al.* 2012, s. 25). Interpretacja takich tworów została uwzględniona w przedstawionym dalej opisie składni przy rozpoznawaniu form składniowych (zob. p. 2.3).

1.9. Lematyzacja, czyli hasłowanie

Celem lematyzacji, inaczej nazywanej hasłowaniem, jest wskazanie dla każdego segmentu opisującej go jednostki słownika fleksyjnego (leksemu). Wymaga to skonstruowania jednoznacznych identyfikatorów wszystkich wyróżnionych leksemów. Typową praktyką słownikową jest identyfikowanie leksemu za pomocą wykładnika konwencjonalnie wybranej formy leksemu, np. bezokolicznika czasownika. Zdarza się jednak, że takie postępowanie daje ten sam identyfikator dla różnych leksemów:

(88) (a) *Jak śmiesz*$_{\text{ŚMIEĆ}}$ *tak się zwracać do matki?*
(b) *Weź ze sobą od razu ten kubełek ze śmieciami*$_{\text{ŚMIEĆ}}$.

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

[18] W artykule Derwojedowa i Rudolf (2003) zaproponowano termin *burkinostka* na oznaczenie konstrukcji takich jak *Burkina Faso* i *burkincząstka* na oznaczenie ich uwikłanych frazeologicznie członów.

Poręcznym elementem identyfikacji leksemów jest podawanie klasy gramatycznej, czyli części mowy[19]. W powyższych przykładach wystarczy to do jednoznacznego wskazania leksemu. Zdarza się jednak, że to za mało:

(89) (a) *Gdy wszyscy pływacy*$_{\text{PŁYWAK}}$ *są nieruchomi, starter podaje sygnał startu.*
(b) *W kałuży uwijały się pływaki*$_{\text{PŁYWAK}}$ *żółtobrzeżki.*
(c) *Pole magnetyczne generowane przez pływak*$_{\text{PŁYWAK}}$ *powoduje lokalną zmianę impedancji akustycznej struny.*

W przykładach występują formy o tej samej charakterystyce liczbowo-przypadkowej trzech różnych leksemów PŁYWAK, różniących się rodzajem (m1, m2 i m3).

W następujących zdaniach występują formy dwóch czasowników o tej samej postaci bezokolicznika, które jednak różnią się zbiorami form:

(90) (a) *Ślemy*$_{\text{SŁAĆ}}$ *doń list za listem już od dwóch miesięcy.*
(b) *Codziennie ścielemy*$_{\text{SŁAĆ}}$ *łóżka.*

Te leksemy można by rozróżnić, podając oznaczenia grup odmiany. To jednak wymagałoby przyjęcia konkretnego systemu wzorów odmiany, co wydaje się niekorzystne.

Bień proponował (Bień *et al.* 1973), aby w takich wypadkach uzupełniać identyfikator o fragmenty formy lub form różniących leksemy w liczbie wystarczającej do jednoznacznego wskazania, o który leksem chodzi. Praktyczne zastosowanie tego schematu wymagałoby wypracowania reguł określających wybór form wchodzących do identyfikatora. Taka procedura dałaby samoobjaśniające się identyfikatory, ale skomplikowałaby sposób ich generowania.

W bieżącej wersji Morfeusza postanowiono dołączać do identyfikatora leksemu literę sygnalizującą część mowy, a jeżeli to nie wystarcza, liczbę numerującą kolejne homonimy. Oczywiście przypisanie takich liczb jest arbitralne i ma wspomniane wcześniej wady[20].

W wypadku nielicznych czasowników z wariantywnym bezokolicznikiem jako lemat jest wybierany (arbitralnie) jeden z nich. Przy generowaniu form konieczne jest podanie właściwego lematu. Na przykład leksem zawierający formy *biec* i *biegnąć* analizator fleksyjny identyfikuje za pomocą lematu BIEC. Wygenerowanie formy inf dla tego lematu da wykładniki zarówno *biec*, jak i *biegnąć*.

---

[19] Oznaczenie klasy gramatycznej jest oczywiście częścią znacznika fleksyjnego, byłoby jednak korzystne, żeby to lemat sam w sobie był jednoznacznym identyfikatorem, żeby można go było w szczególności wykorzystać do powiązania ze słownikiem walencyjnym i dalszą warstwą semantyczną opisu. Ponadto w wypadku leksemów rozpadających się na wiele fleksemów konieczne jest ich zebranie. Tak więc lemat ŚMIEĆ:v będzie wspólny dla form tego czasownika przynależnych do fleksemów oznaczonych fin, praet itd.

[20] Jednym z celów nadania jednoznacznych identyfikatorów leksemom, było przypisanie ich do jednostek słownika walencyjnego Walenty, ponieważ na przykład leksemom SŁAĆ:v1 i SŁAĆ:v2 odpowiadają różne zbiory schematów składniowych. Jednak w chwili pisania tych słów praca ta nie została jeszcze wykonana.

1.10. Struktura znaczników fleksyjnych

Znacznik fleksyjny jest ciągiem wartości rozdzielonych dwukropkami, np.: subst:sg:nom:m1 dla segmentu *chłopiec*. Symbol przed pierwszym dwukropkiem określa klasę fleksemów (za pomocą oznaczeń z tabeli 1.3), następne – wartości kategorii gramatycznych przysługujących danej formie. Wartości są wymieniane w kolejności zgodnej z porządkiem kolumn w tabeli 1.4. Podawane są jedynie wartości kategorii adekwatnych dla danej klasy fleksemów, tak więc znaczenie danej pozycji znacznika zależy od tego, jaka to klasa[21].

W wypadku, gdy dla danego segmentu występuje niejednoznaczność wartości jakiejś kategorii, podawane jest kilka znaczników, np.: subst:sg:nom:m3 i subst:sg:acc:m3 dla segmentu *stół*. Stosowana jest skrócona notacja takich znaczników. Polega ona na wymienieniu alternatywnych wartości danej kategorii rozdzielonych kropką. Na przykład dwa podane znaczniki można zapisać łącznie: subst:sg:nom.acc:m3. Alternatywy na więcej niż jednej pozycji znacznika oznaczają wszystkie możliwe kombinacje wartości. Dlatego na przykład nie da się zapisać łącznie znaczników adj:sg:gen:m1.m2.m3.n:pos i adj:sg:acc:m1.m2:pos, stanowiących interpretacje segmentu *białego*.

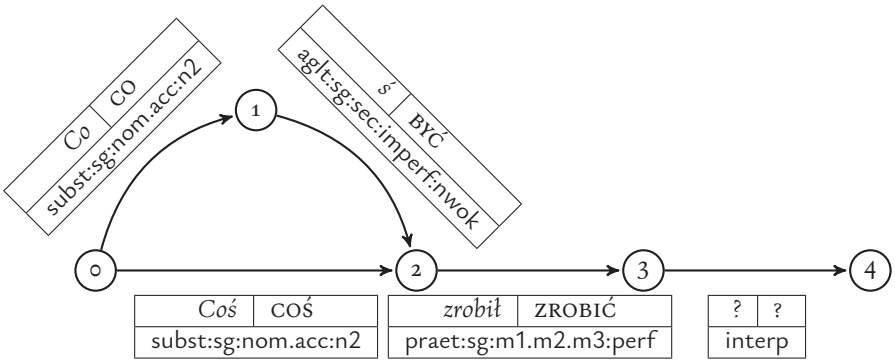
Wartości akcentowości i poprzyimkowości, które przysługują tylko niektórym formom zaimków, oraz wokaliczności, która przysługuje tylko niektórym przyimkom, są notowane na końcu znacznika. Gdy są nieobecne, znacznik jest krótszy o jedną lub dwie pozycje, ale wcześniejsze pozycje nie zmieniają znaczenia.

1.11. Grafowa reprezentacja interpretacji fleksyjnych

Wejściem analizatora fleksyjnego Morfeusz jest ciąg znaków, a wynikiem – lista interpretacji segmentów wykrytych w tym ciągu, reprezentująca skierowany acykliczny graf fleksyjny. Każda z interpretacji obejmuje: oznaczenia początkowej i końcowej pozycji segmentu w tekście, interpretowany segment, lemat, znacznik fleksyjny, ewentualną informację o byciu nazwą własną, ewentualne kwalifikatory stylistyczne i zakresowe z SGJP (np. sygnalizujące, że pewne formy są potoczne lub należą do języka specjalistycznego).

Zdarza się, że podział danego ciągu znaków na segmenty jest niejednoznaczny, a w związku z tym interpretacje generowane przez Morfeusza nie tworzą listy, ale graf acykliczny. Przykładem takiej sytuacji jest zdanie *Coś zrobił?*,

[21] Jednocześnie wartości wszystkich kategorii zostały tak dobrane, aby jednoznacznie odpowiadały kategoriom, w istocie więc z samej wartości wynika, do jakiej kategorii ona należy.



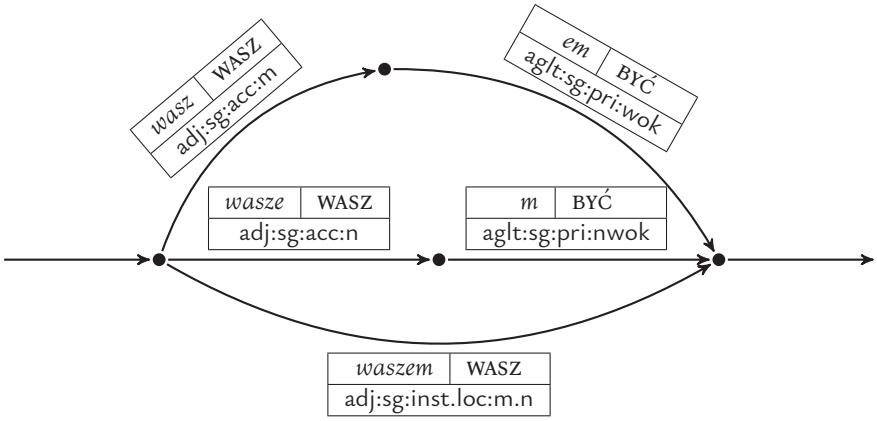
Rysunek 1.3. Graf fleksyjny dla zdania *Coś zrobił?*

0	1	*Co*	CO	subst:sg:nom.acc:n2
1	2	*ś*	BYĆ	aglt:sg:sec:imperf:nwok
0	2	*Coś*	COŚ	subst:sg:nom.acc:n2
2	3	*zrobił*	ZROBIĆ	praet:sg:m1.m2.m3:perf
3	4	?	?	interp

Rysunek 1.4. Reprezentacja grafu fleksyjnego z rys. 1.3 w postaci listy interpretacji analizatora Morfeusz

które można sparafrazować jako *Co zrobiłeś?*, gdzie *coś* jest interpretowane jako rzeczownik CO z doklejoną końcówką czasu przeszłego czasownika, lub też jako *Czy [on] coś zrobił?*, gdzie *coś* jest w całości interpretowane jako rzeczownik. Generowany dla tego przykładu graf fleksyjny przedstawiono na rysunku 1.3. Jest on reprezentowany w postaci ciągu interpretacji przedstawionego na rysunku 1.4. Liczby w pierwszych dwóch kolumnach tabeli oznaczają punkty pomiędzy poszczególnymi segmentami tekstu interpretowanymi przez program. W tym przykładzie fragment tekstu od punktu 0 do 2 (czyli *Coś*) może być interpretowany jako dwa segmenty: 0–1 *Co* i 1–2 *ś* lub pojedynczy segment 0–2. Numery są przydzielane rosnąco, ale, jak widać, nie w każdym wariancie interpretacji tekstu numeracja musi być ciągła.

Bardziej skomplikowane niejednoznaczności segmentacji można znaleźć w tekstach dawnych (Kieraś *et al.* 2017). Na rysunku. 1.5 przedstawiono graf interpretacji dla słowa *waszem*. Słowo to interpretowane w całości jest wykładnikiem historycznej formy narzędnika lub miejscownika leksemu WASZ. Co ciekawe, możliwe są również podziały tego słowa w dwóch różnych wykluczających się miejscach (*wasz·em* i *wasze·m*). W zależności od miejsca podziału pierwszy segment reprezentuje formę leksemu WASZ – męską (*wasz·em widział dom*) albo nijaką (*wasze·m widział dziecko*). Drugi segment jest aglutynacyjną formą czasownika BYĆ, dopasowaną do pierwszej.



Rysunek 1.5. Niejednoznaczna segmentacja słowa *waszem* według reguł historycznych

Przetwarzanie danych fleksyjnych w postaci grafu pozwala w naturalny sposób wyrazić zależności między kolejnymi segmentami. Na przykład poprzyimkowa forma przyimka ON musi stać po przyimku, wykluczone są więc inne interpretacje poprzedzającego segmentu. Manipulując ścieżkami w grafie, można wyrazić tę konieczność (por. Woliński 2004, p. 4.6).

Mechanizm ten można by też wykorzystać do reprezentowania jednostek wieloczłonowych. W odniesieniu do pewnych ciągów członów wiadomo, że muszą one być interpretowane łącznie. Dzieje się tak w szczególności w wypadku ciągów zawierających segmenty zinterpretowane jako frag, np. *na schwał*, i niektóre inne ciągi zawierające elementy, które z pewnością nie wchodzą w zwykłą konstrukcję składniową, np. złożone z przyimka i przysłówka *na pewno*. W takiej sytuacji komponent przetwarzający jednostki wieloczłonowe może zastąpić dwa łuki w grafie nowym łukiem reprezentującym całą jednostkę. W innych wypadkach interpretacja jako jednostka wieloczłonowa jest tylko jedną z możliwości. W takiej sytuacji moduł przetwarzający dokładałby nowy łuk w grafie, nie usuwając istniejących.

Warto zaznaczyć, że podobny sposób reprezentacji tekstu w postaci grafu wariantów jest stosowany przy przetwarzaniu mowy. Opisane tu techniki analizy fleksyjnej i składniowej pracujące na grafach fleksyjnych mogłyby więc być zastosowane do przetwarzania nieujednoznacznionego lub nie w pełni ujednoznacznionego wyjścia analizatora mowy.

Grafowa reprezentacja interpretacji fleksyjnych jest stosowana w pracy Obrębskiego (2002, p. 4.4.1). Używany jest w niej graf dualny do grafów fleksyjnych Morfeusza: wierzchołki reprezentują formy fleksyjne, a krawędzie – ich następstwo, czyli relację poprzedzania.

2

Przyjęta reprezentacja konstrukcji składniowych

Wykłady teorii gramatyk formalnych zwykle skupiają się na problemie, jaki język opisuje dany zbiór reguł gramatyki, a więc jakie zdania są akceptowane, a jakie odrzucane przez daną gramatykę. Jednak gdy mowa o modelowaniu języka naturalnego, problem, czy zdanie należy do opisywanego języka, jest tylko jednym z interesujących zagadnień. Równie istotne, a może nawet istotniejsze jest to, jakie struktury gramatyka przypisuje analizowanym wypowiedzeniom. Nie chodzi więc tylko o to, co jest akceptowane, ale i w jaki sposób. Dlatego przedmiotem tego rozdziału są zjawiska składniowe uwzględnione w gramatyce Świgra 2 i struktury, które w prezentowanym opisie odpowiadają poszczególnym konstrukcjom składniowym. Abstrahuje się przy tym od sposobu wyrażenia gramatyki, a więc mechanizmów, które miałyby tworzyć takie struktury. Mechanizmy te będą przedmiotem rozdziału 4.

Przedstawiane są przede wszystkim struktury przypisywane konstrukcjom bez wnikania w uwarunkowania wpływające na wartości cech przypisywanych poszczególnym frazom. Nie jest to więc pełny wykład prezentowanej gramatyki formalnej. Bardziej szczegółowe analizy były prezentowane w przywoływanych dalej publikacjach, a wszystkie szczegóły można znaleźć w publicznie dostępnej działającej implementacji[1]. Celem tego rozdziału jest zebranie informacji, które mogą być użyteczne dla osób zainteresowanych dalszym użyciem tworzonych interpretacji gramatycznych. Rozdział ten stanowi tym samym dokumentację struktur zawartych w korpusie składniowym Składnica (rozdział 6), które można rozpatrywać w oderwaniu od gramatyki wykorzystanej do ich wygenerowania.

Zasadniczym punktem odniesienia przedstawianego tu opisu jest gramatyka formalna polszczyzny autorstwa Marka Świdzińskiego (1992). Gramatyka ta będzie dalej przywoływana jako GFJP. Stosowane reprezentacje będą więc bliskie tej gramatyce, na ile jest to uzasadnione. Czytając pracę Świdzińskiego, można jednak odnieść wrażenie, że autor skupia się na możliwości wyrażenia reguł za pomocą przyjętego formalizmu gramatyki metamorficznej, traktując jego ograniczenia jako dane z góry i niemodyfikowalne. Niektóre rozwiązania

[1] W lekturze tego rozdziału może pomóc próba analizy wybranych przykładów wypowiedzeń za pomocą sieciowej wersji demonstracyjnej analizatora Świgra 2 dostępnej pod adresem http://swigra.nlp.ipipan.waw.pl/.

są kompromisem na rzecz formalizmu, podczas gdy struktura zgodna z językoznawczą intencją Autora byłaby inna. W odniesieniu do części takich struktur przedstawiono interpretację lepiej ugruntowaną w materiale językowym. Niektóre z przedstawionych zmian stanowią rozwinięcie sugestii zawartych w rozdziale 5 pracy doktorskiej autora (Woliński 2004). Prezentowane pomysły strukturyzacyjne są także konfrontowane z niektórymi innymi formalnymi opisami polszczyzny omawianymi w rozdziale 5.

Rozdział zaczyna się od wprowadzenia koniecznych pojęć z zakresu składni (punkt 2.1), a dalej zawiera omówienie zasad strukturyzacji poszczególnych konstrukcji składniowych. Prezentacja biegnie od kwestii ogólnych jak odejście od strukturyzacji binarnej (p. 2.2) i przyjęcie prostej hierarchii jednostek (p. 2.4), poprzez systematyzację schematów strukturyzacyjnych dla konstrukcji podrzędnych (p. 2.8) i współrzędnych (p. 2.9), aż do omówienia kilku ważnych konstrukcji szczególnych (p. 2.10–2.15).

2.1. Podstawowe pojęcia

2.1.1. Co należy do składni

Pojęcie składni bywa rozumiane rozmaicie (por. Saloni i Świdziński 2001, rozdz. I). W ujęciu filozoficznym w opisie języka wyróżnia się syntaktykę, semantykę i pragmatykę (por. Saloni i Świdziński 2001, s. 15). Syntaktyka opisuje ogół zjawisk związanych ze współzależnościami różnych elementów tekstów. Semantyka – odniesienie elementów tekstów do elementów rzeczywistości pozajęzykowej. Wreszcie pragmatyka dotyczy umieszczenia tekstu w sytuacji komunikacyjnej, w relacji do nadawcy i odbiorcy tekstu. Niniejsza praca poświęcona jest zagadnieniom mieszczącym się w tak rozumianej syntaktyce, jest to jednak pojęcie szersze od językoznawczego rozumienia składni.

Saloni i Świdziński (2001, s. 17) wyróżniają trzy stopnie segmentacji tekstu. Jednostki pierwszego stopnia, najdrobniejsze, to litery (lub szerzej znaki pisarskie) albo fonemy, w zależności od tego, czy rozważany jest tekst pisany czy mowa. Jednostki drugiego stopnia to segmenty lub – po interpretacji – formy fleksyjne. Opisem ich tworzenia z jednostek pierwszego stopnia zajmuje się fleksja (czy odrobinę szerzej – morfologia), której jest poświęcony rozdział 1. Jednostki trzeciego stopnia to wypowiedzenia. Terminem tym określa się twór językowy odpowiadający samodzielnemu komunikatowi, który w pisanym tekście ciągłym jest wyodrębniany za pomocą wielkiej litery na początku i kropki lub podobnego funkcjonalnie znaku interpunkcyjnego na końcu (Saloni i Świdziński 2001, s. 41). Zadaniem składni jest opis tego, w jaki sposób wypowiedzenia mogą być konstruowane z jednostek drugiego stopnia.

Dla określenia zakresu składni przyjętego w tej pracy istotne są więc dwa wymienione elementy: brak odwołań do elementów pozajęzykowych (co na-

zywa się składnią powierzchniową[2]) oraz ograniczenie zakresu opisu do konstrukcji wypowiedzeń z form fleksyjnych.

Szczególnym typem wypowiedzenia jest zdanie. Świdziński (1992, p. 1.2, s. 21) definiuje je jako zorganizowane wokół finitywnej formy czasownika (por. p. 1.7.2) lub złożone z innych zdań, połączonych spójnikami współrzędnymi. Przedstawiona przez Świdzińskiego gramatyka jest ograniczona do wypowiedzeń stanowiących zdania, przy czym przedmiotem szczególnej uwagi są zdania złożone. Niniejszy opis uwzględnia również pewne wypowiedzenia niezdaniowe (zob. p. 2.12), reorganizuje także opis zdań zorganizowanych wokół formy czasownika i wokół spójnika współrzędnego (zob. p. 2.7 i 2.9).

Podstawą opisu składniowego jest zasadnicza intuicja, że konstrukcje składniowe mają naturę hierarchiczną: większe konstrukcje powstają przez składanie ze sobą mniejszych. Takie założenie prowadzi do tworzenia struktur drzewiastych jako reprezentacji konstrukcji składniowych. Dwie zasadnicze koncepcje reprezentowania struktury składniowej, w postaci drzew składnikowych i zależnościowych, zostaną omówione w następnych punktach. Struktury drzewiaste, czyli spójne grafy skierowane bez cykli, są dobrze poznane, odpowiadają dobrze zbadanym formalizmom i mają bardzo jasne intuicje. Ten ostatni fakt jest zapewne najistotniejszym czynnikiem wpływającym na przyjęcie takiej zasady strukturyzacyjnej.

Pojęcie drzewa używane w językoznawstwie zakłada istnienie porządku liniowego na wierzchołkach drzewa odpowiadających formom fleksyjnym (por. Obrębski 2002, s. 52), jest to więc struktura trochę bogatsza od drzewa w teorii grafów. Wiąże się z nią kwestia ciągłości/nieciągłości tworzonych struktur, a więc dopuszczenia możliwości przecinania się krawędzi drzew – przy założeniu, że formy fleksyjne są ustawione zgodnie z porządkiem w wypowiedzeniu (por. p. 2.14).

Saloni i Świdziński (2001, s. 20) wskazują trzy typy związków uznawanych we współczesnej składni: linearne (określające pozycję poszczególnych jednostek składniowych względem siebie), strukturalne (określające wzajemne uwarunkowania jednostek, np. wymagania), semantyczne (zdeterminowane przez znaczenie). W zależności od tego, którym z tych związków da się prymat jako stanowiącym podstawę tworzenia struktury, otrzymuje się różne opisy. Opis zależnościowy skupia się na związkach strukturalnych, opis składnikowy – na

[2] Warto w tym miejscu zaznaczyć, że przymiotnik POWIERZCHNIOWY bywa w kontekście opisu składniowego używany w dwóch istotnie różnych znaczeniach. Przepiórkowski (2008) mówi o powierzchniowym przetwarzaniu, mając na myśli metody, które nie mają na celu tworzenia struktur składniowych dla całych wypowiedzeń, lecz tylko tworzenie częściowych struktur dla interesujących w danym zastosowaniu części wypowiedzenia (tzw. *chunking* lub *shallow processing*). Świdziński (1992) kontrastuje opis powierzchniowo-składniowy, a więc taki, który opiera się na formalnych cechach gramatycznych wyrazów i konstrukcji, ze składnią głębinową, a więc uwzględniającą semantykę. Niniejsza książka dotyczy składni powierzchniowej w sensie Świdzińskiego, ale nie przetwarzania powierzchniowego w sensie Przepiórkowskiego.

związkach linearnych. W istocie jednak każdy opis składniowy pretendujący do pełności musi uwzględniać oba te typy uwarunkowań.

Można postawić pytanie, czy ma sens opis powierzchniowoskładniowy, a więc abstrahujący od związków trzeciego z wymienianych przez Saloniego i Świdzińskiego typów – od związków semantycznych. Czy można uciec od składni głębokiej? Z jednej strony oczywiście bez uwzględnienia semantyki nie rozstrzygnie się do końca, jakie wypowiedzenia są akceptowalne. Świadczy o tym różnica w akceptowalności następujących dwóch przykładów:

(1) *Przyglądał się jej godzinę.*
(2) **Przyglądał się jej poezję.*

Akceptowalność tych przykładów zależy od denotacji rzeczowników stanowiących czwarty segment wypowiedzenia. Rzeczownik GODZINA oznacza jednostkę czasu, co sprawia, że zdanie jest akceptowalne; tak jednak nie jest w wypadku rzeczownika POEZJA.

Z drugiej strony znaczna część zależności stanowiących o poprawności wypowiedzenia (ale nie o jego sensowności) zachodzi na płaszczyźnie formalnych cech gramatycznych łączonych jednostek. Rozdzielenie owych cech formalnych od semantyki widać na przykład w czasownikowych konstrukcjach sfrazeologizowanych. Mimo że ich znaczenie może być odległe od literalnych znaczeń wyrazów tworzących frazeologizm, w przytłaczającej większości wypadków owe składniki, które przestały wnosić swoją indywidualną semantykę do znaczenia frazeologizmu, nie utraciły formalnych cech gramatycznych i powierzchniowoskładniowo zachowują się tak samo jak w konstrukcjach, których znaczenie wynika ze znaczeń składników (por. Świdziński 1996; Lewicki 1976). Następujące zdania ilustrują to na przykładzie frazeologizmu CIOSAĆ KOŁKI NA GŁOWIE:

(3) *Dziadzia Linsrum ciosał [żonie] spokojnie kołki do podwiązywania pomidorów.* [NKJP300]
(4) *Żona zaś ciosała mu kołki na głowie w kwestii jakichś długów czy wspólników.* [NKJP300]
(5) *Starych panów zostawiła w spokoju, nie ciosała im kołków na głowie o wiadro [...]* [NKJP300]

Użycie frazeologiczne czasownika CIOSAĆ w zdaniu (4) bardzo przypomina składniowo użycie niefrazeologiczne (3). Co więcej, w obrębie ustalonego frazeologizmu widać działający standardowy mechanizm składniowy: gdy przy czasowniku CIOSAĆ pojawia się negacja w (5), przypadek argumentu KOŁKI musi być realizowany jako dopełniacz zamiast biernika. Dlatego w niniejszym opisie przyjęto, że konstrukcja składniowa zdań (3) i (4) jest taka sama, a różnić się powinna dopiero ich reprezentacja semantyczna.

W opisie formalnym nastawionym na implementację ostre wydzielenie poziomów opisu wydaje się korzystne, zwłaszcza że do różnych poziomów pasują

różne techniki opisu. Przedstawiony tu opis obejmuje składnię powierzchniową – „grę kształtów", a nie „grę znaczeń" – może jednak zostać rozbudowany o kolejną warstwę opisującą semantykę. O ile struktury drzewiaste wydają się naturalną reprezentacją dla struktur składniowych, to struktura semantyczna być może powinna być reprezentowana inaczej, na przykład za pomocą formuł jakiejś logiki formalnej. Rozdzielenie warstw opisu ułatwia tworzenie takiej heterogenicznej reprezentacji.

2.1.2. Akomodacja i konotacja składniowa

W języku fleksyjnym łatwo jest zauważyć oddziaływania w obrębie wypowiedzenia polegające na tym, że obecność pewnych form fleksyjnych narzuca warunki na dopuszczalną charakterystykę innych współwystępujących form fleksyjnych. Ilustrują to następujące przykłady:

(6) *Widzę willę.*
(7) *Widzę modernistyczną willę.*
(8) **Widzę modernistycznego willę.*
(9) **Widzę modernistyczne willę.*

Forma przymiotnika wprowadzona do zdania (6) musi być w liczbie pojedynczej, w bierniku i w rodzaju żeńskim (sg:acc:f). Źródłem tych wartości jest forma rzeczownika *willę*: jeżeli w zdaniu (7) rzeczownik zostanie zastąpiony rzeczownikiem w innym rodzaju, zmieni się i pasująca forma przymiotnika. Obserwacja innych kontekstów prowadzi do generalizacji: forma przymiotnika występująca w zdaniu w związku z rzeczownikiem musi się z nim zgadzać co do liczby, przypadka i rodzaju.

Saloni i Świdziński (2001) w celu opisania tego rodzaju oddziaływań wprowadzają pojęcie *akomodacji syntaktycznej* i wyróżniają trzy jej rodzaje: morfologiczną[3], słownikową i czysto składniową. Akomodacja, a więc dostosowanie, polega na tym, że „dana jednostka składniowa, o ustalonym oczekiwaniu akomodacyjnym, przebiega listę jakichś innych jednostek w poszukiwaniu takiej, której dyspozycje składniowe odpowiadają temu oczekiwaniu" (s. 111). Akomodacja morfologiczna zachodzi, gdy użycie konkretnego członu akomodującego powoduje, że spośród form fleksyjnych leksemu zostaje wybrana jako drugi człon związku konkretna forma spełniająca oczekiwania akomodacyjne. W podanym przykładzie akomodacja morfologiczna sprawia, że z paradygmatu przymiotnika zostaje wybrana forma zgodna z akomodującą formą rzeczownika. Podobnie akomodacją morfologiczną jest dostosowanie przypadkowe formy rzeczownika do wymagań czasownika lub przyimka. Akomodacja słownikowa polega na tym, że oczekiwana jest jednostka zawierająca formę określonego leksemu. Na przykład przy czasowniku DOWIEDZIEĆ może

[3] W niniejszej pracy byłoby zapewne konsekwentniej nazywać ją akomodacją fleksyjną, przytaczam jednak wymienione terminy zgodnie z oryginałem.

się pojawić fraza z przyimkiem OD (*Dowiedział się od Marii*), ale nie z przyimkiem DO (**Dowiedział się do Marii*). Akomodacja czysto składniowa może być spełniona przez konstrukcję składniową o określonej budowie, porządku linearnym lub innych cechach składniowych. Na przykład oczekiwaniem czysto składniowym czasownika DOWIEDZIEĆ jest zdanie podrzędne typu pytajnego (pytajnozależne), np. *Dowiedział się, kto zacznie*.

Powiązania akomodacyjne nie muszą obejmować wszystkich form występujących w wypowiedzeniu. Saloni i Świdziński (2001, s. 225) wskazują, że niektórych partykuł (np. RACZEJ, CHYBA, MOŻE) i przysłówków (np. WCZORAJ, ZNIENACKA) nie należy uznawać za akomodowane. Przede wszystkim nie zachodzi tu wybór spośród różnych form leksemu, bo leksem ma tylko jedną formę. Trudno też mówić o innych oddziaływaniach akomodacyjnych, formy te bowiem mogą się pojawić w bardzo szerokiej klasie konstrukcji. Sensowniej jest więc przyjąć, że nie zachodzi wybór tych konkretnie jednostek z żadnej puli. Podobnie za nieakomodowane należy uznać niektóre inne konstrukcje, na przykład rzeczowniki w bierniku jak *chwilę, godzinę* i inne określenia czasu, które mogą się pojawić w prawie każdym wypowiedzeniu.

Innym typem powiązań, jakie Saloni i Świdziński (2001) zauważają w wypowiedzeniach polskich, jest konotacja, czyli „zapowiadanie przez daną formę wyrazową innej jednostki składniowej". Jednostkę podlegającą konotacji autorzy nazywają *frazą wymaganą* (rozdział X, s. 232). Obecność wszystkich konotowanych jednostek jest konieczna do tego, aby dane wypowiedzenie było nieeliptyczne. Autorzy wskazują na przykład, że forma czasownikowa *zabił* konotuje frazę rzeczownikową w mianowniku liczby pojedynczej, w trzeciej osobie i rodzaju męskim, oraz frazę nominalną w bierniku. W związku z tym zdanie (10) jest pełne, a zdanie (11) – eliptyczne:

(10) *Morderca zabił dziewczynę.*
(11) *Morderca zabił.*

Różnicę między oddziaływaniami akomodacyjnymi i konotacyjnymi autorzy opisują następująco:

> niespełnienie zapowiedzi konotacyjnej daje wypowiedzenie eliptyczne (ale poprawne gramatycznie), podczas gdy niespełnienie oczekiwania akomodacyjnego – wypowiedzenie niepoprawne (choć, być może, pełne, a nawet zrozumiałe) (s. 264).

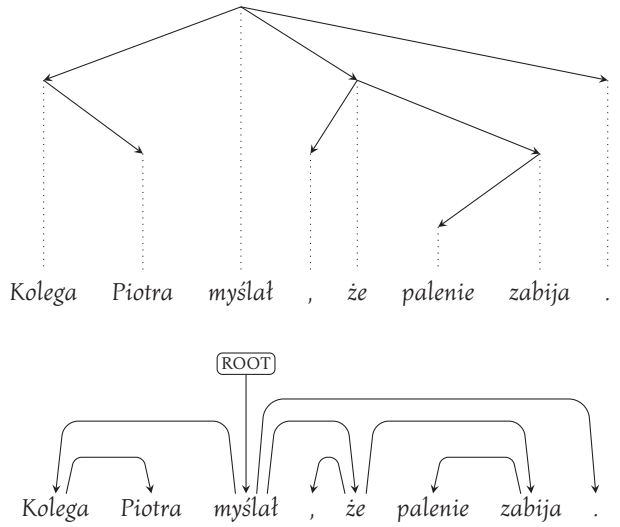
Oznacza to niestety, że kwestia pełności konstrukcji w języku polskim jest nieostra, bo stanowi uwarunkowanie semantyczne, jest związana z tym, czy wypowiedzenie niesie wystarczająco dużo informacji, a nie z wykładnikami formalnymi. Powoduje to problemy przy decydowaniu, które frazy są wymagane. Informacje o oddziaływaniach konotacyjnych poszczególnych jednostek zbiera się w słowniku walencyjnym (por. rozdz. 3).

2.1.3. Analiza zależnościowa

Ideą zależnościowego opisu składni jest zdanie sprawy przede wszystkim z zależności konotacyjnych poprzez wskazanie jednostek, między którymi występują. Strukturę składniową przedstawia się w tym wypadku za pomocą drzewa zależności. Jest to drzewo, którego wierzchołki reprezentują formy fleksyjne tworzące wypowiedzenie (z uwzględnieniem „form" odpowiadających znakom interpunkcyjnym, por. p. 1.5). Krawędzie drzewa łączą element konotujący (a często również akomodujący) z konotowanym (i akomodowanym). Stawiany jest warunek, aby do drzewa należały wszystkie formy tworzące wypowiedzenie, a w związku z tym, aby każda forma – z wyjątkiem jednej, stanowiącej korzeń drzewa – miała rodzica. Oznacza to oczywiście, że relacja zależności musi zostać odpowiednio rozszerzona, aby objąć również te elementy, które nie podlegają oddziaływaniu konotacyjnemu ani akomodacyjnemu.

Przykład drzewa zależnościowego przedstawiono na rysunku 2.1. Analizowane wypowiedzenie zostało uznane za zdanie, w związku z czym korzeń drzewa stanowi forma finitywna czasownika MYŚLEĆ. Przebieg niektórych strzałek wydaje się oczywisty (np. między *myślał* i *Kolega* – konotacja podmiotu; między *Kolega* a *Piotra*), jednak niektóre z nich są przyjęte w sposób mniej lub bardziej arbitralny (np. przyłączenie przecinka do *że*).

Oddziaływania konotacyjne i akomodacyjne są centralne dla reguł gramatyki zależnościowej, jednak reguły takie muszą uwzględniać również pewne



Rysunek 2.1. Reprezentacja zależności składniowych w postaci drzewa (na górze). Na dole: zwykle stosowany diagram wyrażający te same zależności, który jest bardziej zwarty, ale słabiej widać na nim drzewiastą naturę struktury

ograniczenia linearne. Na przykład w regułach Obrębskiego (2002) przyimek „szuka" elementu niosącego wymagany przezeń przypadek jedynie po swojej prawej stronie.

2.1.4. Analiza składnikowa

Inne podejście do opisu struktury składniowej wychodzi od pojęcia członu syntaktycznego. Saloni i Świdziński (2001, s. 24) definiują człon syntaktyczny jako każdą zinterpretowaną składniowo formę fleksyjną oraz każdy kompleks form traktowanych w analizie składniowej jako całostka. Jako kryterium ustalania, jakie ciągi form powinny być traktowane w ten sposób, autorzy proponują badanie, czy ciąg taki może być użyty w izolacji, a więc na przykład w odpowiedzi na pytanie o odpowiedni element wypowiedzi (s. 24) lub jako zawiadomienie – napis sytuowany, a więc umieszczony na przykład na budynku lub na karcie tytułowej książki (s. 32). W tej pracy zamiast terminu człon syntaktyczny będzie używany krótszy termin – składnik.

W zdaniu przedstawionym na rys. 2.1 można za pomocą testu pytań wyróżnić następujące składniki (w nawiasach pytania o te składniki):

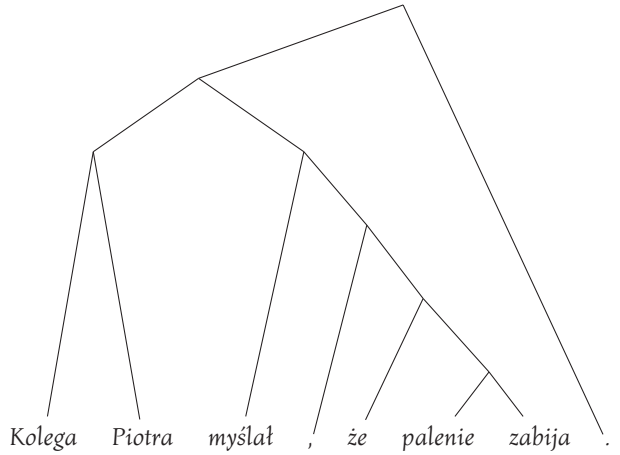
(12) *Kolega* (*Kto myślał?*)
Kolega Piotra (*Kto myślał?*)
Piotra (*Czyj kolega myślał?*)
że palenie zabija (*Co myślał?*)
palenie zabija (*Jaka była myśl kolegi?*)
Kolega Piotra myślał (*Co robił kolega Piotra?*)
myślał, że palenie zabija (*Co robił kolega Piotra?*)
Kolega Piotra myślał, że palenie zabija. (*Co robił kolega Piotra?*)

Warto zauważyć, że nie wszystkich form fleksyjnych w tym zdaniu da się użyć samodzielnie (i nie da się o nie zapytać). W szczególności forma *że* jest składnikiem tylko dzięki pierwszemu członowi definicji. Podobnie traktowane są znaki interpunkcyjne. Całe wypowiedzenie jako samodzielne spełnia definicję składnika.

Wskazane kryteria nie są bardzo precyzyjne, ale warto je przyjąć jako intuicję pojęcia, którego formalna definicja następuje w zasadzie poprzez podanie gramatyki formalnej: składniki są takie, jakie pozwala skonstruować dana gramatyka.

Pojęciem zbliżonym do składnika jest fraza. W niniejszej pracy termin ten będzie używany, gdy chodzi o podkreślenie typu budowy danego składnika (wyróżnione więc zostaną frazy: nominalne, przymiotnikowe, czasownikowe itd.). Nie będzie przestrzegane rozróżnienie między frazą a grupą składniową wprowadzone w pracy Saloniego i Świdzińskiego (2001), w szczególności ten ostatni termin w ogóle nie będzie używany.

Koncepcja analizy składnikowej polega na wyróżnieniu hierarchicznej struktury składników: wypowiedzenie dzieli się na fragmenty stanowiące je-



Rysunek 2.2. Drzewo składników bezpośrednich

go bezpośrednie składniki, a te znowu na składniki drobniejsze, aż do osiągnięcia jednostek uznanych z punktu widzenia składni za atomowe, a więc składników odpowiadających formom fleksyjnym (por. Saloni i Świdziński 2001, rozdz. III). Jeżeli przyjmie się, że składniki muszą być ciągłe, rozłączne i podział musi dokonywać się bez reszty, to rekurencyjny podział na składniki można zilustrować drzewem składników bezpośrednich.

Rysunek 2.2 przedstawia przykładowe drzewo składników bezpośrednich dla zdania z rysunku 2.1. Liśćmi drzewa są składniki odpowiadające poszczególnym formom fleksyjnym, a każdy węzeł wewnętrzny reprezentuje jakiś składnik. Można by go zaetykietować odpowiednim fragmentem tekstu wypowiedzenia. Liście drzewa są wymieniane w kolejności zgodnej z kolejnością wystąpień form w wypowiedzeniu[4].

Każdy wierzchołek wewnętrzny drzewa reprezentuje informację, że odpowiadający mu składnik może zostać rozłożony na pewien ciąg składników, które są dziećmi danego w drzewie (występujących w ustalonej kolejności).

Warto zauważyć, że o ile w drzewie zależnościowym węzły odpowiadają wyłącznie formom fleksyjnym, więc w szczególności ich liczba jest ustalona w wyniku analizy fleksyjnej, o tyle w wypadku opisu składnikowego węzły odpowiadają abstrakcyjnym jednostkom definiowanym przez twórcę gramatyki. Przedstawione drzewo jest więc jedynie jedną z możliwości, jak można widzieć rozkładanie danego wypowiedzenia na składniki.

Analiza na składniki bezpośrednie jest techniką opisu stosowaną w gramatyce Świdzińskiego (1992), a w związku z tym również w prezentowanej tu gramatyce.

---

[4] Saloni i Świdziński (2001, s. 53) zakładają, że kolejność liści nie jest związana z kolejnością form w wypowiedzeniu. Jednak autorzy gramatyk formalnych, w tym Świdziński (1992), zakładają zgodność porządku.

2.1.5. Dystrybucyjne podejście do składni

Podejście dystrybucyjne, o którym była już mowa w punkcie 1.4, ma zastosowanie również w składni. Polega ono na wyróżnianiu klas abstrakcji konstrukcji (składników) mających tę samą dystrybucję, a więc mogących występować w tych samych kontekstach.

Szpakowicz (1983, s. 10) przedstawia to podejście następująco:

> Dystrybucja tworu składniowego jest to zbiór poprawnych językowo kontekstów, w których twór ten może wystąpić. Twory mające tę samą dystrybucję pozostają w relacji równoważności. Reguły składniowe odwołują się najczęściej nie do wyrazów, lecz do klas abstrakcji tej relacji.

W tym ujęciu etykietowanie wierzchołków drzewa 2.2 składnikami należy rozumieć w ten sposób, że wierzchołki są etykietowane klasami abstrakcji równoważności dystrybucyjnej, których dane składniki są reprezentantami (twórca gramatyki nadaje owym klasom abstrakcji bardziej poręczne nazwy). Dystrybucyjną klasyfikacją konstrukcji posługują się Saloni i Świdziński (2001, s. 56 i n.) i Świdziński (1992).

Podobnie jak w wypadku fleksji, równoważność dystrybucyjną trzeba rozpatrywać z pewną dokładnością lub na odpowiednim poziomie abstrakcji. Autor GFJP abstrahuje od semantyki, gdy więc mówi o równoważności konstrukcji, ma na myśli identyczne zbiory dopuszczalnych kontekstów składniowych bez uwzględnienia ograniczeń występowania składników wynikających z semantyki.

Działając w ten sposób, można zauważyć, że składnik *książce Piotra* składa się z dwóch składników (*książce* i *Piotra*), z których pierwszy może występować w takich samych kontekstach, jak całość. Klasę abstrakcji tych składników można nazwać frazą nominalną lub, na bardziej szczegółowym poziomie obserwacji, frazą nominalną o wartości pojedynczej liczby, żeńskiej – rodzaju i celownikowej – przypadka.

2.1.6. Podrzędność, współrzędność i nieredukowalność

Podstawę istotnego sposobu klasyfikacji konstrukcji składniowych stanowi pojęcie reprezentanta. Reprezentantem konstrukcji (ściślej: składnika) nazywa się taki składnik bezpośredni, który jest równoważny dystrybucyjnie całej konstrukcji. O takiej sytuacji mówi się, że konstrukcję można zredukować do danego składnika. W ostatnim przykładzie składnik *książce* jest równoważny frazie *książce Piotra*, której jest składnikiem bezpośrednim, a zatem jest reprezentantem tej konstrukcji.

Konstrukcje składniowe, które mają dokładnie jednego reprezentanta, nazywa się podrzędnymi (por. Saloni i Świdziński 2001, p. I.6, s. 24). Konstrukcje posiadające wśród swoich składników wielu reprezentantów noszą nazwę współrzędnych. Oba te typy łącznie nazywane są konstrukcjami redukowalny-

mi, ich przeciwieństwem zaś są konstrukcje nieredukowalne, których żaden składnik nie jest reprezentantem (Saloni i Świdziński 2001, p. III.5, s. 58 i n.)[5].

W konstrukcji podrzędnej ten składnik, który jest reprezentantem, nazywa się nadrzędnikiem konstrukcji, a jego współskładniki – podrzędnikami. Cała konstrukcja podrzędna nosi więc takie cechy gramatyczne jak nadrzędnik konstrukcji, a podrzędniki modyfikują nadrzędnik semantycznie, ale nie wpływają co do zasady na własności gramatyczne całości, w szczególności mogą mieć własności zupełnie odmienne od nadrzędnika. Na przykład jeśli związek składniowy tworzy forma rzeczownika z frazą przyimkową (np. *dom z tarasem*), różnica ich własności składniowych jest wyrazista. Cała konstrukcja dzieli dystrybucję z rzeczownikiem, a dystrybucja fazy przyimkowej jest ewidentnie inna:

(13) *Widziałem dom z tarasem.*
(14) *Widziałem dom.*
(15) **Widziałem z tarasem.*

Konstrukcje współrzędne (zwane też skoordynowanymi) zwykle składają się z kilku składników o takich samych (na odpowiednim poziomie abstrakcji) własnościach gramatycznych. Pozostałymi współskładnikami takiej konstrukcji są spójniki (współrzędne) i znaki interpunkcyjne. Przejrzyste przykłady współrzędności stanowią szeregi fraz tworzone z wykorzystaniem spójnika I:

(16) *chłopców i dziewczęta*
(17) *mały, biały i zupełnie niegroźny*
(18) *w domu i na ulicy*

Typową konstrukcją nieredukowalną jest fraza złożona z przyimka i rzeczownika lub jego dystrybucyjnego równoważnika, np. *w domu*. We frazie tej nie ma reprezentanta, bo forma przyimkowa nie może wystąpić samodzielnie, a fraza rzeczownikowa ma inną dystrybucję. W tym wypadku całość ma inne własności gramatyczne niż każdy ze składników.

W większości wypadków rozróżnienie konstrukcji podrzędnych, współrzędnych i nieredukowalnych jest wyraziste. Zdarzają się jednak wątpliwości i nietypowości. Nietypowe frazy nominalne, które mają budowę identyczną jak konstrukcje współrzędne, ale są nieredukowalne, omówiono w punkcie 2.9.3. Wątpliwości klasyfikacyjne mogą budzić konstrukcje złożone ze zdań połączonych spójnikami, ponieważ w ich wypadku trzeba zdecydować, czy dystrybucja poszczególnych typów zdań jest na tyle różna, żeby powiedzieć, że łączone elementy nie są sobie równoważne. W wypadku zdań, które z natury są samodzielne, jest to kwestia nieoczywista (por. kwestię spójnika WIĘC w p. 2.10).

[5] W literaturze używane są także określenia *konstrukcje endocentryczne i egzocentryczne*. W niniejszej pracy termin *centrum* nie jest równoważny reprezentantowi (zob. dalej), określenia te nie będą więc używane.

W konstrukcjach składniowych wszystkich wymienionych w tym punkcie typów wygodnie jest wskazać jeden składnik, który w największym stopniu odpowiada za własności składniowe danej konstrukcji. Składnik taki będzie w tej pracy nazywany centrum składniowym (za Świdzińskim 1992). Przyjęto, że w wypadku konstrukcji podrzędnych centrum składniowe stanowi reprezentant, czyli nadrzędnik. W wypadku konstrukcji współrzędnych za centrum obrano spójnik (lub jeden ze spójników). W wypadku konstrukcji nieredukowalnych – składnik bardziej charakterystyczny składniowo, np. dla frazy przyimkowo-nominalnej za centrum uznano przyimek (zob. także p. 2.6). W dalszym ciągu rozważań przez centrum frazy będzie rozumiany albo jej składnik bezpośredni oznaczony jako centrum, albo forma fleksyjna, do której prowadzi gałąź drzewa na każdym poziomie schodząca do centrum w pierwszym rozumieniu (centrum rozumiane rekurencyjnie).

2.1.7. Podrzędniki czasownika

Składnia tradycyjna wypracowała kilka określeń podrzędników finitywnych form i fraz czasownikowych. Wyróżnia się wśród nich podmiot i dopełnienia (łącznie nazywane argumentami) oraz okoliczniki. Argumenty są frazami konotowanymi, w odróżnieniu od okoliczników. Rozróżnieniu temu odpowiadają w GFJP pojęcia fraz wymaganych i luźnych (w pewnym przybliżeniu, pojęcie fraz luźnych obejmuje bowiem w GFJP nie tylko podrzędniki czasownika, ale również wtrącenia).

Jak już wspomniano, kwestia, czy dany podrzędnik jest konotowany przez czasownik, jest nieostra. Niektórzy badacze wskazują, że różne zaproponowane w literaturze testy w tym zakresie dają odpowiedzi wzajemnie sprzeczne, w związku z czym postulują całkowite porzucenie tego rozróżnienia (zob. np. Przepiórkowski 2016, 2017). W niniejszej pracy przyjęto rozstrzygnięcie techniczne: jako frazy wymagane **fw** będą realizowane te podrzędniki, których obecność postuluje słownik walencyjny Walenty (rozdz. 3). Pozostałe podrzędniki uzyskują status fraz luźnych **fl**.

Istotne dla dalszych rozważań jest pojęcie podmiotu. Niestety nie ma wśród badaczy zgodności co do jego definicji. W niniejszej pracy przyjęto, że podmiotem będzie nazywany ten podrzędnik finitywnej formy czasownika, który jest gramatycznie wyróżniony. W języku takim jak polski, z wyraźnie wyrażonymi fleksyjnie kategoriami gramatycznymi, najbardziej fundamentalnym kryterium bycia gramatycznym podmiotem może być występowanie uzgodnienia (akomodacji) rodzaju, liczby i osoby podmiotu z formą czasownika (Saloni 2005). Tak więc akomodująca fraza nominalna w mianowniku zostanie uznana za podmiot. Warto zaznaczyć, że nie każda fraza nominalna w mianowniku ma tę cechę, jako że zdarzają się wymagane frazy mianownikowe, przy których nie ma miejsca uzgodnienie (por. p. 3.1.4).

Powstaje jednak pytanie, czy pojęcie podmiotu warto rozszerzyć na inne typy fraz wymaganych, na przykład takie, które mogą zostać skoordynowa-

ne z frazą nominalną w mianowniku. Takie rozszerzenie faktycznie zostało dopuszczone w słowniku Walenty (zob. p. 3.1.3) i w związku z tym tak samo interpretuje odpowiednie podrzędniki analizator Świgra 2.

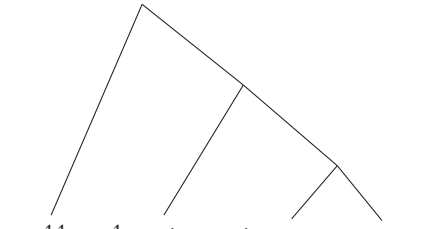
2.2. Niebinarność struktury

Saloni i Świdziński (2001) piszą, że analiza składnikowa w postaci opracowanej przez deskryptywistów amerykańskich (Wells 1947) sprowadza się „do ciągu binarnych, o ile to tylko możliwe, podziałów pewnej całości, np. zdania" (§III.2, s. 50). Nie dyskutują jednak, dlaczego preferowane są podziały binarne. Prowadząc analizę przykładu

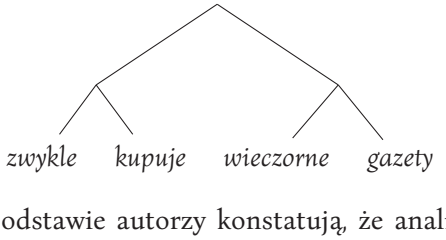
(19) *Moja starsza koleżanka zwykle kupuje wieczorne gazety.*

podają, że dla fragmentu *zwykle kupuje wieczorne gazety* możliwe są dwie strukturyzacje:

(20)



(21)



Na tej podstawie autorzy konstatują, że analiza składnikowa jest niejednoznaczna. Tymczasem można postawić pytanie, czy ta formalna różnica struktur odpowiada jakiemuś rzeczywistemu zróżnicowaniu z punktu widzenia językoznawczego, a więc czy jest pożądane, aby tworzony opis takie rozróżnienie uwzględniał. Jeżeli przedstawiona różnica struktur nie ma wyraźnego uzasadnienia, należy uznać, że rozróżnienie jest sztuczne, i tak zaprojektować reprezentację składniową, żeby się ono nie pojawiało. Tworząc opis składniowy, warto dążyć do tego, żeby niejednoznaczności przypisywanej struktury odpowiadały faktycznej niejednoznaczności (semantycznej) analizowanego wypowiedzenia. W tym przykładzie tak się jednak nie dzieje. O odcinku *zwykle kupuje wieczorne gazety* można powiedzieć tyle, że *wieczorne gazety* są całostką, która opisuje obiekt czynności kupowania, a *zwykle* określa okoliczności tej

czynności. Składniowo obie te frazy są zależne od czasownika, ale do rozstrzygnięcia, która z nich wiąże się z czasownikiem jako pierwsza (wiąże „mocniej"), brak wyrazistych przesłanek.

Aby zachować binarny podział fraz, ale uniknąć pokazanej niejednoznaczności, można arbitralnie wybrać jedną w pokazanych struktur. Odpowiada to na przykład przyjęciu zasady, że podrzędniki po prawej stronie centrum składniowego wiążą się z nim przed podrzędnikami po lewej stronie. Przy takim założeniu ilekroć nadrzędnik ma po jednym podrzędniku z lewej i z prawej, w drzewie pojawia się charakterystyczny układ węzłów. Układ ten można traktować jako „superwęzeł" reprezentujący centrum z dwoma podrzędnikami.

Bardziej czytelnym sposobem uzyskania tego efektu jest jednak odejście od podziału binarnego. Sytuację ze zdania przykładowego można adekwatnie zrelacjonować w strukturze drzewa składnikowego, tworząc wierzchołek o trzech węzłach potomnych.

Koncepcja binarnego podziału fraz współgra z przyjmowanym przez tradycyjną składnię języka polskiego podziałem zdania na grupę podmiotu i grupę orzeczenia. Powoduje ona trudności z interpretacją zdań, w których podmiot linearnie występuje wewnątrz grupy orzeczenia, jak w przykładach:

(22) *Lokalizację sektora* *użytkownicy* *uważają za atrakcyjną*. [Skł.]

(23) *Auto zostało całkowicie zniszczone, a* *kierowcę* *strażacy* *wydobyli za pomocą narzędzi hydraulicznych*. [Skł.]

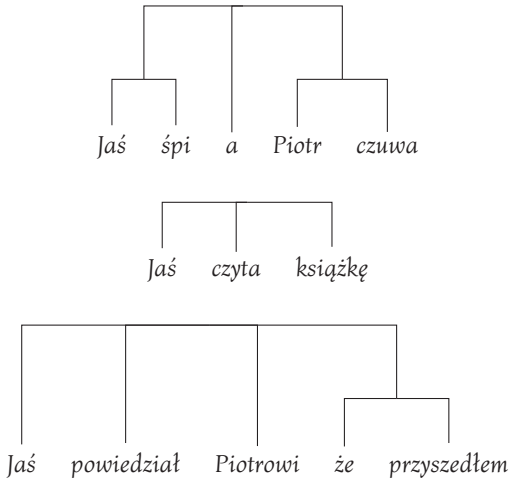
(24) *Co* *on* *robi*? [Skł.]

(25) *Jakąś łzę* *pani* *uroniła*? [Skł.]

Aby zachować binarny podział konstrukcji, trzeba by uznać przytoczone zdania za nieciągłe: składnik stanowiący podmiot (np. *użytkownicy*) rozrywa ciągłość składnika stanowiącego grupę orzeczenia (np. *lokalizację sektora uważają za atrakcyjną*). Odejście od koncepcji grupy orzeczenia pozwala opisać te zdania jako ciągłe.

Kwestia binarnego podziału zdania, a więc specjalnego statusu składnika reprezentującego podmiot, została również zanalizowana w pracy Przepiórkowski *et al.* (2002, rozdz. 2). Autorzy przytaczają kilka argumentów wydających się świadczyć na korzyść podziału binarnego, a następnie zbijają je i przyjmują reprezentację, w której składnik podmiotowy jest w strukturze składniowej na równych prawach ze składnikami dopełnieniowymi.

Podział zdania na grupę podmiotu i grupę orzeczenia jest obecny w gramatyce Szpakowicza (1983). Autor zauważa, że uniemożliwia to analizę zdań w niektórych wariantach szyku. Co ciekawe, na niższym poziomie, gdzie wysycane są wymagania formy czasownikowej, struktura proponowana przez Szpakowicza nie jest już binarna, wszystkie dopełnienia są podczepiane do formy czasownikowej jednocześnie. Wynika z tego także, że gramatyka Szpakowicza nie uwzględnia zjawiska czasowników niewłaściwych (jest to świadome uproszczenie opisu): na poziomie, na którym łączy się grupa podmiotu z gru-



Rysunek 2.3. Koncepcja podziału zdań na składniki bezpośrednie według gramatyki Świdzińskiego (1992, rys. (4)–(6), s. 63–64). Podobnie jak w oryginale pominięto znaki interpunkcyjne

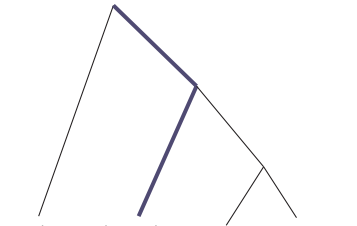
pą orzeczenia, nie ma dostępu do odpowiednich własności czasownika, nie jest rozważane, czy dany czasownik dopuszcza przy sobie podmiot.

Pełne odejście od struktur binarnych dla zdań reprezentuje gramatyka Świdzińskiego (1992). Wprowadzając koncepcję analizy na składniki bezpośrednie (§4.2, s. 62), autor deklaruje, że podział rozważanej konstrukcji (w wariancie analizy rozwijania) może dawać w rezultacie wiele składników bezpośrednich. Inaczej niż w pracy Saloniego i Świdzińskiego (2001) zakłada przy tym (s. 66), że „kolejność rozgałęzień [potomków wierzchołka] jest odwzorowaniem porządku linearnego dominowanych przez odpowiednie wierzchołki składników".

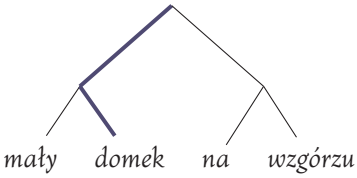
Przykłady struktur postulowanych przez Świdzińskiego dla zdań prezentuje rysunek 2.3. Istotnym elementem tej analizy jest wskazanie centrum konstrukcji wywierającego wpływ implikacyjny na pozostałe składniki (§4.2.2, s. 63), nie jest ono jednak oznaczane w drzewie. Świdziński jako zaletę opisu niebinarnego podaje to, że takie centrum implikacyjne jest współskładnikiem ze wszystkimi swoimi podrzędnikami, co nie jest możliwe, gdy opis jest binarny – jeżeli centrum ma bowiem więcej niż jeden podrzędnik, niektóre z tych podrzędników muszą być zagnieżdżone głębiej w strukturze współskładnika. Przyjęta koncepcja strukturyzacyjna pozwala także dopasować węzeł zdania elementarnego do odpowiedniego schematu zdaniowego (Saloni i Świdziński 2001, rozdz. 12).

Świdziński opowiada się jednak za strukturami binarnymi na innych poziomach opisu, w szczególności w odniesieniu do fraz składnikowych. O ile więc unika niejednoznaczności struktur (20) i (21), to analogiczną frazę nominalną traktuje jako niejednoznaczną:

(26)



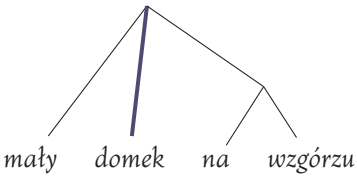
(27)



Tymczasem dotyczy jej ta sama argumentacja, co poprzednio: decyzja, czy to *mały domek* mieści się *na wzgórzu*, czy może *domek na wzgórzu* jest *mały*, nie wydaje się nieść istotnego rozróżnienia semantycznego.

W niniejszej pracy przyjęto konsekwentną interpretację niebinarną. Tak więc konstrukcje zdaniowe otrzymują interpretacje jak na rysunku 2.3, a fraza (26) jest interpretowana jako:

(28)



Ta konwencja strukturyzacyjna zastosowana do fraz może budzić pewne wątpliwości. Zdarza się mianowicie, że jeden z podrzędników z nadrzędnikiem tworzą wyraźną całostkę znaczeniową (w szczególności termin fachowy):

(29) *nieoczekiwany wylew krwi do mózgu*

(30) *rozległy wylew krwi do mózgu*

(31) *złośliwy owczarek niemiecki*

(32) *złośliwy nowotwór mózgu leczony chemią*

(33) *niebieski samochodzik na resorach [=resorak]*

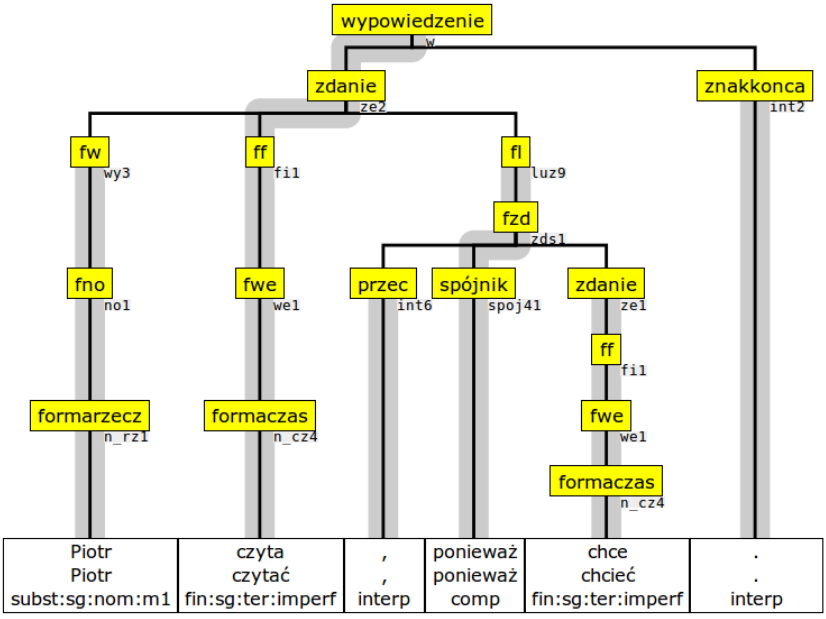
(34) *młoda para żądna zachwytów*

Przy przyjętej interpretacji wszystkie podrzędniki (np. *nieoczekiwany*, *krwi* i *do mózgu*) są traktowane tak samo i nie widać w strukturze drzewa, że niektóre podrzędniki są częścią terminu, a niektóre nie. Można jednak twierdzić, że pokazana struktura dobrze reprezentuje zależności czysto składniowe, natomiast omawiana kwestia należy do poziomu semantycznego, który powinien być reprezentowany osobno właśnie dlatego, że zestawienia słów o takich samych charakterystykach fleksyjnych czasem wiążą się z naddatkiem znaczeniowym, a czasem nie.

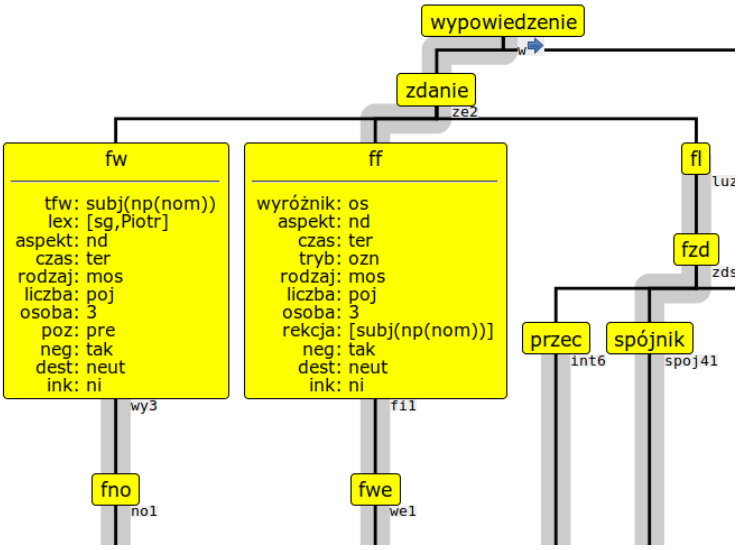
Świdziński przyjął jako narzędzie swojego opisu formalizm gramatyk metamorficznych (ang. *metamorphosis grammars*, Colmerauer 1978) obecnie lepiej znanych w wariancie *definite clause grammars* (DCG Pereira i Warren 1980). Jest to formalizm składnikowy, w związku z czym generowaną przez gramatykę Świdzińskiego reprezentacją struktur składniowych polszczyzny są drzewa składników bezpośrednich. Tę samą reprezentację przyjęto w niniejszym opisie. W bieżącym punkcie zostanie ona przedstawiona pobieżnie, a szczegóły można znaleźć w kolejnych punktach.

Rysunek 2.4 przedstawia przykładowe drzewo składników bezpośrednich wygenerowane przez analizator Świgra 2. Jednostkami terminalnymi gramatyki są formy fleksyjne w sensie przedstawionym w rozdziale 1, czyli trójki uporządkowane złożone z segmentu, lematu identyfikującego leksem i znacznika fleksyjnego identyfikującego cechy gramatyczne danej formy fleksyjnej (por. p. 1.11).

Zbudowane nad jednostkami terminalnymi drzewo składników składa się z węzłów reprezentujących jednostki nieterminalne gramatyki. Wszystkie one są bytami tego samego rodzaju, ale dla poglądowości opisu można je podzielić na kilka klas. Węzły należące do poszczególnych klas zwykle układają się w warstwy, jest to jednak przybliżenie; w bardziej skomplikowanych



Rysunek 2.4. Drzewo dla zdania *Piotr czyta, ponieważ chce.* wygenerowane przez analizator Świgra 2



Rysunek 2.5. Fragment drzewa z rys. 2.4 z uwidocznionymi atrybutami wybranych wierzchołków

konstrukcjach warstwy mogą się przeplatać. Na rysunku 2.4, jak na większości ilustracji w tej książce, pokazywane są jedynie nazwy jednostek nieterminalnych związanych z węzłami, jednak informacja w każdym węźle jest dużo bogatsza. Obejmuje ona wiele atrybutów (cech) o charakterze fleksyjnym i składniowym. Na rysunku 2.5 pokazano wartości atrybutów związanych z niektórymi węzłami przykładowego drzewa.

Najniższa warstwa nieterminali służy do przetransponowania form fleksyjnych na poziom składniowy. Należą do niej takie jednostki, jak: **formarzecz**, **zaimrzecz**, **formaczas**, **przyimek**, **formaprzym**, **zaimprzym**, **formaprzys**, **partykuła**, **przec**, które będą nazywane *formami składniowymi*. Najprostsze jednostki tej warstwy są realizowane przez pojedynczą formę fleksyjną wyróżnioną w analizowanym tekście. Przyjęty w niniejszej pracy podział na segmenty jest jednak bardzo drobny (por. p. 1.5). Na przykład czas przeszły czasowników jest wyrażany dwiema następującymi po sobie formami fleksyjnymi (por. p. 1.7.2). W warstwie form składniowych te formy fleksyjne są łączone, aby w dalszej analizie mogły pełnić funkcję jednostki składniowej. Ta warstwa obejmuje również produktywne złożenia przymiotnikowe typu *polsko-ukraińsko-rosyjski*, różne ustalone zestawienia słowne jak *na pewno*, *z nagła* (które jako całość stanowią realizacje jednostki **formaprzys**), jak również analityczne formy czasowników (ale por. p. 2.13).

Poziom form składniowych plasuje się gdzieś między fleksją a składnią. Niektóre z nich dopuszczają rekurencyjne zagnieżdżenie, co sugerowałoby naturę składniową – dzieje się tak w wypadku wspomnianych złożonych form przymiotnikowych. Konstrukcje te wykazują jednak wyraźne ograniczenia: ich

składnikiem musi być pojedynczy segment przymiotnikowy, nie da się substytuować go frazą (np. możliwa jest formacja *biało-czerwony*, ale już nie **biało-ciemno czerwony*). W wielosegmentowych formach składniowych często trudno jest wskazać centra składniowe. Konstrukcje te są więc odrębne od wyższych poziomów opisu, obejmują także wiele nietypowości i uwarunkowań leksykalnych, które muszą być odnotowane w słowniku (np. *na pewno* jest połączeniem segmentu wyglądającego na wykładnik przyimka z segmentem wyglądającym na wykładnik przysłówka; a formy tych klas nie tworzą razem konstrukcji).

Drugi poziom w drzewie obejmuje frazy składnikowe, na przykład czasownikowe **fwe**, przymiotnikowe **fpm**, nominalne **fno**, zdaniowe **fzd**. Frazy te wyróżnia się ze względu na klasę gramatyczną ich centrum. Składniki na tym poziomie mogą uzyskiwać dowolny poziom komplikacji dzięki modyfikującym podrzędnikom (w szczególności niektóre z nich mogą zawierać podrzędniki zdaniowe) i wchodzeniu w konstrukcje skoordynowane (ze spójnikami takimi jak I lub ALBO).

Trzeci poziom reprezentuje strukturę argumentową. Należy do niego fraza finitywna **ff**, fraza wymagana **fw** i fraza luźna **fl**. Te trzy składniki służą do zbudowania finitywnej realizacji zdania. Wprowadzenie tych fraz pozwala uwidocznić schematy walencyjne czasowników w samej strukturze drzewa. Podmiot jest reprezentowany przez jedną z fraz wymaganych. Frazy wymagane są także używane do reprezentacji wymagań walencyjnych form innych klas gramatycznych.

Czwarty poziom obejmuje jednostki zdaniowe. Są one budowane z jednostek poziomu trzeciego albo też poprzez łączenie zdań. Korzeniem drzewa zgodnie z koncepcją Świdzińskiego jest jednostka **wypowiedzenie**.

Znaki interpunkcyjne są traktowane jako pełnoprawne składniki drzew składniowych. W szczególności gramatyka jest wyposażona w dość rozbudowany mechanizm odpowiadający za wymaganie obecności przecinków w odpowiednich miejscach wypowiedzenia (por. p. 4.4.5).

Istotną nowością w stosunku do opisu Świdzińskiego jest wskazywanie centrów składniowych poszczególnych jednostek. Na rysunku „pogrubiona" szara linia łączy każdy wierzchołek drzewa z jego składnikiem stanowiącym centrum. Linie te łączą się, tworząc ścieżki od korzenia każdego poddrzewa do formy fleksyjnej stanowiącej centrum danej konstrukcji.

Jednostkom nieterminalnym są przypisywane liczne atrybuty (w GFJP zwane parametrami) opisujące cechy fleksyjne i składniowe poszczególnych jednostek. Oprócz kategorii fleksyjnych jak rodzaj, liczba czy osoba pojawiają się tu atrybuty czysto składniowe. Na przykład parametr predestynacji o wartościach „neutralna", „pytajna", „pytajnozależna" i „względna" sygnalizuje konstrukcje, które mogą wystąpić wyłącznie w kontekstach o odpowiedniej charakterystyce (por. p. 4.4.1).

2.4. Hierarchia jednostek

Gramatyka Świdzińskiego (1992), podobnie jak gramatyka Szpakowicza (1983), operuje bardzo rozbudowaną hierarchią jednostek. Każdemu z typów fraz składnikowych odpowiada hierarchia kilku jednostek składniowych. Podobnie jest ze zdaniami. Jednostki tworzące owe hierarchie w GFJP wymieniono na rysunkach 2.6 i 2.7[6].

Jednostki tworzące hierarchie zdaniowe odpowiadają typom konstrukcji współrzędnych według GFJP (zob. p. 2.9). Jednostki z hierarchii składnikowych są związane z binarną koncepcją strukturyzacji tych konstrukcji. Typy jednostek odpowiadają poszczególnym podrzędnikom, które mogą wystąpić przy danym nadrzędniku. Tak więc konstrukcja nominalna z dopełniaczem została przewidziana, aby zapewnić podłączanie podrzędników nominalnych w dopełniaczu do centrum nominalnego. Podobnie konstrukcje z frazą przyimkową, z atrybutem (podrzędnikiem przymiotnikowym), z frazą nominalną i z frazą przysłówkową służą do podłączania wymienionych typów podrzędników.

Taka organizacja jednostek powoduje tworzenie bardzo długich gałęzi w drzewach składniowych. Jeżeli na przykład fraza nominalna jest realizowana bez podrzędników przez jedną rzeczownikową formę fleksyjną, musi ona zostać zinterpretowana jako wszystkie pośrednie poziomy hierarchii (zob. poddrzewo dla jednostki **fno** realizowane przez segment *obrazkami* na rysunku 2.9, s. 93).

Co więcej, pięć z wymienionych hierarchii (dla zdań, fraz zdaniowych, fraz nominalnych, fraz przymiotnikowych i fraz przysłówkowych) stanowi cykle. Reguły GFJP przewidują, że każda z wymienionych jednostek w najprostszym przypadku może być realizowana przez samotne wystąpienie jednostki z poziomu o jeden niższego. Dodatkowo jednostka najniższa w hierarchii może być realizowana przez jednostkę najwyższą. Taka organizacja powoduje znaczne trudności realizacji komputerowej gramatyki. Przede wszystkim jednak wydaje się nie mieć uzasadnienia teoretycznego. Cykl oznacza bowiem, że jednostki te można swobodnie na siebie przepisywać, a więc że każde wystąpienie można zastąpić dowolnie wybraną inną jednostką z tego cyklu.

Świdziński pisze, że w jego ujęciu jednostka składniowa jest rozumiana jako „klasa abstrakcji wyrażeń niemal równoważnych dystrybucyjnie" (1992, §3.3, s. 59). Jednak możliwość swobodnego przepisywania jednostek oznacza, że są one dystrybucyjnie w pełni równoważne, mogą występować w dokładnie tych samych kontekstach i pełnią tę samą funkcję. Można z tego wysnuć wniosek, że wprowadzone typy jednostek są zbyt szczegółowe i przekraczają granicę szczegółowości reprezentowalnej w stosowanym formalizmie, jaką jest równoważność dystrybucyjna (por. Woliński 2004, p. 5.1).

---

[6] W GFJP termin *fraza* jest używany tylko na określenie jednostki najbardziej rozbudowanej, jej mniej rozbudowane czy też niepełne warianty są nazywane *konstrukcjami*.

zdanie równorzędne **zr**
zdanie szeregowe **zsz**
zdanie jednorodne **zj**
zdanie proste **zp**
zdanie elementarne **ze**

fraza zdaniowa **fzd**
fraza zdaniowa szeregowa **fzdsz**
fraza zdaniowa jednorodna **fzdj**
fraza zdaniowa z korelatem **fzdkor**
fraza zdaniowa elementarna **fzde**

Rysunek 2.6. Hierarchie zdań i fraz zdaniowych w GFJP

fraza nominalna **fno**
konstrukcja nominalna z dopełniaczem **knodop**
konstrukcja nominalna z frazą przyimkową **knopm**
konstrukcja nominalna z atrybutem **knoatr**
konstrukcja nominalna z inkorporacją **knoink**
konstrukcja nominalna **knom**

fraza przymiotnikowa **fpt**
konstrukcja przymiotnikowa z frazą nominalną **kptno**
konstrukcja przymiotnikowa z frazą przyimkową **kptpm**
konstrukcja przymiotnikowa z frazą przysłówkową **kptps**
konstrukcja przymiotnikowa z inkorporacją **kptink**
konstrukcja przymiotnikowa **kprzym**

fraza przysłówkowa **fps**
konstrukcja przysłówkowa z frazą przyimkową **kpspm**
konstrukcja przysłówkowa z frazą przysłówkową **kpsps**
konstrukcja przysłówkowa z inkorporacją **kpsink**
konstrukcja przysłówkowa **kprzysł**

Rysunek 2.7. Hierarchie fraz składnikowych w GFJP

W prezentowanej tu gramatyce przyjęto skromniejszy zbiór jednostek. Jest tylko jedna jednostka reprezentująca zdanie (nazwana **zdanie**). Podobnie wszystkie konstrukcje realizujące dany typ frazy składnikowej zostały połączone w jedną jednostkę odpowiadającą temu typowi (por. Świdziński i Woliński 2009). Dzięki temu tworzone drzewa składnikowe mają dużo mniejszą wysokość niż w gramatyce Świdzińskiego, unika się sztucznych poziomów hierarchii, a tworzone struktury są czytelniejsze i bardziej zrozumiałe, bez utraty zawartości informacyjnej.

2.5. Typy fraz składnikowych

Typy fraz składnikowych zostały wyróżnione ze względu na ich budowę wewnętrzną, przede wszystkim ze względu na charakterystykę gramatyczną centrum frazy. Świdziński (1992, §4.3.7, s. 76) wyróżnia następujące typy fraz składnikowych: werbalna, nominalna, przymiotnikowa, przysłówkowa, przyimkowa i zdaniowa. Frazy werbalne, nominalne, przymiotnikowe i przysłówkowe mają jako centrum odpowiednio formę czasownika, rzeczownika, przymiotnika lub przysłówka. Frazy te są redukowalne do swojego centrum składniowego. Fraza przyimkowa (w istocie przyimkowo-nominalna) jest nieredukowalna, składa się z przyimka (który jest oznaczany jako centrum frazy) i frazy nominalnej w przypadku wymaganym przez ten przyimek.

Frazy zdaniowe są pojęciem wprowadzonym w celu reprezentowania zdań podrzędnych. Uwzględniono frazy zdaniowe spójnikowe, względne i pytajnozależne:

(35) *Pamięta , że pił alkohol.*

(36) *Trudno jest też prywatyzować , ponieważ roszczenia dawnych właścicieli zawsze będą wracały.*

(37) *Za sobą mieli martwą rzekę , która leżała wśród piasków jak rozbity dzban.*

(38) *Milicja szuka teraz po tamtej stronie , gdzie ją znaleźli.*

(39) *Nie napisał mi Pan , dlaczego się Pan przenosi do Warszawy.*

(40) *Mimo że nadal nie wiadomo , jak będzie wybierany przyszły prezydent miasta, salony polityczne Szczecina są pełne spekulacji na ten temat.*

Frazy zdaniowe spójnikowe (przykłady (35) i (36)) składają się ze spójnika podrzędnego (uznanego za centrum składniowe) i wprowadzanego przezeń zdania, a tym samym są nieredukowalne. Frazy względne (przykłady (37) i (38)) i pytajnozależne (przykłady (39) i (40)) są redukowalne do swojego jedynego (oprócz interpunkcyjnych przecinków) składnika, którym jest zdanie o szczególnej postaci – względne lub pytajnozależne. Zdanie takie jako inicjalny składnik musi mieć leksykalny element względny lub pytajny.

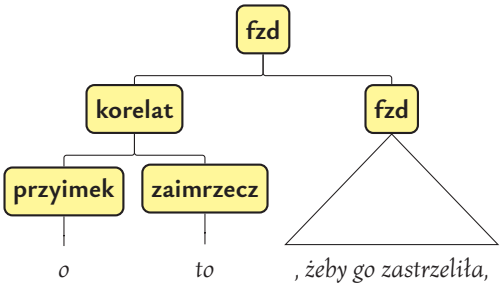
Wariantem frazy zdaniowej jest fraza zdaniowa z korelatem. Korelat to forma zaimka TO lub fraza przyimkowo-nominalna z formą zaimka TO (Świdziński 1992, §8.5, s. 292).

(41) *W dokumentach dowódcy szczycą się tym, że stłumili bunt bez użycia broni.*

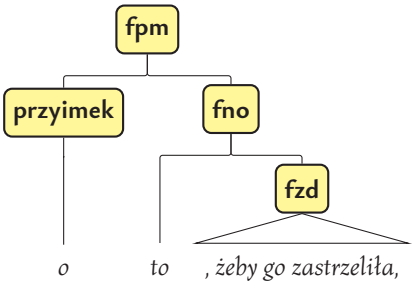
(42) *– No wie Pan, jeśli prokurator prosi postronną osobę o to, żeby go zastrzeliła, to chyba jest o czym pisać...*

Frazy zdaniowe z korelatem są osobnymi typami wymagań, w szczególności od konkretnego czasownika zależy postać dopuszczalnego korelatu.

Strukturę wewnętrzną fraz z korelatem przyjętą przez Świdzińskiego można przedstawić na przykładzie frazy *o to, żeby go zastrzeliła*, następująco:



Jednostka **korelat** może być realizowana wyłącznie przez formy zaimka TO (ewentualnie z przyimkami). Alternatywnie frazę tę można by strukturyzować następująco (taką interpretację przyjęto w gramatyce POLFIE):



W tej strukturze umieszczono frazę zdaniową jako podrzędnik TO (co oznacza uznanie istnienia frazy *to, żeby...*), a całość jest interpretowana jako fraza przyimkowa. Należy jednak zwrócić uwagę, że taka fraza nie jest równoważna typowej frazie przyimkowo-nominalnej. Niemożliwe jest na przykład zdanie

(43) **Rozłupała ją o to, żeby go zastrzeliła.*

ponieważ czasownik ROZŁUPAĆ dopuszcza jako podrzędnik frazę przyimkowo-nominalną, ale nie frazę zdaniową z korelatem. Oznacza to, że niezależnie od tego, która interpretacja zostanie przyjęta, konstrukcja z korelatem musi być odróżniana przez gramatykę jako twór dopuszczalny w specyficznych kontekstach: albo jako szczególny typ frazy zdaniowej, albo jako specyficzny wariant frazy przyimkowo-nominalnej. W niniejszej pracy przyjęto za GFJP pierwszą strukturę.

Repertuar fraz składnikowych został uzupełniony o frazy przyimkowo-przymiotnikowe:

(44) *Zarząd uznał dowcip za wysoce niestosowny i zareagował śmiertelnie poważnym pismem.*

(45) *Wobec wniosku o uznanie tej odpowiedzi za wystarczającą zgłoszono sprzeciw.*

Ważnym rozwinięciem gramatyki w stosunku do GFJP jest uwzględnienie w opisie form liczebnikowych (Świdziński i Woliński 2009). Formy takie zamie-

niają się na poziomie składniowym w wystąpienia frazy liczebnikowej **flicz**. Reprezentuje ona składniowe liczebniki proste (np. *dziewiętnastoma*) i złożone (np. *czterysta dwudziestu dwóm tysiącom siedemset trzydziestu jeden*). Fraza nominalna może być realizowana przez frazę liczebnikową i (opcjonalną) frazę rzeczownikową. W takiej realizacji nadrzędnikiem i reprezentantem jest fraza liczebnikowa (por. p. 2.8.3).

Imiesłowy przymiotnikowe czynne i bierne są traktowane jako przymiotniki, a więc stają się centrami fraz przymiotnikowych, natomiast odsłowniki (gerundia) są traktowane jako rzeczowniki i stają się centrami fraz nominalnych. Przy tym dziedziczą one właściwości walencyjne od odpowiednich schematów dla form finitywnych czasownika, od którego są derywowane (por. 3.1.10).

Pełną listę typów fraz składnikowych przyjętych w niniejszej pracy przedstawiono w tabeli 2.1.

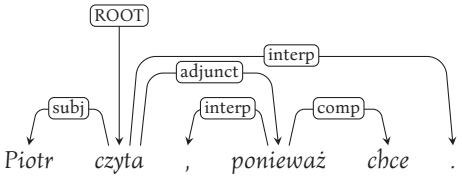
Tabela 2.1. Typy fraz składnikowych, odpowiadające im jednostki nieterminalne gramatyki oraz formy składniowe mogące stanowić centra odpowiednich fraz

typ frazy	jednostka	centrum
fraza werbalna	**fwe**	**formaczas**
fraza nominalna	**fno**	**formarzecz/zaimos**
fraza liczebnikowa	**flicz**	**formalicz**
fraza przymiotnikowa	**fpt**	**formaprzym**
fraza przysłówkowa	**fps**	**formaprzys**
fraza przyimkowo-nominalna	**fpm**	**przyimek**
fraza przyimkowo-przymiotnikowa	**fpmpt**	**przyimek**
fraza zdaniowa	**fzd**	(**spójnik**)

2.6. Wyróżnianie centrów

Istotną innowacją proponowanego opisu jest jawne oznaczanie centrów składniowych. Świdziński, komentując reguły GFJP, mówi o istotnym wpływie nadrzędnika konotacyjnego, jednak nie oznacza go *explicite*. W regułach opisujących konstrukcje podrzędne jednostka po lewej stronie reguły z zasady dziedziczy komplet parametrów od tego elementu prawej strony, który stanowi nadrzędnik i reprezentanta konstrukcji. Sprawia to, że frazy rozbudowują się niejako wokół składnika stanowiącego centrum, co przypomina nieco teorię X-bar (Jackendoff 1977). W drzewach generowanych przez analizator Świgra 2 centrum konstrukcji oznaczane jest jawnie.

Jawne oznaczenie centrów umożliwia automatyczną konwersję uzyskiwanych struktur na drzewa zależnościowe dzięki temu, że oznaczenia centrów na poszczególnych poziomach tworzą ścieżki od korzenia danego poddrzewa do segmentu stanowiącego rekurencyjnie rozumiane centrum konstrukcji.



Rysunek 2.8. Drzewo zależnościowe odpowiadające drzewu składnikowemu przedstawionemu na rysunku 2.4 (s. 83)

Każdy wierzchołek wewnętrzny drzewa składnikowego wnosi więc informację pozwalającą utworzyć krawędzie zależnościowe od segmentu stanowiącego rekurencyjne centrum danej frazy do rekurencyjnie rozumianych centrów pozostałych składników.

Na przykład w drzewie przedstawionym na rysunku 2.4 (s. 83) ścieżka rekurencyjnie wyznaczająca centra schodzi z korzenia (**wypowiedzenie**) do formy *czyta*, więc to ona stanowi korzeń drzewa zależnościowego. Wierzchołek **wypowiedzenie** ma podrzędnik **znakkonca**, więc w drzewie zależnościowym znajdzie się krawędź łącząca *czyta* z końcową kropką. Wierzchołek **zdanie** ma niebędące centrum składniki **fw** i **fl**, więc ich rekurencyjnie rozumiane centra – odpowiednio formy fleksyjne *Piotr* i *ponieważ* – staną się punktami docelowymi kolejnych krawędzi wychodzących od formy *czyta*. Uwzględnienie informacji wprowadzanych przez wszystkie wierzchołki drzewa składnikowego prowadzi do drzewa zależnościowego przedstawionego na rysunku 2.8. Warto zauważyć, że struktura zależnościowa „spłaszcza" frazy: w drzewie z rysunku 2.8 trzy równoprawne krawędzie wychodzą z wierzchołka *czyta*, jednak ich pochodzenie jest różne – dwie odpowiadają zależności między składnikami zdania, trzecia zaś – między całym zdaniem a znakiem interpunkcyjnym.

2.7. Struktura zdania elementarnego

Za Świdzińskim (1992) zdaniem elementarnym będzie w tej pracy nazywana realizacja zdania, której centrum jest finitywna forma czasownika. Jednak struktury przypisywane zdaniom elementarnym w niniejszej pracy są istotnie inne niż w GFJP.

Zdanie elementarne w opisie Świdzińskiego (Świdziński 1992, rozdz. 6 i Aneks p. 5.6–5.8) składa się z czasownikowej frazy finitywnej **ff** i trzech fraz wymaganych **fw**. Składnikiem zdania elementarnego może być także fraza luźna **fl**, jeżeli występuje na początku zdania. Frazy wymagane odpowiadają argumentom czasownika (podmiotowi i dopełnieniom), fraza luźna reprezentuje podrzędniki nie będące argumentami (okoliczniki) oraz inne elementy występujące w strukturze zdania, które nie są podrzędnikami czasownika

(np. wtrącenia). Każda z fraz wymaganych może mieć realizację pustą. Elementy niewymagane zajmujące pozycje inne niż inicjalna są reprezentowane w postaci fraz luźnych będących składnikami frazy finitywnej lub frazy wymaganej, bezpośrednio po której występują w zdaniu. Jeżeli obok siebie występują co najmniej dwie frazy luźne, to każda kolejna z nich staje się składnikiem bezpośrednim poprzedniej. Przykład takiej struktury przedstawia rysunek 2.9.

Można wskazać kilka wad tego opisu (pomijając problemy techniczne powodujące generowanie zbyt dużej liczby interpretacji, gdyby reguły Świdzińskiego traktować literalnie, por. Woliński 2004, p. 5.2.2).

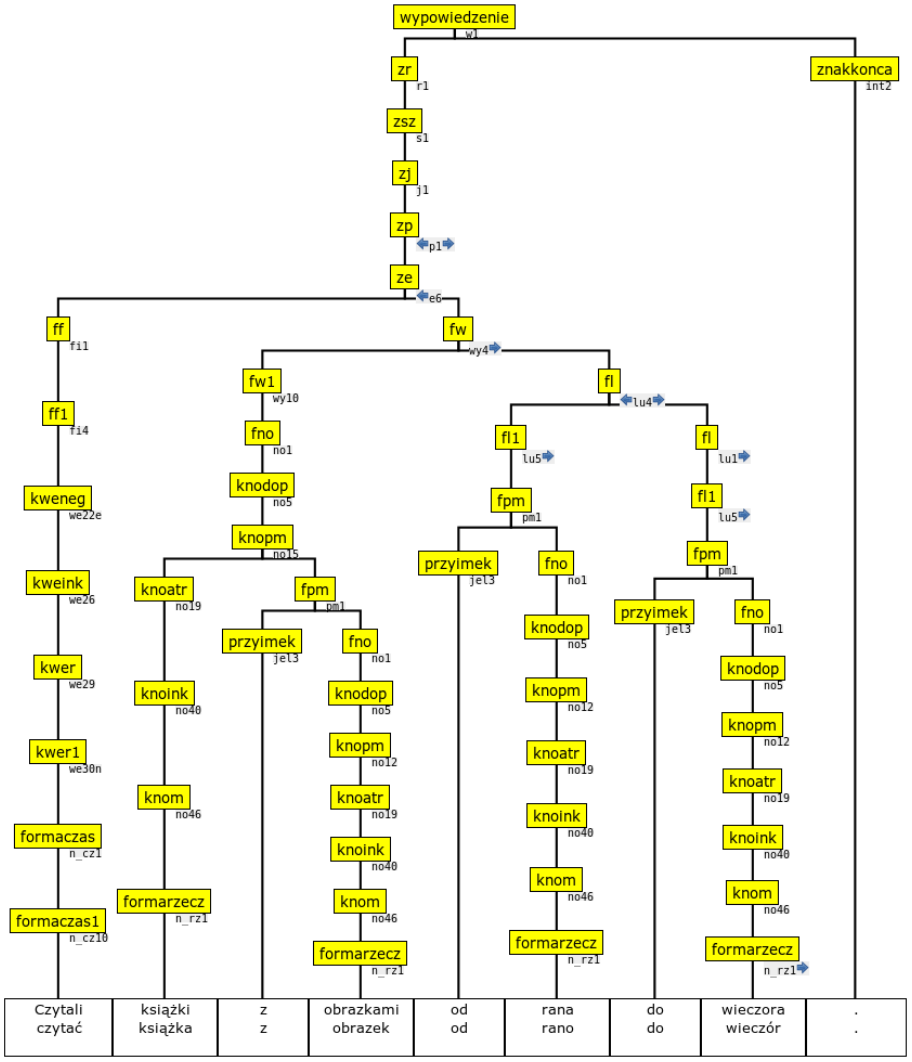
Uczynienie z fraz luźnych składników fraz wymaganych spowodowało pojawienie się w drzewie wierzchołków, które dominują nad fragmentami wypowiedzenia, co do których trudno byłoby wykazać, że są składnikami. Na przykład w strukturze przedstawionej na rysunku 2.9 węzeł frazy wymaganej **fw** dominuje nad fragmentem wypowiedzenia *książki z obrazkami od rana do wieczora*. Według przedstawionej tam interpretacji jest on składnikiem bezpośrednim zdania elementarnego **ze**. Z semantycznego punktu widzenia określenie *od rana do wieczora* z pewnością nie odnosi się do książek. A również składniowo określenie to (stanowiące dwa osobne podrzędniki będące frazami przyimkowo-nominalnymi) powinno mieć jako nadrzędnik frazę finitywną *czytali* lub jakiś składnik z tą frazą jako centrum.

Kolejnym problemem opisu jest sztywne ustalenie liczby fraz wymaganych – trzy, wliczając podmiot. Słownik Walenty przewiduje schematy zawierające do siedmiu argumentów. Tak więc część z nich nie może być objęta przez reguły GFJP.

Można postawić hipotezę, że ta postać opisu jest wynikiem ograniczeń stosowanego formalizmu. Przede wszystkim gramatyki metamorficzne operują regułami o ustalonej liczbie składników po prawej stronie. Każda reguła gramatyki tworzy więc wierzchołki drzewa o ustalonej liczbie wierzchołków podrzędnych. Aby zapewnić tworzenie węzłów drzewa o różnej liczbie dzieci, konieczne byłoby wprowadzenie odpowiednio wielu osobnych reguł, z których każda realizowałaby daną liczbę składników. Tak zapisana gramatyka byłaby trudna w utrzymaniu i nieczytelna.

Wprowadzenie w GFJP fraz luźnych do wnętrza fraz wymaganych i frazy finitywnej pozwoliło ograniczyć skład zdania oraz ułatwiło zapis i tak skomplikowanych reguł opisujących zdanie elementarne. Aby uniknąć tego samego problemu w obrębie fraz wymaganych, Świdziński zdecydował się zagnieżdżać frazy luźne rekurencyjnie w sobie. Inicjalna fraza luźna została dopuszczona, ponieważ w tej pozycji nie ma poprzedzającej frazy, w którą mogłaby się wczepić. Powstałe w ten sposób reguły opisują właściwe ciągi form fleksyjnych, gramatyka w tym sensie faktycznie modeluje więc język polski. Niestety tworzonych struktur nie można uznać za adekwatne drzewa składników bezpośrednich.

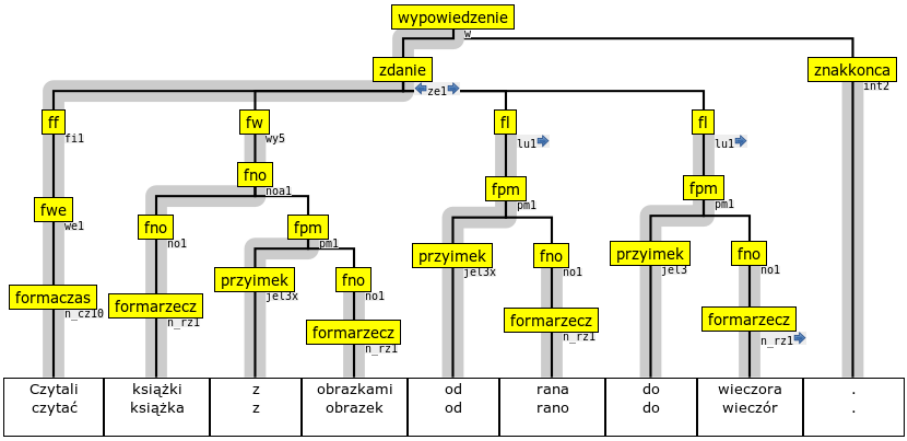
Przyjęta w niniejszej pracy struktura zdań jest znacząco inna (por. rys. 2.10). Mianowicie fraza finitywna i wszystkie frazy wymagane i luźne składające się



Rysunek 2.9. Analiza zdania zawierającego frazę wymaganą i frazy luźne według GFJP

na dane zdanie stały się składnikami na tym samym poziomie. Liczba podrzędników czasownika nie jest ograniczona. Frazy wymagane są dopuszczane w liczbie wynikającej ze schematu walencyjnego danego czasownika. Frazy luźne dopuszczane są dowolnie[7]. Kolejność fraz względem siebie jest dowolna, co dobrze modeluje swobodny szyk zdania polskiego. (Uwzględniono przy tym

[7] Analizator ogranicza liczbę fraz luźnych, które mogą wystąpić w jednym zdaniu, do trzech. Stanowi to sensowną heurystykę ograniczającą liczbę nadmiarowych interpretacji. Parametr ten można łatwo zmienić, gdyby ograniczenie okazało się zbyt silne.



Rysunek 2.10. Analiza zdania z rys. 2.9 według analizatora Świgra 2

ograniczenie położenia nietypowych nominalnych fraz podmiotowych względem frazy finitywnej, por. p. 2.9.3). Dodatkowym typem frazy, który może pojawić się jako składnik zdania elementarnego, jest jednostka **posiłk** realizująca formy posiłkowe stanowiące części nieciągłych form analitycznych czasowników (por. p. 2.13).

W tym miejscu warto też wspomnieć o atrybutach fraz wymaganych **fw** i luźnych **fl**, są one bowiem nadawane na innej zasadzie niż w pozostałych jednostkach gramatyki. O ile w wypadku fraz składnikowych ich atrybuty zdają sprawę z cech konkretnej frazy, to atrybuty fraz wymaganych i luźnych są kopiami odpowiednich atrybutów frazy finitywnej i zdania (zob. np. atrybuty frazy **fw** na rysunku 2.5, s. 84). Są więc dziedziczone „w dół", a nie „w górę". Rozwiązanie to, przejęte z GFJP, pozwala na dostęp do wybranych atrybutów zdania w regułach opisujących realizację fraz wymaganych i luźnych. Mechanizm ten został wprowadzony, aby umożliwić uzgodnienie atrybutów osoby, rodzaju i liczby w regule opisującej realizację frazy wymaganej podmiotowej. Służy on także do opisania tzw. dopełniacza negacji (por. p. 4.4.2) i interakcji parametru negacji wymaganej frazy bezokolicznikowej z negacją zdania (także p. 4.4.2). W odniesieniu do fraz luźnych pozwala opisać następującą zależność: jeżeli fraza finitywna jest w drugiej osobie, to musi zachodzić zgodność rodzaju między formą czasownika i luźną frazą wołaczową:

(46) *Przyszedłeś do mnie, chłopcze.*
(47) **Przyszedłeś do mnie, dziewczyno.*

2.8. Schematy strukturyzacyjne konstrukcji podrzędnych

Cechą charakterystyczną konstrukcji podrzędnych jest to, że dokładnie jeden ze składników konstrukcji stanowi jej dystrybucyjnego reprezentanta, a więc ma cechy identyczne lub bardzo zbliżone do konstrukcji jako całości (por. p. 2.1.6). Reprezentanta tego uznano jednocześnie za jej centrum składniowe. Ważnym przykładem konstrukcji podrzędnej jest zdanie elementarne omówione w poprzednim punkcie. Schemat ten jest również stosowany we wszystkich typach fraz składnikowych.

W tym punkcie najpierw zostaną omówione typowe podrzędniki we frazach składnikowych, a w osobnych podpunktach będą wymienione konstrukcje, które stanowią rozszerzenie opisu względem GFJP.

2.8.1. Typowe konstrukcje podrzędne

W konstrukcji podrzędnej repertuar dopuszczalnych podrzędników zależy od typu nadrzędnika, a także od konkretnej jednostki leksykalnej stanowiącej centrum nadrzędnika. W GFJP leksykalnie uwarunkowane wymagania są przypisywane jedynie czasownikom. Potrzeba opisu wymagań konkretnej jednostki jest jednak widoczna i dla leksemów innych klas.

Przykład stanowią leksemy derywowane od czasowników, które w tym procesie przejęły niektóre wymagania swoich podstaw derywacyjnych: rzeczownik DOWÓD w przykładzie (48) wymaga jako podrzędnika frazy zdaniowej ze spójnikiem ŻE – niemożliwej przy innych rzeczownikach (por. **stół, że...*). Rzeczownik PRZEJAZD w (49) wymaga frazy nominalnej w narzędniku, która nie jest typowym podrzędnikiem rzeczowników. W zdaniu (50) przymiotnik ŻĄDNY wymaga frazy nominalnej w narzędniku, w (51) zaś fraza zdaniowa ze spójnikiem ŻE jest wymagana przez przymiotnik WŚCIEKŁY.

(48) *Kotlarczyk zdobył dowody, że Majda kradnie teksty do swojej habilitacji.* [NKJP300]

(49) *Podjeżdżał już do kasjerki pobierającej opłatę za przejazd autostradą.* [NKJP300]

(50) *Tutaj żądny mięsa lub skór łowca sam stawał się obiektem łowów.* [NKJP300]

(51) *[...] za nimi ciągnął zawsze sznur ubeków, wściekłych, że zza krzaków czy łanów zboża niczego nie można podsłuchać.* [NKJP300]

Tego rodzaju wymagania uwzględniono w analizatorze Świgra 2, a ich opis został następnie rozwinięty w słowniku walencyjnym Walenty, który obejmuje jako typy nadrzędników czasowniki, rzeczowniki, przymiotniki i przysłówki (zob. rozdz. 3).

Warto zaznaczyć, że o ile w wypadku czasowników zróżnicowanie dopuszczalnych przy nich podrzędników jest duże i praktycznie każda jednostka czasownikowa powinna znaleźć się w słowniku, to w wypadku leksemów innych klas obecność elementów wymaganych jest raczej wyjątkiem i tylko niektóre leksemy wymagają szczegółowego opisu.

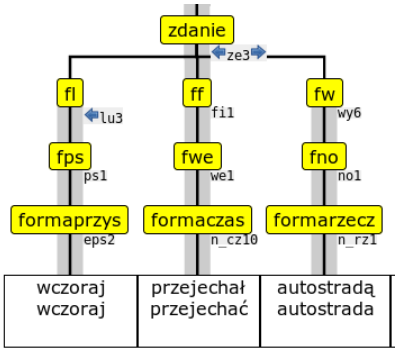
Dla leksemów nieczasownikowych rozróżnienie podrzędników wymaganych i luźnych nie jest tak wyraziste jak dla czasowników. O ile podrzędniki nietypowe postulowane przez słownik walencyjny dla danej jednostki leksykalnej należy interpretować jako wymagane, to status podrzędników typowych jest problematyczny. Uznanie ich za wymagane może prowadzić do niespójności z traktowaniem ich przy tych jednostkach, których nie ma w słowniku walencyjnym i w związku z tym nie ma dla nich informacji o statusie podrzędników.

Dlatego w strukturach generowanych przez analizator Świgra 2 przyjęto różną reprezentację w zależności od typu frazy. Co do zasady stosowane jest przyjęte za GFJP pojęcie fraz wymaganych i luźnych, pozwalające pokazać w strukturze drzewa strukturę argumentową. W wypadku konstrukcji, których centrum jest forma czasownika, a więc fraz werbalnych **fwe** i zdań **zdanie**, podrzędnikami mogą być frazy wymagane **fw** i luźne **fl** (oraz wspomniana już jednostka **posiłk**, zob. p. 2.13). Dla fraz składnikowych nominalnych, przymiotnikowych i przysłówkowych frazy dopuszczone przez słownik walencyjny są oznaczane jako wymagane **fw**, a pozostałe frazy, których typy zależą od typu nadrzędnika, są wprowadzane bez pośredniczącego poziomu frazy wymaganej lub luźnej (por. struktury na rys. 2.12 i 2.13). Nadrzędnikiem zdania jest fraza finitywna **ff**, nadrzędnikiem fraz składnikowych jest prostsza fraza składnikowa tego samego typu. Przykłady struktur utworzonych według tych zasad przedstawiono na rysunkach 2.11–2.13.

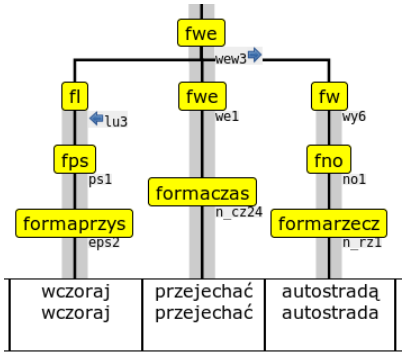
Podczepianie podrzędników wewnątrz frazy werbalnej jest dopuszczone w dwóch sytuacjach: gdy stanowiąca centrum forma jest niefinitywna, więc gdy nie ma ona szansy stać się centrum zdania, oraz gdy jest ona składnikiem frazy skoordynowanej, której tylko niektóre składniki są wspólne w obrębie koordynacji (por. p. 2.9.5). Podrzędniki form finitywnych podczepiane są na poziomie zdania.

Typy fraz wymaganych przewidziane w słowniku Walenty wymieniono w punkcie 3.1.2. Frazy luźne realizujące niewymagane podrzędniki czasownika obejmują następujące konstrukcje:

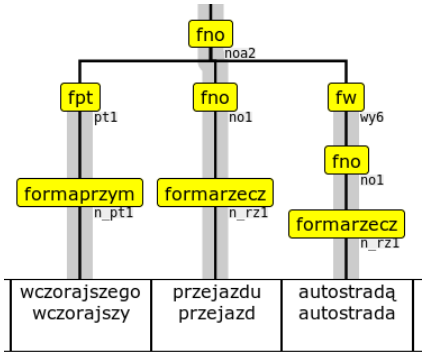
– frazy przysłówkowe (w tym z centrum będącym imiesłowem przysłówkowym współczesnym i uprzednim),
– frazy przyimkowe,
– modyfikatory partykułowe (zob. p. 2.8.4),
– frazy przyimkowo-przymiotnikowe (np. *wolą płacić* *po staremu*),
– frazy nominalne w przypadkach zależnych (w zasadzie w celowniku, np. *opiszę* *ci* *to*; bierniku, np. *milczeć* *długą chwilę* i narzędniku, np. *miejscami* *wystąpią opady*),



Rysunek 2.11. Zdanie: nadrzędnikiem jest fraza finitywna **ff**, podrzędnikami – frazy luźne i wymagane



Rysunek 2.12. Fraza werbalna **fwe**: nadrzędnikiem jest prostsza fraza werbalna, podrzędnikami – frazy luźne i wymagane



Rysunek 2.13. Fraza nominalna **fno**: nadrzędnikiem jest prostsza fraza nominalna, podrzędnikami – fraza wymagana przewidziana przez słownik walencyjny i uzgodniona fraza przymiotnikowa **fpt** jako typowy podrzędnik nominalny. W wypadku tej ostatniej struktura nie determinuje, czy jest ona wymagana czy luźna

– frazy przymiotnikowe (np. *Dziewczyna przyszła zmęczona*.),
– frazy zdaniowe z niektórymi spójnikami (np. ŻEBY – *Zostawiła naczynia, żeby obeschły*.) i względne typu CO (np. *Było to błędem, co pokazał wynik wyborczy*.).

Frazy luźne realizujące elementy niebędące podrzędnikami to różnego rodzaju wtrącenia, m.in. frazy nominalne w wołaczu (, *chłopcze,*), wykrzykniki (, *psiakość,*) i wtrącenia zdaniowe (, *być może,*).

Dla frazy nominalnej **fno** dopuszczalne podrzędniki to (por. Świdziński i Woliński 2009, 2007):
– fraza przymiotnikowa **fpt** uzgodniona co do przypadka, liczby i rodzaju (np. *domek fiński*);
– fraza przyimkowa **fpm** (*domek na wzgórzu*);
– fraza nominalna **fno** w dopełniaczu (*domek Marysi*);
– fraza zdaniowa **fzd** względna (*domek, który chciał kupić,*);
– modyfikator partykułowy **modpart** (*jakby domek*, zob. p. 2.8.4);
– fraza nominalna apozycyjna (*domek zabawka*, zob. p. 2.8.2).

Podrzędnikiem we frazie przymiotnikowej **fpt** mogą być:
– fraza przysłówkowa **fps** (*bardziej zielony*);
– fraza przyimkowa **fpm** (*zielony od spodu*);
– modyfikator partykułowy **modpart** (*jakby zielony*);
– nietypowy modyfikator przymiotnikowy **modjaki** (*jakiś zielony*, p. 2.8.5).

Możliwymi podrzędnikami frazy przysłówkowej **fps** są:
– fraza przysłówkowa **fps** (*ogromnie ożywczo*);
– fraza przyimkowa **fpm** (*w miarę ożywczo*);
– modyfikator partykułowy **modpart** (*jakby ożywczo*).

Najmniejszy repertuar podrzędników jest możliwy we frazie przyimkowej **fpm**:
– fraza przysłówkowa **fps** (*wprost z rondla*);
– modyfikator partykułowy **modpart** (*jakby z rondla*).

2.8.2. Konstrukcje apozycyjne i pokrewne

Specyficznym typem fraz rzeczownikowych są w języku polskim konstrukcje apozycyjne (Kallas 1978, 1980). Oto przykłady konstrukcji tego typu:

(52) *A teraz ma krowę piwoszkę*. [Skł.]
(53) *Dziękuję, panie pośle*. [Skł.]
(54) *Może pan weryfikator nie wie, co podpisuje?* [Skł.]
(55) *Ośrodkowi patronować ma zawodowy mistrz świata Dariusz Michalczewski*. [Skł.]

(56) *Przedostatni figuruje* Artur Balazs, poseł i minister rolnictwa, *a zamyka ją* Stefan Oleszczuk z UPR-u, starosta kamieński. [Skł.]

Konstrukcja apozycyjna powstaje przez zestawienie dwóch fraz nominalnych o podobnej charakterystyce. Saloni i Świdziński (2001) podają, że między członami musi zachodzić uzgodnienie przypadka. Konstrukcja ma charakter podrzędny, jej nadrzędnikiem jest pierwszy składnik, co widać, gdy składniki różnią się wartością rodzaju (Saloni i Świdziński 2001, s. 184):

(57) *Wygrała* dziewczyna chirurg.
(58) **Wygrał* dziewczyna chirurg.
(59) Dziewczyna chirurg *zgrabnie otworzyła mu jamę brzuszną.* [Saloni i Świdziński 2001]
(60) * Dziewczyna chirurg *zgrabnie otworzył mu jamę brzuszną.*

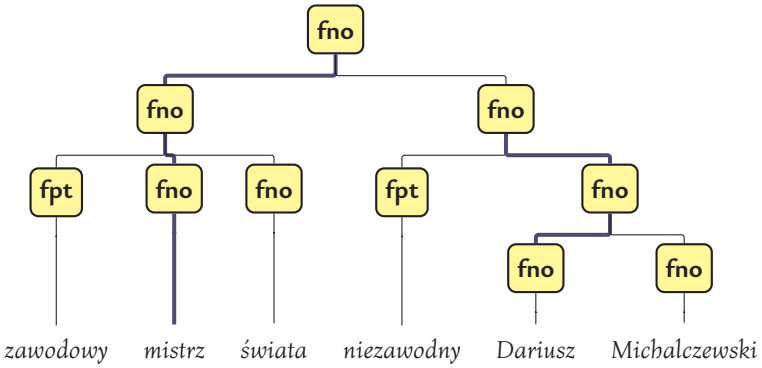
Uzgodnienia z czasownikiem w tych przykładach wskazują, że reprezentantem konstrukcji apozycyjnej jest tylko pierwszy składnik. Składnik podrzędny może być wydzielony przecinkami ortograficznymi, jak w przykładzie (56).

Jako apozycję postanowiono także traktować konstrukcję złożoną z imienia i nazwiska. W konsekwencji nadrzędnikiem tej frazy jest imię.

Istnieją podobne konstrukcje, w których drugi składnik jest w mianowniku (Saloni 2004). W naturalnie brzmiących frazach tego typu składnik ten jest najczęściej nazwą (przykłady (61)–(63)), ale może on być nazwą gatunkową pisaną małą literą (przykłady (64)– (66)). Konstrukcja ta jest również używana do wyrażenia wyniku pomiaru, np. (67). Niestety uwarunkowania te trudno uchwycić w gramatyce powierzchniowej.

(61) Agencja detektywistyczna „Nemezis" *dobrze prosperuje.* [Skł.]
(62) *Wystartował m.in. znany kolarz zawodowej* grupy Mróz, *Piotr Wadecki, [...]* [NKJP1M]
(63) *Historyczny Park Stanowy Fort Churchill, znajduje się na terenie* hrabstwa Lyon *w* stanie Nevada. [NKJP1M]
(64) *Poznaj* język asembler *dla* procesorów Intel *i kompatybilnych* [NKJP300]
(65) *Kierowca jadący* samochodem toyota corolla *najprawdopodobniej wpadł w poślizg i jego auto dachowało.* [NKJP300]
(66) *Właściciel Ośrodka wolałby, żeby zamiast* nazwy wczasy odchudzające *używać spa.* [NKJP1M]
(67) *W poniedziałek miał* gorączkę 39 stopni *i skarżył się na bóle brzucha.* [Skł.]

Konstrukcje omawiane w tym punkcie nie były uwzględnione w GFJP. Ze względu na ich charakter, polegający na zestawieniu dwóch kompletnych fraz nominalnych, w prezentowanej tutaj gramatyce przyjęto ich reprezentację w postaci osobnego węzła w drzewie. Tak więc podrzędnik apozycyjny nie występuje jako współskładnik innych podrzędników rzeczownika. W przykładzie



Rysunek 2.14. Przykład zagnieżdżonej konstrukcji apozycyjnej

na rysunku 2.14 są dwie konstrukcje apozycyjne: imię *Dariusz* łączy się na zasadzie apozycji z nazwiskiem *Michalczewski* oraz fraza *zawodowy mistrz świata* z frazą *niezawodny Dariusz Michalczewski*. Podrzędniki składnika *mistrz* (czyli *zawodowy* i *świata*) łączą się z nim w kolejny składnik, który staje się nadrzędnikiem konstrukcji apozycyjnej. Podobnie składnik *niezawodny* zostaje przypisany jako podrzędnik do frazy apozycyjnej *Dariusz Michalczewski* jako całości, a nie do jej nadrzędnika *Dariusz*.

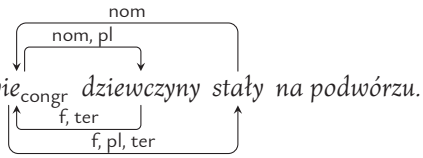
2.8.3. Frazy liczebnikowe i nominalne ze składnikiem liczebnikowym

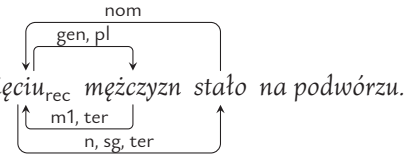
Frazy zawierające formy liczebnikowe nie zostały uwzględnione w GFJP. Ten brak naprawiono poprzez wprowadzenie do gramatyki jednostki **flicz** reprezentującej frazę liczebnikową (por. Świdziński i Woliński 2009). W najprostszej postaci jest ona realizowana przez pojedyncze formy liczebnika (*dwóm, pięciorgiem, cztery*). Bardziej skomplikowane realizacje obejmują konstrukcje wyrażające liczność, np. *dwudziestu sześciu, dwustu dwudziestoma sześcioma, dwadzieścia cztery tysiące sto dwadzieścia jeden*, złożone z form liczebników i form rzeczowników TYSIĄC, MILION itd. Jednostka **flicz** może stanowić składnik frazy nominalnej **fno**, przy czym przykłady (68)–(70) świadczą o tym, że w konstrukcji złożonej z formy liczebnika i formy rzeczownika za reprezentanta (a więc nadrzędnik) musi zostać uznana forma liczebnika:

(68) *Pięciu mężczyzn stało na podwórzu.*
(69) *Pięciu stało na podwórzu.*
(70) * *Mężczyzn stało na podwórzu.*
(71) *Dwie dziewczyny stały na podwórzu.*
(72) *Dwie stały na podwórzu.*
(73) *Dziewczyny stały na podwórzu.*

Zależności akomodacyjne liczebników z formami rzeczownikowymi i czasownikowymi w języku polskim są dość skomplikowane. Ich dwa zasadnicze układy ilustrują zdania (68) i (71). To zróżnicowanie jest związane z kategorią akomodacyjności wprowadzoną w punkcie 1.7.6. Forma *dwie* uzgadnia się co do przypadka z formą *dziewczyny*, a zatem jej wartość akomodacyjności to congr. Forma *pięciu* należy do rządzących (rec), co oznacza, że ma ona względem rzeczownika dopełniaczowy rząd przypadka.

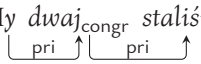
Za Salonim i Świdzińskim (2001, s. 207) przyjęto interpretację, w której układ zależności akomodacyjnych jest taki sam dla wszystkich form liczebników (por. także Derwojedowa 2011). Interpretację tę dla przykładów (68) i (71) przedstawiają następujące schematy:

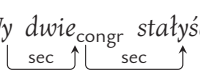
(74) 

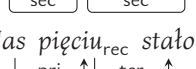
(75) 

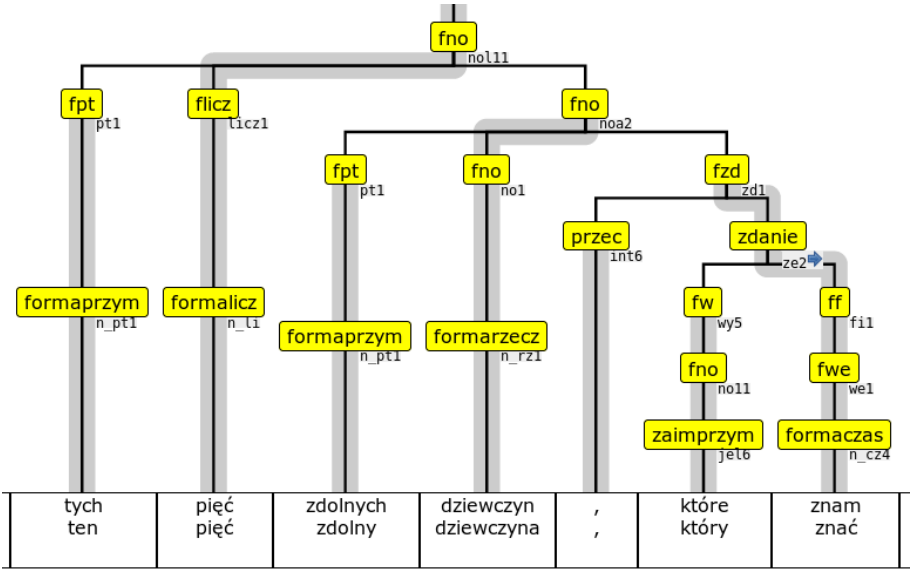
Forma czasownika narzuca formie liczebnikowej wartość przypadka (tutaj: nom). Forma liczebnikowa narzuca formie rzeczownikowej swoją wartość liczby oraz wartość przypadka zależną od cechy akomodacyjności (p. 1.7.6) danej formy liczebnikowej. Forma uzgadniająca congr narzuca wartość przypadka równą swojej, a forma rządząca rec – wartość dopełniacza. Podrzędna forma rzeczownikowa narzuca formie liczebnikowej wartość rodzaju i osoby. W wypadku pozycji podmiotu forma liczebnikowa narzuca formie czasownikowej wartość liczby, osoby i rodzaju, znowu zależnie od akomodacyjności: forma congr – swoje wartości, forma rec – ustaloną wartość rodzaju nijakiego, liczby pojedynczej i osoby trzeciej.

Formy liczebnikowe mają nieustaloną wartość osoby, mogą bowiem łączyć się z frazami nominalnymi i formami czasownikowymi o różnych wartościach osoby. Opisane „przekazanie" wartości osoby od podrzędnika liczebnika do jego nadrzędnika czasownikowego widać, gdy podrzędnikiem jest zaimek osobowy:

(76) 

(77) 

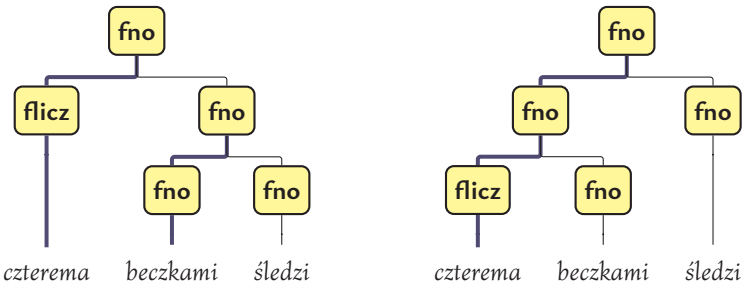
(78) 



Rysunek 2.15. Przykład rozbudowanej frazy nominalnej z liczebnikiem

Strukturyzacja fraz nominalnych zawierających jako składnik frazę liczebnikową jest uwarunkowana tym, że składnik ten musi być nadrzędnikiem frazy, a zatem trzeba w osobnych węzłach drzewa uwzględnić podrzędniki składnika liczebnikowego i składnika nominalnego (por. Świdziński i Woliński 2009; Świdziński 2005). Rysunek 2.15 przedstawia przykład wynikającej z tego strukturyzacji.

Jeżeli we frazie z liczebnikiem pojawia się jednostka miary lub pojemnik (szklanki, butelki, puszki, skrzynki, beczki itp.), to powstaje wątpliwość interpretacyjna zilustrowana na rysunku 2.16. Źródłem problemu jest to, że liczebnik jest składniowym nadrzędnikiem, co uniemożliwia zinterpretowanie tych trzech elementów „płasko" jako jednego wierzchołka. Wybór między



Rysunek 2.16. Dwie możliwe strukturyzacje fraz nominalnych z liczebnikiem i pojemnikiem

tymi strukturami można uznać za arbitralny i nie przypisywać im różnicy semantycznej. W tej pracy przyjęto wariant drugi, co pozwala nadać podobną strukturę następującym konstrukcjom:

(79) *dwa lata więzienia*
(80) *siedmiu synów kowala*
(81) *trzech synów i dwie córki kowala*
(82) *trzy tysiące złotych*
(83) *dwa tuziny jajek*
(84) *dwie kopy i pięć jajek*
(85) *dwadzieścia i pół jabłka*

2.8.4. Modyfikatory partykułowe

GFJP nie uwzględnia w przedstawionej analizie partykuł, nie opisuje więc na przykład takich zdań polskich:

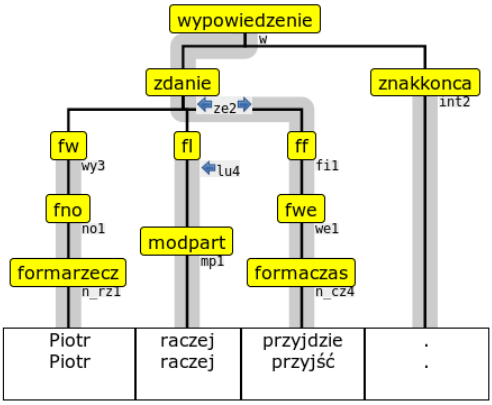
(86) *Piotr raczej przyjdzie...*
(87) *Byle uczniak to wie.*
(88) *Czekała prawie dwie godziny.*
(89) *Mieszka w zbyt luksusowym domu, który kosztował zbyt dużo.*
(90) *Zatrzymał się na miejscu niemal pod mostem.*

W tej pracy przyjęto, że partykuły wchodzą wyłącznie w konstrukcje podrzędne i tylko jako podrzędnik. Oznacza to, że wokół partykuł nie rozbudowują się frazy, partykuły stanowią składnik tylko same w sobie. Partykuły są mocno zróżnicowane w zakresie jednostek, dla których mogą stanowić podrzędnik. Dlatego wprowadzono następujące klasy, wskazujące jakich jednostek dana partykuła może być składnikiem:

– pwe – frazy czasownikowej i zdania (np. CZY, CZYŻBY, NARESZCIE, PRZECIEŻ, RACZEJ, por. (86), ZNOWU),
– pno – frazy nominalnej (np. BYLE, por. (87), NAWET, TYLKO),
– plicz – frazy liczebnikowej (np. AŻ, NIESPEŁNA, PRAWIE, por. (88)),
– ppt – frazy przymiotnikowej i przysłówkowej (np. CAŁKIEM, NIEZBYT, ZA, ZBYT, por. (89)),
– ppm – frazy przyimkowo-nominalnej (np. AKURAT, NIEMAL, por. (90), RZEKOMO, TUŻ).

Wiele partykuł należy do więcej niż jednej klasy.

W gramatyce wprowadzono jednostkę **modpart**, która może być realizowana przez pojedynczą formę partykułową. Jedynym modyfikatorem centrum partykułowego może być partykuła NIE, a i ona jest możliwa jedynie przy niektórych jednostkach, mianowicie BYLE, CAŁKIEM, DOSYĆ, DOŚĆ, NAZBYT, TYLKO, WPROST, ZA. Niektóre partykuły (co najmniej CZY, CZYŻ, CZYŻBY, AZALI, AZALIŻ, LI) wprowadzają do wypowiedzenia pytajność. Mogą one tworzyć



Rysunek 2.17. Przykład modyfikatora partykułowego na poziomie zdania

konstrukcję pytajną lub pytajnozależną. W regułach zdefiniowano również realizację jednostki **modpart** przez dwuczłonowe konstrukcje *co najwyżej* i *co najmniej*:

(91) *Znają co najwyżej encyklopedię.*

(92) *Co najmniej przeczyta, a może skomentuje.*

Jednostka **modpart** może stanowić realizację frazy luźnej (o ile partykuła należy do klasy pwe, por. rys. 2.17), jak również jest dopuszczalnym składnikiem fraz składnikowych o centrach poszczególnych typów: partykuła klasy pno – składnikiem frazy nominalnej **fno** (por. rys. 2.18), klasy plicz – frazy liczebnikowej **flicz** itd. Partykuła klasy pwe może być także składnikiem frazy werbalnej **fwe** z centrum niefinitywnym.

2.8.5. Modyfikatory taki, jaki(ś)

Nietypowym zachowaniem składniowym cechują się niektóre wystąpienia form przymiotników taki, jaki i jakiś. W wyniku analizy przeprowadzonej na potrzeby znakowania korpusu Składnica po dyskusji w zespole przyjęto, że fragmenty zaznaczone w następujących przykładach stanowią składniki:

(93) *– O, jaki śliczny chłopczyk! – wykrzyknął i spalił się do reszty.* [NKJP1M]

(94) *– Ojej, jakie wspaniałe czereśnie.* [NKJP1M]

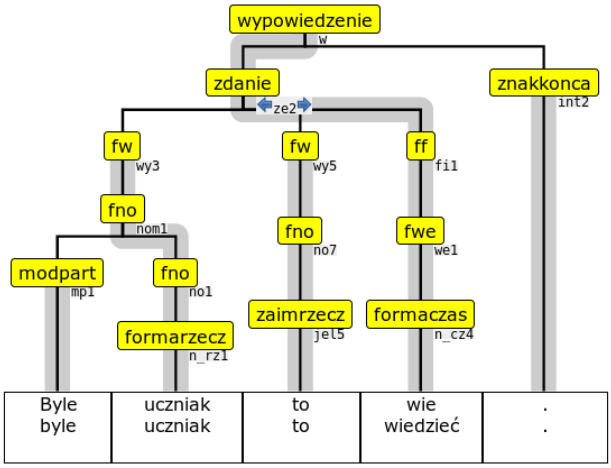
(95) *Był jakiś dziwny.* [Skł.]

(96) *Być może, w takim przypadku powstałaby jakaś porządna, prawdziwa lewica.* [Skł.]

(97) *– Masz takie męskie dłonie, January.* [Skł.]

(98) *Cmentarz jest taki pusty!* [Skł.]

(99) *Dlaczego Ania jest taka niesamodzielna?* [Skł.]



Rysunek 2.18. Przykład modyfikatora partykułowego we frazie nominalnej

(100) – *Staramy się opanować sytuację, ale nie jest to takie proste*. [Skł.]

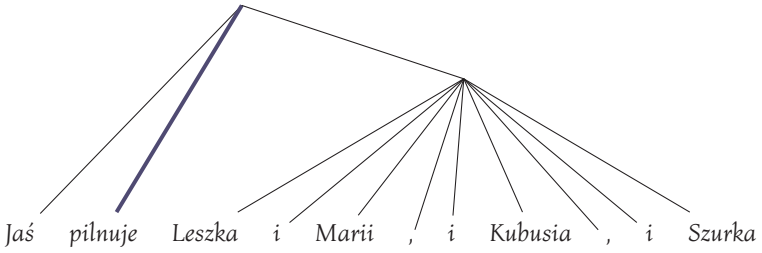
(101) *Mieszkał z ojcem i nie chciał, żeby ktoś wiedział, że płaci takie duże rachunki.* [Skł.]

W zdaniach tych formy leksemów TAKI, JAKI i JAKIŚ wykazują zróżnicowania rodzajowo-liczbowo-przypadkowe, wypada więc uznać je za formy przymiotników. We wszystkich wypadkach można by interpretować te formy jako podrzędniki rzeczowników, jednak z różnym stopniem wyrazistości test redukcji wskazuje, że zaznaczone fragmenty tworzą całość. Na przykład *takie męskie dłonie* to nie są dłonie, które są *takie* i *męskie*, one są *takie męskie*. Interesującą wskazówką jest też być może przykład (98), w którym postulowany podrzędnik *taki pusty* został w całości oderwany od rzeczownika *cmentarz*.

Przyjęto więc, że formy wymienionych przymiotników mają nietypową możliwość występowania jako podrzędnik innych przymiotników. W gramatyce modyfikator taki jest realizowany przez jednostkę nieterminalną o nazwie **modjaki**. W konstrukcji tej następuje uzgodnienie przypadka, rodzaju i liczby.

2.9. Schematy strukturyzacyjne konstrukcji współrzędnych

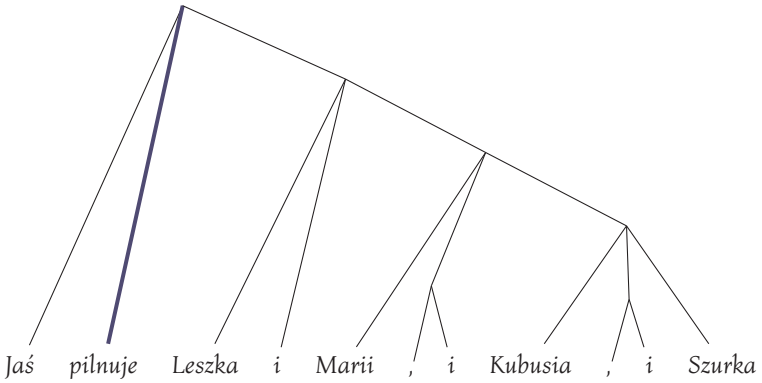
Konstrukcje współrzędne charakteryzują się tym, że wśród ich składników bezpośrednich daje się wskazać więcej niż jeden składnik o własnościach bardzo zbliżonych do własności całej konstrukcji. Można więc wskazać wielu kan-



Rysunek 2.19. Ilustracja (111) z rozdziału III pracy Saloni i Świdziński (2001)

dydatów na reprezentanta konstrukcji. Oprócz takich składników w skład konstrukcji wchodzą spójniki współrzędne i ewentualne znaki interpunkcyjne.

Jak odnotowują Saloni i Świdziński (2001, §III.5, s. 58), w analizie składnikowej konstrukcje współrzędne nie dają się adekwatnie reprezentować za pomocą węzłów binarnych. Z konstrukcji *brata albo Marii* nie da się usunąć pojedynczego wyrazu, zachowując poprawność wypowiedzenia. W szczególności *albo*, *brata albo*, ani *albo Marii* nie są składnikami. Tak więc konstrukcja ma trzy składniki bezpośrednie i nie daje się zbinaryzować. Dlatego autorzy uznają konieczność wprowadzenia wierzchołków o trzech potomkach. Jednak dalej, omawiając konstrukcje szeregowe (§III.6.4, s. 70), autorzy uznają dziewięć gałęzi wychodzących z tego samego wierzchołka (por. rys. 2.19) za „niepożądane konsekwencje" przyjętej zasady opisu, choć „drzewo to oddaje intuicje semantyczne". Zamiast tego autorzy proponują interpretację (por. rys. 2.20), w której z jednego wierzchołka odchodzą co najwyżej trzy gałęzie. Osiągają to poprzez rekurencyjne zagnieżdżenie jednostek o mniejszym stopniu rozgałęzienia oraz techniczny zabieg zebrania przecinka i następującego spójnika w nową jednostkę. Jak piszą, taka struktura, choć daleka od intuicyjnej interpretacji, operuje prostszymi rozgałęzieniami, przez co „byłaby preferowana



Rysunek 2.20. Ilustracja (112) z rozdziału III pracy Saloni i Świdziński (2001)

w technicznych rozwiązaniach informatycznych". Jest to zapewne echo problemu ustalonej liczby wierzchołków potomnych w regułach gramatyk bezkontekstowych (por. p. 2.7).

Reguły GFJP ze względu na sztywną liczbę elementów prawej strony reguły opisują konstrukcje szeregowe w taki właśnie rekurencyjny sposób, z użyciem pomocniczej jednostki nazywanej konstrukcją jednorodną. Przykład zdania szeregowego (reprezentującego inny układ niż na rys. 2.20) zanalizowanego według reguł GFJP przedstawia rysunek 2.21. Zdanie szeregowe przepisuje się na wystąpienie zdania jednorodnego, które z kolei składa się ze spójnika, zdania elementarnego, przecinka i kolejnego zdania jednorodnego, realizującego pozostałą część szeregu. Przy interpretacji tej szereg złożony z czterech elementów jest opisany trzema zagnieżdżonymi wystąpieniami jednostki zdanie jednorodne.

W prezentowanej gramatyce szeregi są realizowane bezpośrednio, poprzez tworzenie wierzchołków zbierających wszystkie składniki szeregu. Strukturę odpowiadającą rysunkowi 2.21 przedstawia rysunek 2.22. Jako centrum struktury przyjęto spójnik. W wypadku konstrukcji szeregowych jest to jeden ze spójników – początkowy w konstrukcji zaczynającej się spójnikiem, a w pozostałych konstrukcjach – szeregowy końcowy (jako bardziej charakterystyczny).

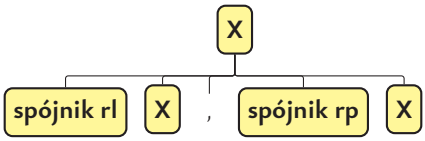
W GFJP schematy strukturyzacyjne współrzędne są stosowane do łączenia zdań i fraz zdaniowych. W niniejszej pracy schematy te są stosowane również w odniesieniu do fraz nominalnych, przymiotnikowych, werbalnych, przysłówkowych i przyimkowych. Przykłady realizacji z udziałem fraz różnych typów przedstawiono przy poszczególnych schematach strukturyzacyjnych.

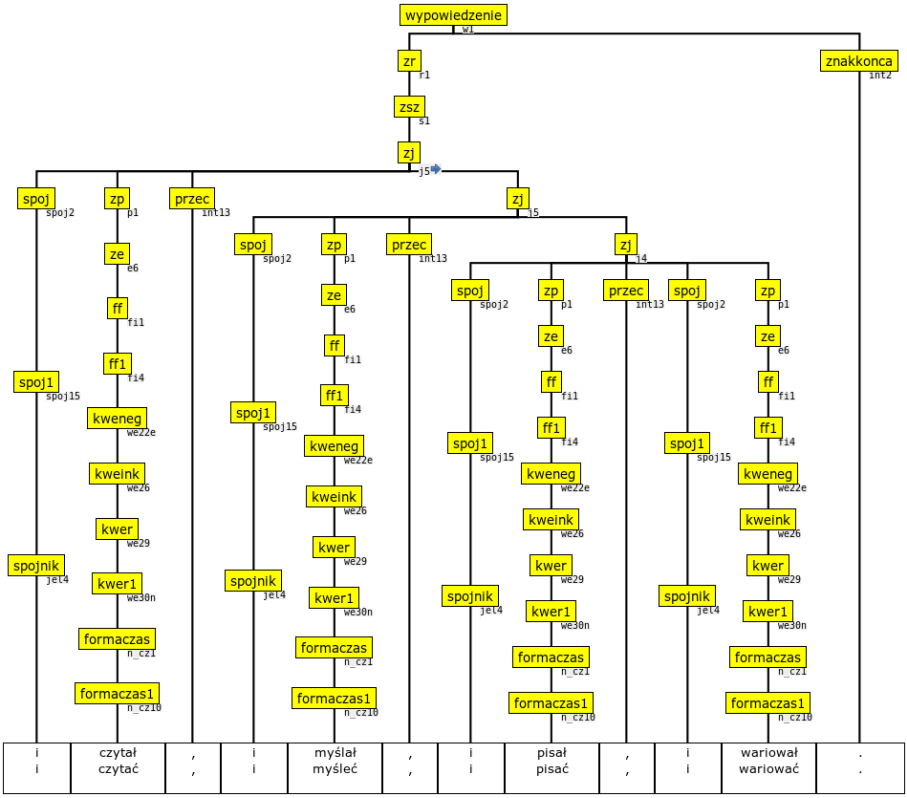
2.9.1. Konstrukcje równorzędne

Konstrukcje nazwane przez Świdzińskiego równorzędnymi są dwumiejscowe – obejmują dwie frazy tego samego typu (oznaczanego na poniższych schematach **X**) i spójnik. Zgodnie z ogólną zasadą wynikowa konstrukcja jest też typu **X**. Świdziński wyróżnia trzy układy konstrukcji równorzędnych (1992, §5.2.1, s. 107–108).

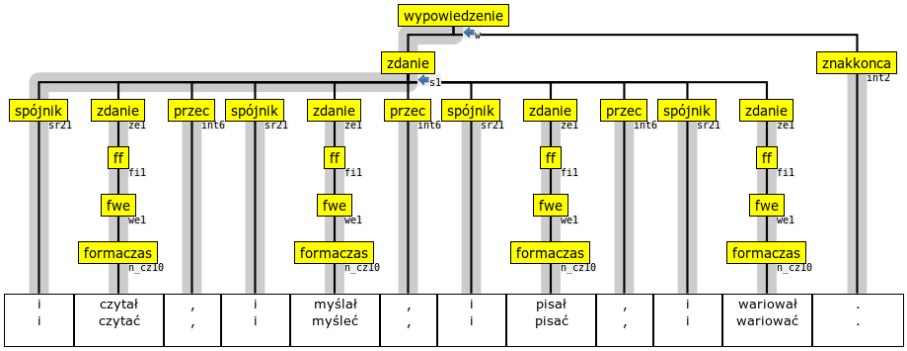
Konstrukcja równorzędna układ I

W pierwszym układzie elementem spajającym jest spójnik dwuczęściowy, np. NIE TYLKO – ALE TAKŻE. Jego części są nazywane przez Świdzińskiego spójnikiem równorzędnym lewym i spójnikiem równorzędnym prawym.





Rysunek 2.21. Zdanie szeregowe *I czytał, i myślał, i pisał, i wariował* zinterpretowane według reguł GFJP



Rysunek 2.22. To samo zdanie w interpretacji analizatora Świgra 2

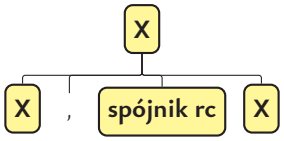
(102) zdanie *Nie tylko pobierzemy się wkrótce, ale także rodzice kupili nam mieszkanie*. [GFJP, przyk. (3), s. 107]

(103) *W ciągu ostatniego minionego tygodnia* fno *zarówno dolar, jak i marka straciły wobec polskiej waluty.* [Skł.]

(104) *Widywano go z nimi* fpm *zarówno przed szałasem w Małej Łące, jak i przed swoją chałupą we wsi*. [Skł.]

Konstrukcja równorzędna układ II

W tym układzie występuje spójnik równorzędny centralny, np. A. Spójnik taki obowiązkowo poprzedzony jest przecinkiem. Do tej kategorii należą spójniki nietworzące szeregów (por. dalej).



(105) zdanie *Pobierzemy się wkrótce, a rodzice kupili nam mieszkanie*. [GFJP, przyk. (4), s. 108]

(106) zdanie *On nie uczestniczy w tych grach i rywalizacjach, ale ma instynkt państwowy*. [Skł.]

(107) *Główną jednak atrakcją pozostał* fno *seks mimowolny, czyli koncepcje babci dotyczące naszych ukrytych stosunków*. [Skł.]

(108) *To by wyjaśniało* fno *obecność tlenu, ale nie sprawę „Kondora"*. [Skł.]

(109) *A Religa ma sprawny sztab* fpm *w Warszawie, ale nie poza nią*. [Skł.]

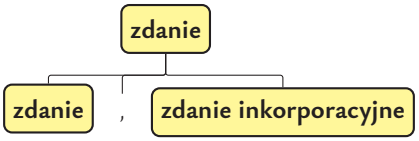
(110) *– Pytam* fzd *, co sądzisz, a nie czy byłeś*. [Skł.]

(111) *Jego anegdoty traktowały o najrozmaitszych księżach dziwakach* fzd *, których w okolicy nie brakło, a których przygody tchnęły często humorem z epoki Seweryna Soplicy czy nawet Paska*. [Skł.]

(112) *Musi umieć* fwe *scalać ludzi, a nie ich dzielić*. [Skł.]

Konstrukcja równorzędna układ III

Układ trzeci obejmuje zdania łączone spójnikami inkorporacyjnymi. Są to jednostki nietypowe wśród spójników, ponieważ nie występują pomiędzy łączonymi zdaniami, ale wewnątrz drugiego z łączonych członów.



(113) *Pobierzemy się wkrótce*, *rodzice zaś kupili nam mieszkanie*.
[GFJP, przyk. (5), s. 108]

(114) *Dziś byłoby to raczej niemożliwe*, *są natomiast inne sposoby, o czym świadczy los biblioteki PAN*. [Skł.]

(115) *W czwartek do egzaminu ustnego z historii przystąpiło 94 maturzystów*, *w piątek natomiast 59 miłośników biologii popisywało się wiedzą z tej dziedziny*. [Skł.]

Spójniki inkorporacyjne są z natury zdaniowe, nie można nimi połączyć na przykład fraz nominalnych (niemożliwy jest ciąg **Piotr, Jan zaś* interpretowany jako fraza nominalna).

Układy równorzędne I i II są stosowane we wszystkich typach fraz wymienionych w poprzednim punkcie. Układ III – jedynie do łączenia zdań.

2.9.2. Konstrukcje szeregowe

Konstrukcje szeregowe charakteryzują się możliwością połączenia wielu równoważnych elementów. Elementem spajającym szereg są spójniki szeregowe (oznaczone w poniższych schematach **spójnik sz**), a ostatni element szeregu musi być poprzedzony szczególnym spójnikiem szeregowym końcowym (**spójnik szk**), przy czym spójniki szeregowe mogą pełnić funkcję szeregowych końcowych (wtedy wszystkie spójniki w szeregu są takie same), ale nie na odwrót. Na przykład według klasyfikacji Świdzińskiego do spójnika szeregowego typu i, który może być realizowany przez formę *i*, pasują następujące spójniki szeregowe końcowe: *i*, *i nawet*, *i także*, *oraz* (Świdziński 1992, s. 95). Na przykład:

(116) *Piotr i Jan, i Maria, i Radek*
(117) *Piotr i Jan, i Maria, oraz Radek*
(118) *Piotr i Jan, i Maria, albo i Radek*
(119) *Piotr lub Jan, lub Maria, lub też Radek*
(120) *Piotr ani Jan, ani Maria, ani nawet Radek*

Repertuar spójników szeregowych został rozszerzony w stosunku do GFJP o spójnik CZY, który ma tę nietypową cechę, że może łączyć frazy, ale nie może łączyć zdań:

(121) *Chodzi o to, żeby demonstracje odbywały się w spokoju, żeby w minimalnym stopniu zakłócały życie innych ludzi i żeby nie było polską normą* **fno** *bicie policjantów, okupowanie budynków, wyrywanie znaków drogowych czy niszczenie samochodów*. [Skł.]

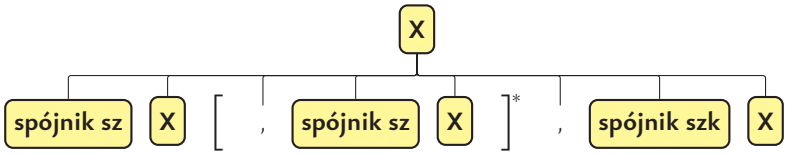
(122) – **fpm** *O dwunastej czy o pierwszej*. [Skł.]

Jako szeregowy końcowy występuje też wariant CZY TEŻ.

Świdziński (1992, §5.3.1 s. 118) wyróżnia trzy układy konstrukcji współrzędnych szeregowych.

Konstrukcja szeregowa układ I

W tym układzie każde wystąpienie jednostki **X** w szeregu jest poprzedzone spójnikiem tego samego typu. Niepierwsze wystąpienia spójnika poprzedza przecinek. W poniższym schemacie fragment ujęty w oznaczenie $[\,]^*$ może być powtórzony dowolną liczbę razy (w szczególności zero).



(123) zdanie *I błyska, i grzmi, i pada.* [GFJP, przyk. (41a), s. 118]

(124) zdanie *Ani my ciebie nie pobudzamy, ani ty się nie dzielisz.* [Skł.]

(125) *Bracia Kaczyńscy chcą w tej chwili* fno *albo koalicji na swoich warunkach, albo przyspieszonych wyborów.* [Skł.]

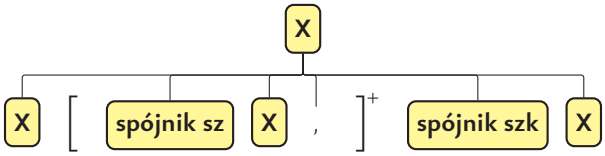
(126) fno *I w jednym, i w drugim wypadku zakres i tematyka informacji cierpią na tym ogromnie.* [Skł.]

(127) *– Zresztą ci sami należą też* fpm *i do czeskiego, i do bułgarskiego.* [Skł.]

W tym układzie niemożliwe jest użycie spójnika LUB.

Konstrukcja szeregowa układ II

Tutaj spójniki występują jedynie między łączonymi frazami, ostatni spójnik może być różny od poprzednich (oczywiście przy zachowaniu tego samego typu spójnika). W poniższym schemacie oznaczenie $[\,]^+$ sygnalizuje, że ten fragment schematu może wystąpić jeden raz lub więcej razy.



(128) zdanie *Błyska i grzmi, i pada.* [GFJP, przyk. (42), s. 118]

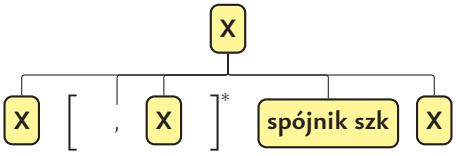
(129) zdanie *Stanął na grobli i zadumał się głębokim zadumaniem, i zafrasował się wielkim zafrasowaniem.* [Skł.]

(130) *Nie było pastucha i jego bata, nie było* fno *psów ani łańcucha w dusznej oborze, ani zdradliwych rogów zawistnych towarzyszek.* [Skł.]

(131) *Popluł w żylaste, wykrzywione dłonie, zżarte od* fno *cementu i wapna, i gipsu.* [Skł.]

Wymaganie co najmniej jednokrotnego wystąpienia fragmentu ujętego w nawiasy pozwala uniknąć zbieżności z minimalną realizacją układu III.

W ostatnim układzie szereg jest spajany za pomocą przecinków z wyjątkiem ostatniego wystąpienia jednostki **X**, które jest poprzedzone spójnikiem szeregowym końcowym.



Najczęstsza realizacja tego schematu to połączenie dwóch fraz spójnikiem szeregowym końcowym (a więc bez wystąpień powtarzalnej części schematu). Ta realizacja dominuje wśród konstrukcji współrzędnych, a w niej z kolei przeważa użycie spójnika I. Spójnik szeregowy końcowy w tym układzie może być przecinkiem (wtedy cały szereg jest czysto przecinkowy).

Poniższe przykłady ilustrują bogactwo użyć tego schematu w zdaniach i frazach składnikowych różnych typów.

(132) zdanie *Błyska, grzmi* , *pada* [GFJP, przyk. (43a), s. 118]

(133) zdanie *Błyska, grzmi* i *pada.* [GFJP, przyk. (44), s. 118]

(134) *Mam swoje ulubione* fno *tytuły* , *fragmenty* ... [Skł.]

(135) *Miejsc tych tragedii nie wizytują* fno *premierzy, ministrowie* czy *samorządowi notable*. [Skł.]

(136) *– Czy mam* fno *garb na plecach* albo *długą szyję*? [Skł.]

(137) *Przy ascendencie w* fno *Skorpionie* lub *Koziorożcu*, *nieoczekiwane wydarzenia mogą skomplikować życie.* [Skł.]

(138) *Zanieczyszczenia* fpt *cytoplazmatyczne* i *błonowe* *są również groźne, stanowiąc prawdopodobne źródło zarówno aktywności proteolitycznej, jak i białek cytoplazmatycznych w puli niehistonowej.* [Skł.]

(139) *Polska jest* fpt *etnicznie jednorodna, heteroseksualna, katolicka* i *wolnorynkowa*. [Skł.]

(140) *Potem często pokazywano ich* fpt(pact) *przytulających się* lub *całujących*. [Skł.]

(141) *Nie są* fpt(ppas) *dokarmiane* ani *pojone*. [Skł.]

(142) fps *Prędzej* czy *później* *Unię szlag trafi.* [Skł.]

(143) *U kotów lekiem z wyboru są pochodne benzodwuazepiny podawane* fps *dożylnie* lub *doodbytniczo*. [Skł.]

(144) *Dlatego też cenny towar trafi prawdopodobnie* fpm *do Niemiec* lub *do Danii*. [Skł.]

(145) *Myślałem* fzd *, że jest bierny* , *że wypłynął na fali*. [Skł.]

(146) *Nie wie nawet* fzd *, kiedy jest sobota* i *kiedy wypadają żydowskie święta*. [Skł.]

(147) *Pisze* fzd *, iż jego rządy trwały lat 36* i *że jego następcą był Włodzimierz*.

(148) *Zabroniono im* fwe *mieszkać w niektórych dzielnicach, chodzić niektórymi ulicami* oraz *podróżować koleją*. [Skł.]

(149) *Ludzie zaś mają oprócz gestów swoje emocje i tym emocjom dają wyraz,* fwe *mówiąc czułe słówka* albo *klnąc*. [Skł.]

(150) fwe *Drżąc na ciele* i *wzbudzając obydę w duszy, rozpocząłem wędrówkę myszką, by uruchomić bank informacji.* [Skł.]

(151) *Cała rodzina* fwe *była* i *pozostała łemkowska.* [Skł.]

2.9.3. Nieredukowalne frazy nominalne

Wśród fraz nominalnych zgodnych formalnie ze schematami konstrukcji współrzędnych występują konstrukcje niezgodne z definicją współrzędności. W konstrukcjach tych reprezentantem jest albo dokładnie jedna z fraz składnikowych, albo też reprezentanta w ogóle nie da się wskazać.

Zjawisko to najłatwiejsze jest do zaobserwowania dla fraz podmiotowych:

(152) *Przyszli i Jan, i Maria, i Piotr*.

Fraza *i Jan, i Maria, i Piotr* występuje jako podmiot przy formie czasownikowej cechującej się wartością mnogą liczby i męskoosobową rodzaju. Jednak każdy ze składników jest w liczbie pojedynczej i nie wszystkie są męskoosobowe. Fraza jest więc nieredukowalna, jej własności (męskoosobowa, mnoga) nie pokrywają się z własnościami żadnego ze składników. Że fraza jako całość jest męskoosobowa, można wywnioskować z tego, że w tym zdaniu wykluczona jest forma czasownika *przyszły*.

Tworzenie tego rodzaju konstrukcji jest możliwe jedynie z użyciem niektórych spójników. Spójniki można podzielić na koniunkcyjne i alternatywne. Tylko te pierwsze wydają się tworzyć frazy nieredukowalne:

(153) *Przyszli Jan i Maria.*
(154) *Przyszedł Jan i Maria.*
(155) **Przyszli Jan lub Maria.*
(156) *Przyszedł Jan lub Maria.*

Frazy ze spójnikiem koniunkcyjnym mogą być redukowalne lub nie. W niniejszym opisie przyjęto, że koniunkcyjne są spójniki o oznaczeniach I, ANI, NIE TYLKO, TAK, ZARÓWNO, ALE I, PRZEC. Frazy z innymi spójnikami są zawsze redukowalne.

We frazach redukowalnych, jeżeli składniki różnią się wartością jednej z cech gramatycznych uzgadnianych z otoczeniem, to występuje tzw. uzgodnienie do sąsiada. Jest to mianowicie uwarunkowanie porządku linearnego frazy polegające na tym, że nadrzędnik, z którym fraza się uzgadnia, musi wystąpić linearnie po stronie uzgadniającego się składnika frazy (być jego sąsiadem). Przy tym takie uzgodnienie jest możliwe w wypadku fraz ze spójnikiem

koniunkcyjnym, jedynie jeżeli fraza finitywna poprzedza podmiot. Ograniczenie szyku nie występuje dla spójników alternatywnych (Szpakowicz i Świdziński 1990). Następujące przykłady ilustrują takie uzgodnienie liczby i rodzaju podmiotu z orzeczeniem:

(157) *Przyszedł Jan i Marysia.*
(158) *Przyszła Marysia i Jan.*
(159) **Jan i Marysia przyszła.*
(160) *Przyszedł Jan lub Marysia.*
(161) *Jan lub Marysia przyszła.*

Przytoczone przykłady nie pasują do definicji konstrukcji współrzędnych. Skoro reprezentant jest dokładnie jeden, należałoby uznać je za podrzędne. Być może lepiej mówić o nich jako konstrukcjach z koordynacją, a więc takich, gdzie centrum stanowi spójnik (współrzędny/koordynacyjny).

W przykładach przedstawiono użycie fraz mianownikowych w pozycji podmiotu, bo w takim kontekście widać wyraźnie uzgodnienie. Nieredukowalne frazy nominalne można zaobserwować również w innych przypadkach, gdy składniki mają wspólny podrzędnik przymiotnikowy. Przykłady takie są jednak nietypowe i zwykle brzmią sztucznie.

(162) *Widzę naszych Jana i Marię.*

W prezentowanej gramatyce modyfikację przymiotnikiem dopuszczono tylko w wypadku fraz nieredukowalnych, uznając, że w wypadku fraz redukowalnych przymiotnik dołącza się wewnątrz frazy składnikowej.

W konstrukcjach nieredukowalnych powstaje pytanie o sposób wyznaczenia wartości kategorii gramatycznych, które są różne od składników. Wartość liczby frazy nieredukowalnej jest mnoga – wydaje się to mieć uwarunkowanie semantyczne, fraza denotuje grupę obiektów. Jako wartość osoby przyjmowana jest najmniejsza z wartości osoby składników frazy:

(163) *Przyszliśmy chłopiec i ja.*
(164) **Przyszli chłopiec i ja.*
(165) *Przyszliśmy Maria, ty i ja.*
(166) *Przyszłyście ty i Maria.*

Jeśli chodzi o wartość rodzaju, to ze względu na mnogą wartość liczby istotne jest jedynie rozróżnienie między formą czasownika właściwą dla rodzaju męskoosobowego i formą pasującą do wszystkich pozostałych rodzajów (nazywaną na zasadzie skrótu niemęskoosobową). W takiej konstrukcji rodzaj męskoosobowy dominuje inne, to znaczy obecność jednego składnika męskoosobowego wystarczy, żeby nadać taką samą wartość frazie:

(167) *Przyszli facet i kobiety.*
(168) *Przyszli Maria, Jan i dziecko.*
(169) **Przyszły facet i kobiety.*

Jeśli żaden ze składników nie jest męskoosobowy, dla całej frazy przyjmowana wartość rodzaju zapewniająca uzgodnienie z niemęskoosobową formą czasownika.

Nieredukowalność wydaje się możliwa również w wypadku koordynacji fraz przymiotnikowych:

(170) *Poprawki pierwsza i osiemnasta nie uzyskały akceptacji Izby.*

Mnoga wartość liczby formy *poprawki* wymaga uzgodnienia z mnogą wartością frazy przymiotnikowej. Taką więc wartość trzeba przypisać skoordynowanej frazie przymiotnikowej *pierwsza i osiemnasta*.

Innym zbliżonym zjawiskiem są konstrukcje z przyimkiem z zachowujące się analogicznie do nieredukowalnych fraz skoordynowanych:

(171) *Przyszli Jan z Marią.*

2.9.4. Połączenia współrzędne fraz składnikowych różnych typów

Ciekawym zjawiskiem składniowym w języku polskim jest możliwość utworzenia konstrukcji współrzędnej z fraz różnych typów. Na przykład w następującym zdaniu składnikami konstrukcji współrzędnej są fraza werbalna w bezokoliczniku i fraza nominalna:

(172) *Chce pić i sałatkę.*

Zjawisko to zostało w GFJP uwzględnione w bardzo ograniczonym zakresie. Świdziński przewidział mianowicie możliwość łączenia współrzędnego wymaganych fraz zdaniowych różnych typów (Świdziński 1992, §8.6.5, s. 301 oraz §5.15.5.3, s. 420), np.:

(173) *Wiadomo, kto będzie, a także że warto wydać przyjęcie.* [GFJP, przyk. (45), s. 420]

(174) *Mówiono nam, żeby przyjść oraz że będzie przyjęcie.* [GFJP, przyk. (47), s. 420]

(175) *Jak śpiewasz bądź kiedy tańczysz, wolimy.* [GFJP, przyk. (49), s. 420]

Możliwość takiej realizacji fraz wymaganych musi być przewidziana w słowniku walencyjnym. W GFJP została ona opisana w sposób bardzo nieporęczny, poprzez wprowadzenie mieszanych typów fraz zdaniowych oznaczonych mie1, mie2 i mie3 odpowiadających konkretnym kombinacjom fraz (np. mie1 to fraza typu że skoordynowana z frazą pytajnozależną, jak w przykładzie (173)). Wadą tego rozwiązania jest konieczność przewidzenia i nazwania każdej kombinacji możliwych do skoordynowania fraz. Symbole typu mie1 są przy tym zupełnie nieczytelne, a to nimi trzeba by operować w słowniku walencyjnym, aby oznaczyć, że przy danym czasowniku możliwa jest koordynacja danych typów fraz.

W słowniku walencyjnym Walenty zastosowano inne rozwiązanie. Mianowicie w opisie jednej pozycji składniowej może zostać użyte wiele specyfikacji typów frazy wymaganej (por. p. 3.1.3). Zapis taki oznacza, że pozycja może zostać zrealizowana przez frazę dowolnego z wymienionych typów, ale także przez frazę skoordynowaną złożoną z wymienionych składników. W ten sposób formalizm dopuszcza możliwość wyrażenia koordynacji dowolnych typów fraz, możliwe jest też określenie dla konkretnego czasownika zbioru typów fraz, które mogą zostać skoordynowane, co pozwala na dużą elastyczność i opis nietypowości.

W przedstawianym tu opisie są stosowane dane słownika Walenty, a zjawisko jest opisywane poprzez dopuszczenie koordynacji fraz wymaganych **fw**. Fraza skoordynowana otrzymuje jako oznaczenie typu frazy wymaganej listę typów fraz składowych. Przy wysycaniu pozycji składniowej przez daną frazę, jest sprawdzane, czy wszystkie typy są dopuszczane przez słownik walencyjny dla tej pozycji (zob. p. 4.5.4).

Dogłębną analizę koordynacji tego typu można znaleźć w pracy Patejuk (2015). Autorka analizuje także konstrukcje, w których koordynowane są frazy stojące na różnych pozycjach składniowych.

W analizatorze Świgra 2 ograniczono się do analizy koordynacji w obrębie pozycji w zakresie objętym opisem w Walentym, uznając, że koordynacja różnych pozycji jest zjawiskiem zbyt rzadkim, by warto było je uwzględniać na tym etapie rozwoju analizatora.

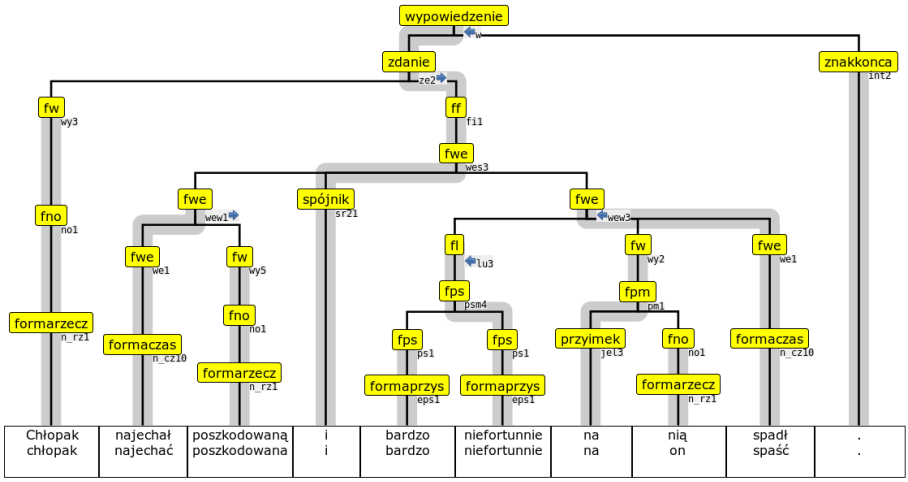
2.9.5. Konstrukcje uwspólniające podrzędniki

Dla pewnych zdań polskich naturalna wydaje się struktura, w której wszystkie lub tylko niektóre podrzędniki są wspólne dla wielu czasowników. Oto przykłady wypowiedzeń zawierających takie zdania:

(176) *Wszystko błyszczy, kręci się, tańczy.* [Skł.]
(177) *– Jeżeli to jest ten przypadek, o którym myślę, to ten chłopak najechał poszkodowaną i bardzo niefortunnie na nią spadł.* [Skł.]
(178) *Kilka miesięcy temu sytuację pokazali i opisali dziennikarze.* [Skł.]
(179) *– Ona zrobiła sobie operację plastyczną i przebiera się za mężczyznę – mówi chłopczyk, nie patrząc mi w oczy.* [Skł.]
(180) *Wczorajsze minęło i nie wróci, a dzisiaj mnie to ni ziębi, ni grzeje.* [Skł.]

Na przykład w wypowiedzeniu (176) podmiot *Wszystko* może być interpretowany jako wspólny dla wszystkich trzech form czasownikowych. W wypowiedzeniu (177) w zdaniu nadrzędnym podmiot *ten chłopak* jest wspólny dla form finitywnych *najechał* i *spadł*. Forma *najechał* wymaga dodatkowo frazy biernikowej *poszkodowaną*. Forma *spadł* dodatkowo wymaga frazy przyimkowej *na nią*, ma też luźny podrzędnik *bardzo niefortunnie*.

Takim zdaniom przypisywana jest struktura z koordynacją fraz werbalnych **fwe**, w których część wymagań (specyficznych dla danego czasownika)



Rysunek 2.23. Przykład zdania, w którym podmiot jest wspólny dla dwóch finitywnych form czasowników

jest realizowana wewnątrz składowej frazy werbalnej, a część – jako podrzędniki frazy skoordynowanej jako całości. Przykład takiej struktury odpowiadającej fragmentowi wypowiedzenia (177) przedstawia rysunek 2.23.

Przyjęto, że takie uwspólnienia są możliwe przy odpowiednim układzie linearnym, a mianowicie – skoordynowana fraza werbalna jest ciągła, a uwspólnione podrzędniki muszą linearnie znajdować się poza nią. Tak więc obecna wersja gramatyki nie opisuje następującego zdania z uwzględnieniem uwspólnienia podmiotu:

(181) *Błyszczy wszystko, kręci się, tańczy.*

W tego rodzaju skoordynowanych frazach werbalnych może pojawić się komplikacja związana z przypisanymi im wartościami kategorii gramatycznych. Jeżeli przyjąć, że w zdaniu (182) podrzędniki *tej jesieni* oraz *książki* odnoszą się do obu czasowników, to wypada uznać, że występuje w nim skoordynowana fraza werbalna *czytałem, a one pisały*:

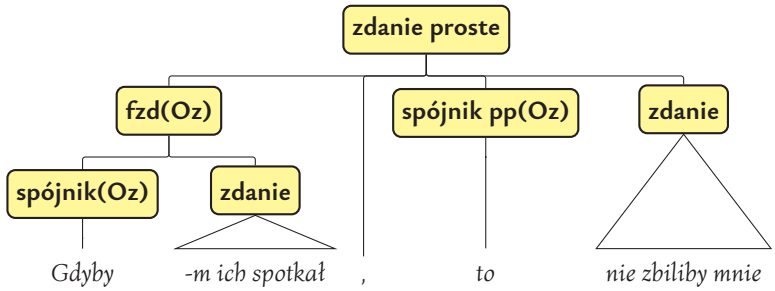
(182) *Tej jesieni czytałem, a one pisały książki.*

Powstaje pytanie, jakie wartości osoby, liczby i rodzaju powinny przysługiwać tej frazie. Składowe frazy różnią się co do wartości każdej z tych cech. W analizatorze Świgra 2 fraza ta otrzymuje specjalną wartość mie (mieszana) każdej z wymienionych kategorii. Wartość ta nie uzgadnia się z żadną wartością przewidzianą dla fraz nominalnych, co powoduje, że fraza taka nie może uzgodnić się z żadną frazą podmiotową. A więc fraza *czytałem, a one pisały* jako całość nie może mieć zrealizowanego podmiotu. W zdaniu (182) podmiot jednego z czasowników składowych został zrealizowany wewnątrz frazy współrzędnej.

2.10. Zdania proste

Oprócz już wymienionych schematów strukturyzacyjnych Świdziński wyróżnia jeszcze jeden rodzaj konstrukcji zdaniowych, które nazywa zdaniami prostymi. Są to konstrukcje złożone ze zdania elementarnego i frazy zdaniowej lub dwóch zdań elementarnych połączonych spójnikiem podrzędnym. Świdziński uznaje te konstrukcje za podrzędne (GFJP, §5.5.1, s. 139), redukowalne do jednej frazy finitywnej (GFJP, §4.3.5, s. 75). Ich cechą szczególną i powodem wyodrębnienia w GFJP w osobną jednostkę składniową jest to, że składowe zdania podrzędne mają ustaloną pozycję linearną, w odróżnieniu od sytuacji typowej, gdzie fraza zdaniowa zawierająca zdanie podrzędne może wystąpić w dowolnej pozycji względem innych składników zdania (elementarnego). Konstrukcje takie występują wyłącznie na poziomie zdań, nie ma konstrukcji analogicznych dla innych typów fraz składnikowych. Można więc powiedzieć, że zdania proste zdają w GFJP sprawę ze szczególnych cech składniowych pewnych spójników podrzędnych. Świdziński wyróżnia cztery układy zdań prostych.

Zdania proste w układzie I to konstrukcje, w których spójnikowi podrzędnemu towarzyszy specyficzny dla niego „spójnik odpowiednik", np. JEŻELI – TO, przy czym zdanie wprowadzane przez odpowiednik (nazywany spójnikiem podrzędnym prawym) musi następować linearnie po frazie zdaniowej, jak na rysunku 2.24. Oznaczenie Oz typu spójnika podrzędnego w inicjalnej frazie zdaniowej musi być zgodne z typem spójnika podrzędnego prawego. Na przykład spójnikowi podrzędnemu DOPÓKI lub PÓKI odpowiada podrzędny prawy DOPÓTY lub PÓTY, a spójnikowi JEŻELI – spójnik TO.

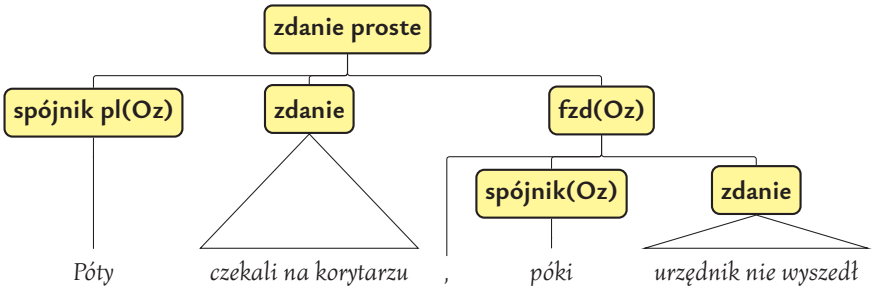


Rysunek 2.24. Zdanie proste w układzie I

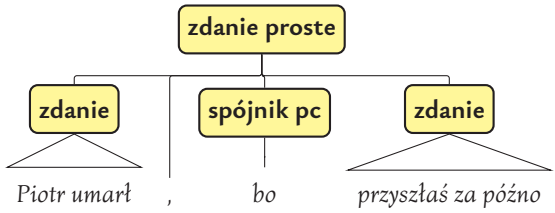
(183) – *Dopóki rząd i koalicja nie podejmą decyzji, dopóty każdy może wypowiadać swoje zdanie.* [Skł.]

(184) *Jeśli chodzi o majątek, to Najder zdążył przepisać swoje mieszkanie na Obożnej na syna.* [Skł.]

(185) *Jeżeli bandzior chce ograniczyć wolność innych, to powinien za to zapłacić.* [Skł.]



Rysunek 2.25. Zdanie proste w układzie II



Rysunek 2.26. Zdanie proste w układzie III

(186) *Gdybym miał pieniądze, to kupiłbym bez namysłu.* [Skł.]

Układ II zdania prostego stanowi lustrzane odbicie układu I, zob. rys. 2.25. Spójnik odpowiednik nazywany podrzędnym lewym wprowadza zdanie nadrzędne, które musi poprzedzać zdanie podrzędne. Oznaczenie Oz typu spójnika podrzędnego lewego musi uzgadniać się z typem spójnika podrzędnego w finalnej frazie zdaniowej.

(187) *Póty czekali, póki urzędnik nie wyszedł, i w końcu poszli.* [GFJP, (139), s. 140]

Układ III realizuje konstrukcje ze spójnikami nazywanymi przez Świdzińskiego podrzędnymi centralnymi. Wprowadzane przez nie zdanie podrzędne musi być drugim składnikiem konstrukcji, jak na rysunku 2.26.

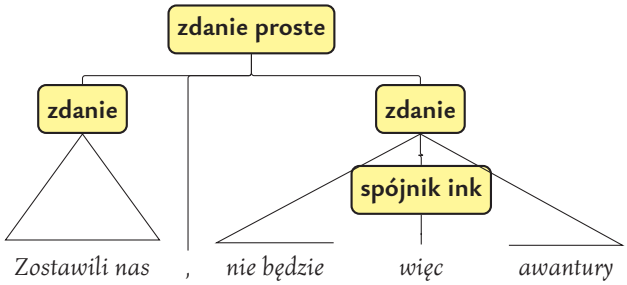
(188) *Piotr umarł, bo przyszłaś za późno.* [GFJP, (140), s. 140]

(189) *W trakcie remontu zauważył go nowy właściciel mieszkania, bo z piasku wystawała ręka.* [Skł.]

(190) *Ci oczywiście nie zrobili mu najmniejszej krzywdy, gdyż gość był szmalowny.* [Skł.]

(191) *Stary nie wiedział, o co chodzi, więc zatrzymał samochód niedaleko budynku telewizji.* [Skł.]

Układ IV zdania prostego został wprowadzony dla opisania podrzędnych spójników inkorporacyjnych WIĘC i BOWIEM. Podobnie jak w układzie III,



Rysunek 2.27. Zdanie proste w układzie IV

wprowadzane przez nie zdanie podrzędne musi być drugim składnikiem konstrukcji (por. rys. 2.27).

(192) *Pragnie pomagać ludziom, myśli więc też o zdawaniu na resocjalizację.* [Skł.]

(193) *W ten sposób mógł zarobić nawet 3,5 mln zł, z ustaleń gazety wynika bowiem, że chodzi o około 1500 osób.* [Skł.]

Dyskusja

Można postawić pytanie, czy pokazane struktury są optymalne, a w szczególności, czy ograniczenia szyku są wystarczającym powodem wprowadzenia nowej jednostki składniowej.

Na przykład zdania w układzie III ze spójnikiem BO są bardzo podobne do zdań z podrzędnikiem wprowadzanym spójnikiem PONIEWAŻ. Rysunek 2.28 przedstawia zarys drzewa dla następującego zdania analogicznego do przykładu (188):

(194) *Piotr umarł, ponieważ przyszłaś za późno.*

Jak się zdaje, zdanie (188) różni się od (194) jedynie usztywnionym porządkiem linearnym składników. Składnik wprowadzany przez PONIEWAŻ może zająć różne pozycje w zdaniu nadrzędnym, podczas gdy składnik wprowadzany przez BO musi być ostatni[8]:

(195) *Piotr, ponieważ przyszłaś za późno, umarł.*

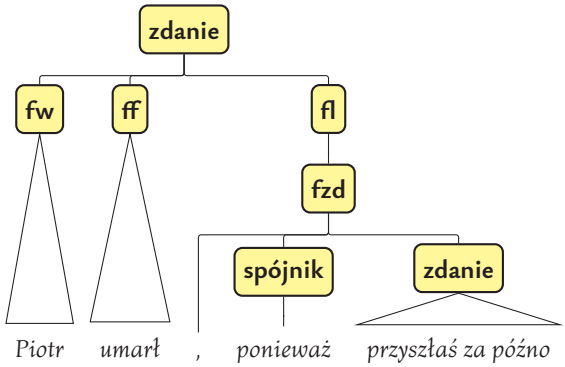
(196) *Ponieważ przyszłaś za późno, Piotr umarł.*

(197) **Piotr, bo przyszłaś za późno, umarł.*

(198) **Bo przyszłaś za późno, Piotr umarł.*

Również układ ograniczeń formy czasownikowej w zdaniu podrzędnym jest identyczny dla zdania wprowadzonego spójnikiem PONIEWAŻ w obrębie

[8] Tak przynajmniej przyjmują wydawnictwa poprawnościowe. W korpusie daje się znaleźć zdania ze spójnikiem BO użytym w innych pozycjach.



Rysunek 2.28. Struktura przypisywana zdaniu (194)

zdania elementarnego i w zdaniu wprowadzonym spójnikiem BO w obrębie zdania prostego. Mianowicie w obu wypadkach wykluczony jest tryb rozkazujący.

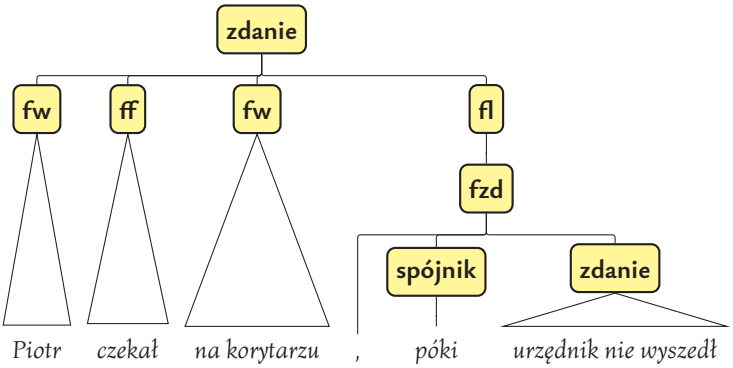
Jednak drzewa postulowane przez Świdzińskiego różnią się diametralnie. W zdaniu ze spójnikiem PONIEWAŻ (rys. 2.28) tworzy on frazę zdaniową. Z punktu widzenia zdania nadrzędnego (elementarnego) fraza zdaniowa jest podłączona jako fraza luźna stanowiąca współskładnik zarówno frazy finitywnej *umarł*, jak i podmiotu *Piotr*. Dla odmiany w zdaniu z *bo* (przykład (188)) nie ma frazy zdaniowej, fragment *, bo przyszłaś za późno* w ogóle nie jest składnikiem, co więcej w strukturze nie widać, żeby fragment z BO był podrzędnikiem czasownika UMARŁ, ponieważ występuje on na innym (wyższym) poziomie drzewa niż pozostałe podrzędniki czasownika (frazy werbalnej). W tej pracy przyjęto, że struktury dla obu rozpatrywanych zdań powinny być bardziej zbliżone do siebie.

Podobne wątpliwości budzą zdania proste w układzie II. Usztywnienie pozycji linearnej składnika wprowadzanego przez PÓKI jest spowodowane przez obecność spójnika podrzędnego lewego PÓTY. Bez niego bowiem szyk jest swobodny, np.:

(199) *Piotr czekał na korytarzu, póki urzędnik nie wyszedł.*
(200) *Piotr, póki urzędnik nie wyszedł, czekał na korytarzu.*

Zdania takie są interpretowane jako zawierające luźną frazę zdaniową ze spójnikiem PÓKI. Rysunek 2.29 przedstawia zarys takiego drzewa. Składnik wprowadzany przez spójnik PÓKI realizuje w nim frazę luźną, stanowiącą składnik zdania elementarnego. Podobnie jak w wypadku zdań ze spójnikami PONIEWAŻ i BO podobne zdania otrzymują istotnie różne interpretacje.

Obecność elementu PÓTY faktycznie powoduje, że składnik wprowadzany przez PÓKI musi znaleźć się na końcu zdania. Jednak struktura zaproponowa-



Rysunek 2.29. Struktura przypisywana zdaniu (199)

na w układzie II jest zbyt sztywna, składnik PÓTY może się bowiem przemieszczać w obrębie zdania nadrzędnego.

(201) *Póty Piotr czekał na korytarzu, póki urzędnik nie wyszedł.*
(202) *Piotr póty czekał na korytarzu, póki urzędnik nie wyszedł.*
(203) *Piotr czekał póty na korytarzu, póki urzędnik nie wyszedł.*
(204) *Piotr czekał na korytarzu póty, póki urzędnik nie wyszedł.*
(205) *Gdybyśmy dopóty wołali, dopóki urzędnik nie wyszedł, sprawa byłaby załatwiona.* [GFJP, (207b), s. 155]
(206) **Póty Piotr czekał na korytarzu.*

Niepoprawność ostatniego z przykładów pokazuje, że obecność elementu PÓTY wymaga pojawienia się składnika wprowadzanego przez PÓKI.

Zdanie (205) Świdziński przytacza jako niesłusznie nieakceptowane przez GFJP z powodu tego, że aglutynant *-śmy* stanowi pierwszy linearnie składnik zdania wprowadzanego przez spójnik GDYBY. Przez to zaś składnik DOPÓTY został zepchnięty z pierwszej pozycji linearnej w zdaniu prostym.

Wydaje się więc, że opis zdań prostych w układzie II powinien dopuszczać ruchomość spójnika lewego w obrębie zdania nadrzędnego, a jednocześnie zapewniać zdaniu podrzędnemu ostatnią pozycję w układzie linearnym. Preferowana byłaby jednak struktura analogiczna do przedstawionej na rysunku 2.29.

Dla kontrastu w układzie I ruchomość spójnika prawego w obrębie zdania nadrzędnego nie jest możliwa, nawet w wypadku tych samych spójników co w układzie II:

(207) *Póki urzędnik nie wyszedł, póty Piotr czekał na korytarzu.*
(208) **Póki urzędnik nie wyszedł, Piotr czekał póty na korytarzu.*

GFJP uwzględnia specyficzne ograniczenia trybu, negacji i aspektu wywoływane obecnością spójników tworzących zdania proste (GFJP, s. 143–158). Te

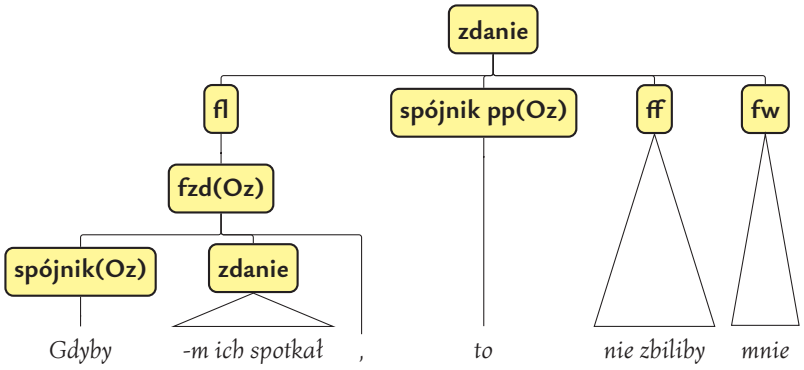
ograniczenia można jednak opisać za pomocą odpowiednich warunków bez wprowadzania specjalnej jednostki składniowej.

Postulowane struktury

Ponieważ w prezentowanej gramatyce stosowana jest tylko jedna jednostka **zdanie** dla wszystkich typów konstrukcji zdaniowych, więc i zdania proste są nią objęte. Przyjęto także, że zdania proste powinny przypominać inne realizacje zdań elementarnych, a więc że składnik wprowadzany przez spójnik podrzędny niosący ograniczenia linearne powinien znaleźć się wśród typowych składników zdania elementarnego (fraz finitywnej, wymaganych i luźnych). Sam ten składnik jest realizowany przez frazę luźną. Na potrzeby układu I i II potrzebne jest wprowadzenie spójnika odpowiednika jako składnika zdania elementarnego. Przecinki stanowiące jawne składniki zdań prostych stają się w tej interpretacji składnikami odpowiednich fraz zdaniowych.

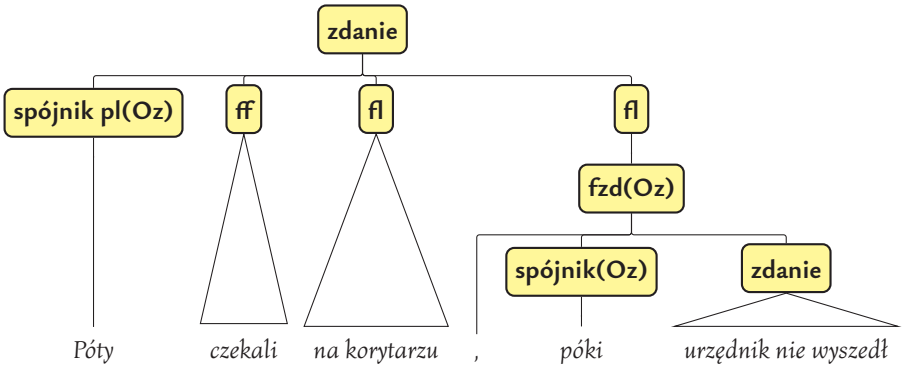
Nowe struktury zdań prostych są pozbawione sztucznie stworzonego dodatkowego poziomu zdaniowego, a cechująca się ograniczeniami pozycyjnymi fraza zdaniowa staje się współskładnikiem frazy finitywnej i pozostałych jej podrzędników[9].

Struktura odpowiadająca układowi I zdania prostego ma ustalony szyk początkowej części. Składają się na nią fraza zdaniowa pewnego typu (realizująca frazę luźną) i spójnik podrzędny prawy tego samego typu, zob. rys. 2.30. Dalej następują składniki zdania elementarnego: fraza finitywna, frazy wymagane i luźne.



Rysunek 2.30. Struktura odpowiadająca zdaniu prostemu w układzie I

[9] Postulowane tu struktury zostały zaproponowane w dość późnym stadium rozwoju gramatyki, dlatego w korpusie Składnica można jeszcze (koniec r. 2018) zobaczyć struktury zgodne z GFJP. Struktury te zostaną zmienione w jednej z przyszłych aktualizacji Składnicy.

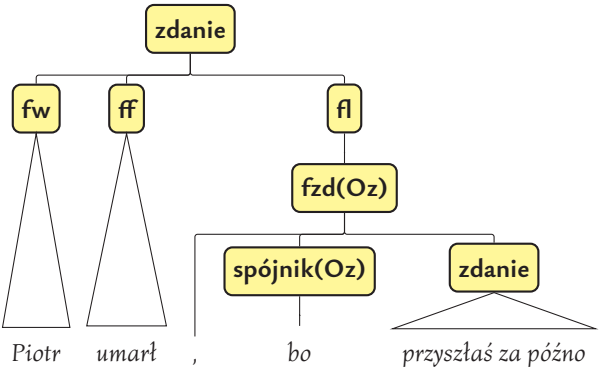


Rysunek 2.31. Struktura odpowiadająca zdaniu prostemu w układzie II

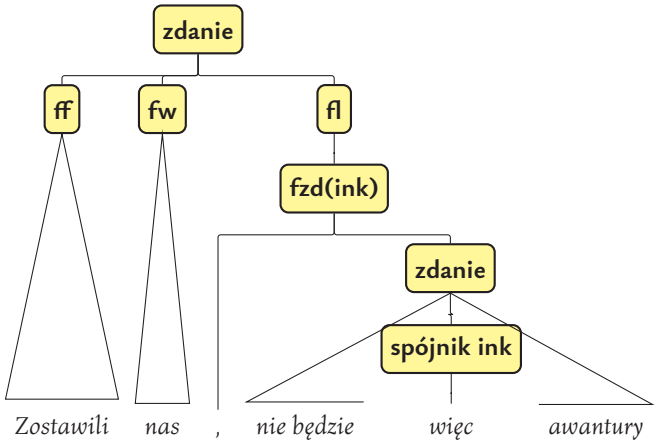
W układzie II fraza zdaniowa znajduje się na końcu konstrukcji, a poprzedzają ją składniki zdania elementarnego oraz spójnik podrzędny lewy zgodnego typu, w dowolnej kolejności (na rysunku 2.31 spójnik lewy zobrazowano jako pierwszy, ale nie jest to jedyna dopuszczalna możliwość).

Schemat ten może być traktowany jako instancja typowych schematów zdania elementarnego, jeżeli zostanie w nich przewidziany jako składnik spójnik podrzędny lewy. W gramatyce muszą być wyrażone następujące warunki: spójnik podrzędny lewy jest opcjonalny, jednak jeśli się pojawi, to musi wystąpić również fraza zdaniowa ze zgodnym typem spójnika i musi ona zajmować pozycję linearnie ostatnią. Przy braku spójnika lewego fraza zdaniowa jest opcjonalna i może zajmować dowolną pozycję.

Układ III przypomina typowy skład zdania elementarnego (zob. rys. 2.32) z dodanym ograniczeniem: fraza zdaniowa wprowadzana przez spójnik klasyfikowany jako podrzędny centralny musi być ostatnim składnikiem konstruk-



Rysunek 2.32. Struktura odpowiadająca zdaniu prostemu w układzie III



Rysunek 2.33. Struktura odpowiadająca zdaniu prostemu w układzie IV

cji. Spójniki podrzędne centralne zostają przy tym dodane do grupy spójników tworzących frazy zdaniowe (inaczej niż w GFJP).

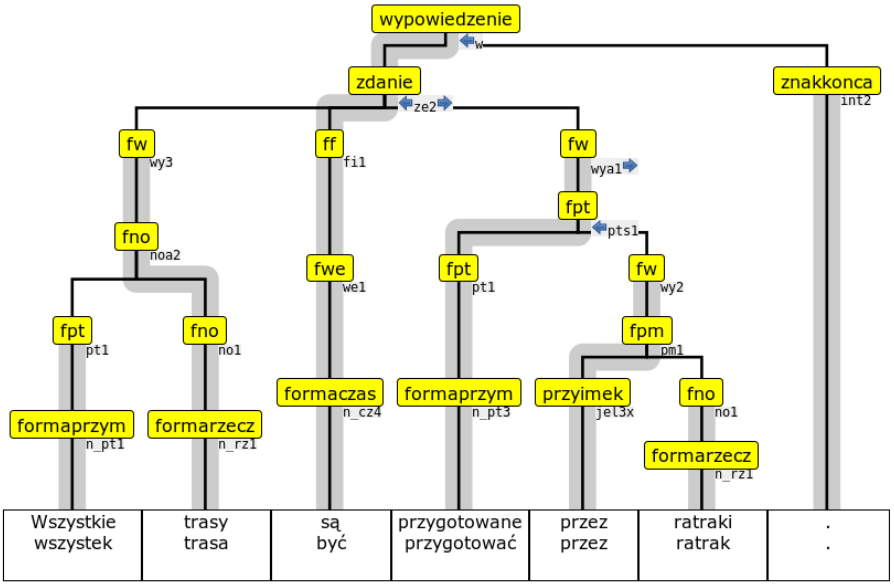
Układ IV (rys. 2.33) jest analogiczny do układu III, przy czym do jego realizacji potrzebne jest wprowadzenie nieobecnej w GFJP frazy zdaniowej inkorporacyjnej. Fraza taka jest realizowana przez zdanie ze spójnikiem podrzędnym inkorporacyjnym.

2.11. Strona bierna

Świdziński (1992) nie rozważa zjawiska strony biernej (w szczególności w całym tekście książki nie pojawia się ani jedna forma leksemu BIERNY). W implementacji GFJP przyjęto, że strona bierna jest realizowana z użyciem schematu dla czasowników wprowadzających tę konstrukcję (BYĆ i ZOSTAĆ) dopuszczającego podmiot i wymaganą frazę przymiotnikową o oznaczeniu adjp(pred) (zob. p. 3.1.7), której centrum staje się forma imiesłowu przymiotnikowego biernego. Rysunek 2.34 przedstawia przykład takiej analizy dla zdania (210), stanowiącego wynik przekształcenia na stronę bierną zdania (209).

(209) *Ratraki przygotowały wszystkie trasy.*
(210) *Wszystkie trasy są przygotowane przez ratraki.*

W tym przykładzie forma *są* ma zrealizowany podmiot nominalny *wszystkie trasy* i wymaganą frazę przymiotnikową w mianowniku, której centrum jest imiesłów przymiotnikowy bierny *przygotowane*. W wyniku przejścia na stronę bierną schemat dla czasownika PRZYGOTOWAĆ jest realizowany przez podrzędniki formy *przygotowane* w ten sposób, że podmiot nominalny zamienia się we



Rysunek 2.34. Przykład struktury dla zdania w stronie biernej

frazę przyimkowo-nominalną *przez ratraki*, natomiast fraza biernikowa *wszystkie trasy* występująca w konstrukcji czynnej nie może być zrealizowana przy imiesłowie – staje się ona podmiotem nadrzędnego czasownika *są*.

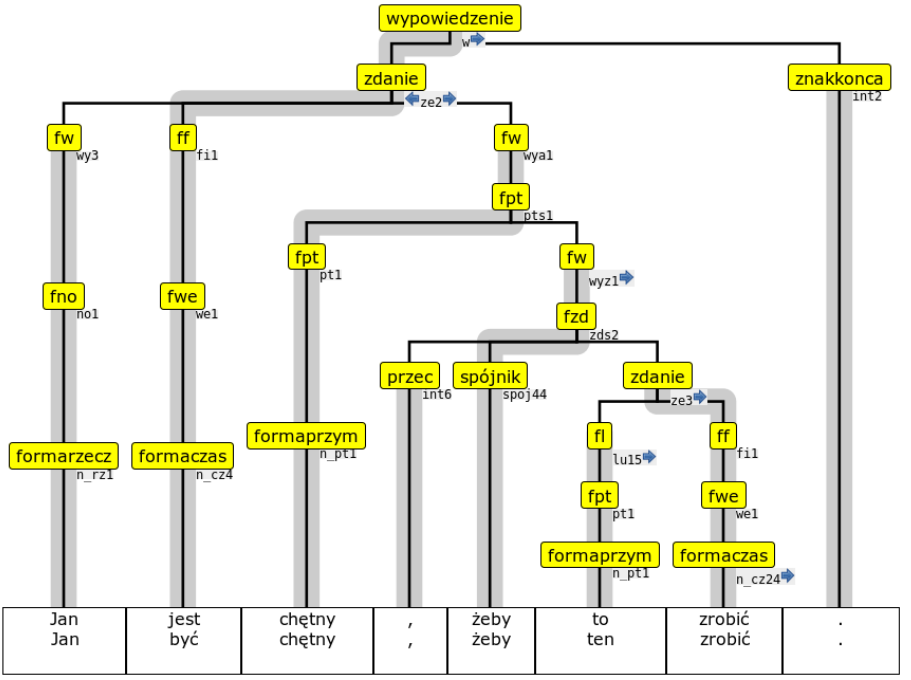
Tworzona struktura jest analogiczna do zdań z czasownikiem BYĆ lub ZOSTAĆ, w których występuje wymagana fraza przymiotnikowa w mianowniku, której centrum jest regularny przymiotnik. Przykład takiej struktury dla zdania (211) obrazuje rysunek 2.35.

(211) *Jan jest chętny, żeby to zrobić.*

Słownik walencyjny używany w implementacji GFJP nie zawierał informacji o możliwości realizacji biernych. W związku z tym zakładano, że jeżeli istnieje forma oznaczona jako imiesłów bierny, to może ona zostać użyta w konstrukcji biernej. Było to rozwiązanie przybliżone. Część czasowników, nazywanych w SGJP quasi-tranzytywnymi, tworzy formę o kształcie imiesłowu biernego, która jednak nie wchodzi w konstrukcję bierną, np. CHYBIONY, KLAPNIĘTY, NABURMUSZONY:

(212) *Piotr chybił celu.*
(213) **Cel został chybiony przez Piotra.*
(214) *Okazało się to pomysłem chybionym, bo chłopiec nie mógł się odnaleźć w świecie małych protestantów z zamożnych rodzin.* [NKJP300]

Jednak nawet jeśli czasownik tworzy konstrukcję bierną, to nie wszystkie związane z nim schematy składniowe muszą w niej być dopuszczalne.



Rysunek 2.35. Przykład struktury dla zdania z wymaganą frazą przymiotnikową w mianowniku

Dokładniejsza informacja o możliwości tworzenia strony biernej jest dostępna w słowniku Walenty. Wyróżniana w schematach pozycja obj wskazuje podrzędnik, który odpowiada podmiotowi czasownika nadrzędnego w konstrukcji biernej (zob. p. 3.1.3). Obecność tej pozycji wskazuje pośrednio, że dany schemat może być realizowany w stronie biernej.

Przyjęto, że konstrukcja bierna jest tworzona przez czasowniki BYĆ i ZOSTAĆ. Wykorzystywany jest przy tym dostępny dla obu czasowników schemat z podmiotem i frazą przymiotnikową adjp(pred). Ten typ frazy dopuszcza w tym wypadku realizacje w mianowniku lub narzędniku, możliwe są więc również konstrukcje typu:

(215) *Za młody! – sprzeciwił się wielki książę, gdy mu podano lisię podchorążych mających być przedstawionymi cesarzowi do nominacji.* [Leon Kruczkowski, *Kordian i cham*; za NKJP300]

(216) *Oszołomiony przepychem naczyń i mebli, które zdawały się być stworzonymi dla świątyni, [...] usiłował zachować na twarzy wyraz doskonałej obojętności* [Stefan Żeromski, *Popioły*; za NKJP300]

(217) *Ale pamiętaj, jesteś związanym przysięgą i pod żadnym pozorem nie wolno ci nikomu wspominać o Konklawe, nawet jemu.* [Raymond E. Feist, *Król Lisów*; za NKJP300]

Schemat ze słownika Walenty uwzględnia kontrolę argumentu adjp(pred) przez podmiot, co na poziomie składniowym przekłada się na uzgodnienie rodzaju i liczby frazy przymiotnikowej z podmiotem czasownika finitywnego. Zjawisko to zachodzi zarówno fraz przymiotnikowych, których centrum jest faktyczny przymiotnik (np. (211)), jak i zawierających imiesłów bierny i realizujących stronę bierną. Obecny opis niestety nie uwzględnia tego, że w zdaniach z ZOSTAĆ możliwe jest jedynie użycie imiesłowów czasowników dokonanych.

Schematy bierne mogą być realizowane także poza kontekstem strony biernej, na przykład gdy fraza przymiotnikowa z centrum imiesłowowym jest podrzędnikiem rzeczownika:

(218) *Po raz pierwszy taka sytuacja zdarzyła się w majątku administrowanym przez kapitana Charlesa Boycotta, stąd podobne akcje nazwano bojkotem.* [NKJP300]

2.12. Zdanioidy

W gramatyce uwzględniono konstrukcje, które składają się z elementów właściwych dla zdania, ale którym brakuje finitywnego nadrzędnika. Oto przykład wypowiedzenia, które można usłyszeć w rejestracji przychodni lekarskiej:

(219) *– Ja do internisty.*

Jego interesującą cechą jest to, że nie bardzo wiadomo, jakiego czasownika w nim „brakuje" (*przyszedłem, chcę, życzę sobie*?). Wypowiedzenie to jest kompletne bez finitywnej formy czasownika (choć należy do rejestru potocznego). Wyrazisty jest również jego podział na składniki – są to fraza nominalna w mianowniku *ja* i fraza przyimkowo-nominalna *do internisty*. Obie frazy można intuicyjnie zinterpretować jako wymagane, przy czym pierwsza stanowiłaby podmiot niewidocznego predykatu. O drugą z fraz w kontekście przychodni z pewnością zada się pytanie *Pan do kogo*? a nie *Pan dokąd*?, co wskazuje na jej typ (por. p. 3.1.5).

Oto więcej przykładów konstrukcji bezczasownikowych stanowiących samodzielne wypowiedzenia, które mogą stanowić samoistny komunikat (a nie uzupełnienie wcześniejszej wypowiedzi z jawną formą czasownika). W części z nich byłoby bardzo trudno dodać jakąkolwiek formę czasownikową, nie burząc jednocześnie zupełnie struktury (np. trudno sobie wyobrazić takie przekształcenie wypowiedzenia (221)). Niektóre wypowiedzenia można by uzupełnić formą leksemu BYĆ lub TO, ale dałoby to wypowiedzenia brzmiące mniej naturalnie (np. (226) i (227)).

(220) *– Ja do biblioteki.* [Skł.]
(221) *Do zobaczenia w Krakowie!*

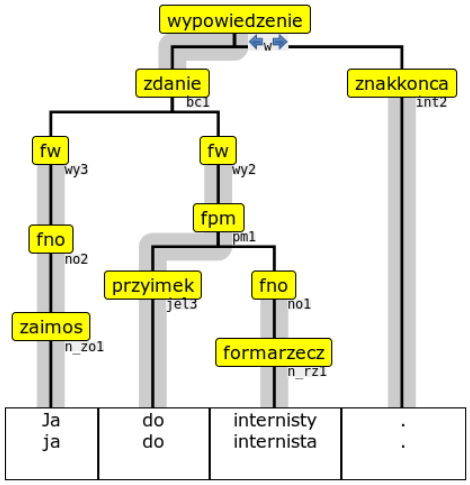
(222) *Wszystko przez czynsze, bilety i kartofle.* [Skł.]
(223) *Tyle relacja Franciszka Ancewicza.* [Skł.]
(224) *Ręce do tyłu.* [Skł.]
(225) *Pytanie, czy IPN ma prawo do tworzenia takiej listy?* [Skł.]
(226) *Sęk w tym, że nie wiadomo, czy Kazancew w ogóle wsiadł do śmigłowca.* [Skł.]
(227) *Początek meczu o godz. 12 na stadionie w Maszewie.* [Skł.]
(228) *– Po diabła ci nauki realne?* [Skł.]
(229) *Prezes jak zwykle w znakomitej formie.* [Skł.]
(230) *W programie wiele konkursów z odjazdowymi nagrodami!* [Skł.]
(231) *Dziś kolejna próba wyłonienia kandydata.* [Skł.]
(232) *Czy pan poseł sekretarz?* [Skł.]
(233) *– Mamo, dziś zebranie rodzicielskie w szkole.* [Skł.]
(234) *Dla nas koniec jazdy.* [Skł.]
(235) *Przed kasami kolejowymi tłum spoconych i zrezygnowanych podróżnych.* [Skł.]
(236) *W jednym przedziale tyle osób...* [Skł.]
(237) *Za nim wyjątkowy rok.* [Skł.]
(238) *Za nią obszar bez dna, bez echa.* [Skł.]
(239) *Przede mną krawędź.* [Skł.]
(240) *Przy nim dwaj blondyni.* [Skł.]
(241) *Na pozostałym obszarze po rannych mgłach pogodnie.* [Skł.]
(242) *Przy nim krzesło, które też ma swoją bogatą historię...* [Skł.]
(243) *Za łąkami bór ogromny, niezmierzony, niezdeptany.* [Skł.]

Tego rodzaju konstrukcje, nazwane roboczo zdanioidami bezczasownikowymi, opisano jako szczególne realizacje jednostki **zdanie** (zob. rys. 2.36). Konstrukcja taka jest ciągiem fraz wymaganych i luźnych. Nie ma ona centrum – jest nim nieobecny czasownik. W gramatyce Świgra 2 dopuszczono taką realizację jedynie dla zdanioidów samodzielnie stanowiących wypowiedzenie (a więc w szczególności nie pozwala się na połączenia współrzędne zdanioidów). W tym zakresie opis może zostać ostrożnie rozszerzony.

Opis w postaci zdanioidu przyjęto także dla wypowiedzeń niesamodzielnych, mających postać frazy lub fraz „oderwanych" od poprzedniego wypowiedzenia i stanowiących jakby podrzędniki czasownika z tego poprzedniego zdania:

(244) *– Co jest za tym murem?*
– Szarość. Magma.
(245) *– Czy ty myślisz o mnie, kiedy się nie widzimy?*
– Myślę wtedy, gdy ty o mnie myślisz. Codziennie.
(246) *List muszę w tej sprawie napisać. Do męża. Ślubnego zresztą.*

W przykładzie (244) konstrukcje *szarość* i *magma* są interpretowane jako zdanioidy realizowane przez pojedynczą frazę wymaganą podmiotową, bo kon-



Rysunek 2.36. Drzewo dla wypowiedzenia (219)

strukcje te można by rozbudować do pełnych odpowiedzi *Za tym murem jest szarość/magma.* W przykładzie (245) *codziennie* jest traktowane jako zdanioid realizowany przez frazę luźną, bo frazą luźną byłoby w zdaniu *Codziennie myślę.* W przykładzie (246) wypowiedzenie *do męża* można traktować jako dopowiedzianą frazę wymaganą dla poprzedzającego *napisać*. Drugi zdanioid jest pod tym względem ciekawszy, ponieważ *ślubnego* nie może być interpretowane jako podrzędnik czasownika, tylko rzeczownika *męża*. Przyjęto konwencję interpretowania takiej konstrukcji jako zdanioidu zawierającego wymaganą frazę przymiotnikową.

Zdanioidy stanowiące rozwinięcie poprzedniego wypowiedzenia mogą zawierać wszelkie typy fraz występujące w zdaniach jako wymagane lub luźne, w szczególności trafiają się w tekstach usamodzielnione frazy zdaniowe wprowadzane spójnikiem podrzędnym lub frazy względne. Zdanioidy stanowiące wypowiedzenia samodzielne wydają się być pod tym względem dużo bardziej ograniczone.

Opis w postaci zdanioidu został przyjęty również dla wszelkich usamodzielnionych fraz, na przykład fraz nominalnych stanowiących tytuł artykułu prasowego (np. *Skok na SKOK-i*) lub didaskalium (np. *(Oklaski)*). Na zasadzie konwencji przyjęto, że jeżeli nie ma wyraźnych wskazówek wynikających z kontekstu, taka samodzielna fraza jest traktowana jako wymagana. W tym wypadku dopuszczany jest brak znaku interpunkcyjnego na końcu wypowiedzenia.

2.13. Nieciągłe formy analityczne czasowników

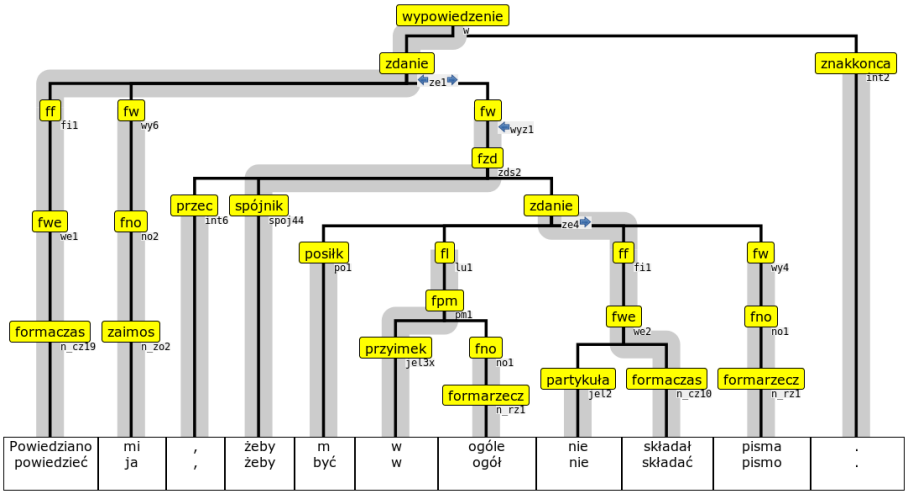
Jak wspomniano na s. 45, w kontekście pewnych spójników (kończących się cząstką *-by*: CHOĆBY, CZYŻBY, GDYBY, JAKBY, JAKOBY, ŻEBY) następuje rozbicie formy czasu przeszłego czasownika na aglutynant (np. *-ście*) i pseudoimiesłów (np. *czytali*). Elementy te mogą być oddzielone od siebie innymi składnikami zdania:

(247) – *Najpierw powiedziano mi, żeby m w ogóle nie składał pisma w tej sprawie, bo takich jak ja jest wielu.* [Skł.]

Zjawisko to zostało uwzględnione przez Świdzińskiego (1992, §6.2.1 i §6.2.5.1) w ten sposób, że inicjalnym składnikiem zdania elementarnego może być fraza luźna realizowana nietypowo przez aglutynant. Za pomocą odpowiednich wartości atrybutów reguły wymuszają, aby tak realizowane zdanie mogło występować jedynie w kontekście odpowiednich spójników.

Interpretacja aglutynantu jako frazy luźnej wydaje się zabiegiem czysto technicznym ujednolicającym skład zdania elementarnego. Trudno dla niej znaleźć uzasadnienie lingwistyczne. Z punktu widzenia językowego aglutynant taki stanowi oderwaną część analitycznej formy czasownika. Taką interpretację postulują Saloni i Świdziński (2001, s. 62), jednocześnie zakładając nieuwzględnianie szyku w analizie składnikowej. Podejście takie było niemożliwe do zastosowania w GFJP.

W niniejszej gramatyce uwzględniono nieciągłe formy analityczne w większym zakresie. Dla ich opisu wprowadzono jednostkę składniową **posiłk**, która reprezentuje element pomocniczy formy analitycznej. Jednostka ta staje się składnikiem bezpośrednim zdania (por. rys. 2.37). Przy jej obecności pewne



Rysunek 2.37. Drzewo dla zdania *Powiedziano mi, żebym w ogóle nie składał pisma.*

atrybuty zdania (np. tryb) są obliczane na podstawie opisu frazy finitywnej i jednostki **posiłk**.

Jednostka **posiłk** może być realizowana przez:

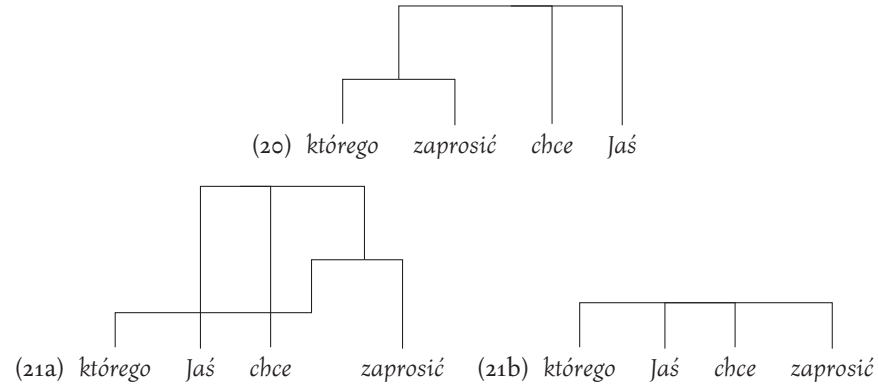
– aglutynant – dla analitycznego czasu przeszłego, por. przykład (247);
– partykułę NIECH – dla analitycznego trybu rozkazującego, por. (248);
– partykułę BY i ewentualny aglutynant – dla analitycznego trybu warunkowego, por. (249) i (250);
– formę czasu przeszłego czasownika BYĆ – dla analitycznego wariantu nierzeczywistego trybu warunkowego, por. (251);
– formę czasu przyszłego czasownika BYĆ – dla analitycznego czasu przyszłego, por. (252) i (253).

(248) *– Panie podchorąży, niech się pan pośpieszy.* [Skł.]
(249) *Nigdy by mężowi coś takiego nie przyszło do głowy!* [Skł.]
(250) *Może byśmy tak załatwili Mafię wspólnie?* [Skł.]
(251) *Czy naprawdę byłbym jej coś zrobił za Zucha?* [Skł.]
(252) *– Ale ich liczba w Europie będzie szybko rosnąć.* [Skł.]
(253) *– Będzie pan rozmawiał z doktorem Kazniczem.* [Skł.]

W regułach gramatyki przyjęto, że jednostka **posiłk** może wystąpić w zdaniu tylko raz, na dowolnej pozycji, ale nie w bezpośrednim sąsiedztwie frazy finitywnej **ff** (w tym wypadku forma analityczna byłaby ciągła). Wyjątkowo jako **posiłk** jest zawsze traktowany aglutynant na pozycji inicjalnej. Informacja o jego wystąpieniu jest odzwierciedlana w charakterystyce zdania, aby dopuścić jego wystąpienie tylko w kontekście odpowiednich spójników (por. p. 4.4.1). Uwzględniane są także ograniczenia szyku: element posiłkowy nie może stać za czasownikiem głównym w wariancie analitycznym czasu przeszłego (**czytał dlaczego·m*) ani trybu rozkazującego (**czyta książki niech*). Bardziej naturalnie brzmi wariant z elementem posiłkowym przed czasownikiem, jest to jednak raczej preferencja niż kategoryczna zasada, reguły dopuszczają więc konstrukcje typu *czytać to będzie*.

Przedstawiona tu reprezentacja z jednostką **posiłk** stosowana jest tylko wtedy, gdy element posiłkowy jest odsunięty od formy czasownikowej. Ciągłe formy analityczne czasowników są opisywane w obrębie jednostki **formaczas**, podobnie jak w GFJP (por. Świdziński 1992, §7.2.6, s. 233).

Warto również w tym miejscu zwrócić uwagę na interpretację składniową segmentu *się*. Podobnie jak w NKJP w niniejszym opisie przyjęto, że wszystkie jego wystąpienia są interpretowane na poziomie fleksyjnym jako jedyna forma partykuły SIĘ. W tradycyjnych opisach niektóre wystąpienia *się* (tzw. *się* inherentne) są traktowane jako część analitycznej zwrotnej formy czasownika (np. *bać się*). W niniejszym opisie w konsekwencji przyjętego opisu fleksyjnego postanowiono nie wyróżniać tych wystąpień. Forma *się* jest realizacją frazy wymaganej **fw** o typie frazy wymaganej się. Zaniedbano także obecne w Walentym rozróżnienie *się* wzajemnościowego.



Rysunek 2.38. Struktury dla przykładów (255) i (254) postulowane przez Świdzińskiego (1992, ilustracje (20) i (21) na s. 67 i 68)

2.14. Problem nieciągłości

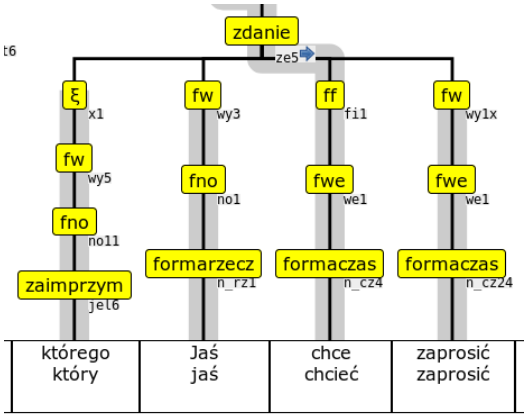
Konstrukcję składniową nazywa się nieciągłą, jeżeli jej składniki nie sąsiadują linearnie ze sobą (Świdziński 1996, s. 113). Przedstawiony tu opis takich konstrukcji jest niejednorodny, odzwierciedla bowiem stopniowe wprowadzanie ich do gramatyki.

Przyjęte zasady strukturyzacji sprawiają, że pewne konstrukcje, które można widzieć jako nieciągłe, nie są tak traktowane. Dotyczy to rozbijania „grupy orzeczenia" przez podmiot (zob. s. 80), opisu spójników inkorporacyjnych (zob. s. 109) i form analitycznych czasowników, w których partykuły SIĘ, NIECH, BY lub aglutynant są odseparowane od formy czasownika (p. 2.13). Rozwiązanie takie przyjęto, aby uniknąć rozbudowy stosowanego formalizmu gramatycznego. Jednak wymienione typy nie wyczerpują palety nieciągłości w języku polskim, które są częste tekstowo.

Świdziński (1992, §4.2.4, s. 66–68) zauważa, że w pewnych konstrukcjach, szczególnie „w nienacechowanych zdaniach pytajnych, we wszystkich zdaniach składowych podrzędnych pytajnozależnych lub względnych składnik pytajny lub względny zajmuje pozycję inicjalną, bez względu na to, jak usytuowany jest strukturalnie". Taka wymuszona pozycja linearna składnika powoduje, że często bywa on „odrywany" od swojego nadrzędnika, co powoduje, że pewne konstrukcje w naturalny sposób są nieciągłe. Świdziński przytacza następujące przykłady i rozważa dla nich struktury przedstawione na rysunku 2.38:

(254) *którego zaprosić chce Jaś*
(255) *którego Jaś chce zaprosić*

Pierwszy z przykładów jest ciągły i w naturalny sposób odpowiada mu pierwsza z pokazanych struktur. Podrzędnik *którego* stoi obok swojego nadrzędni-



Rysunek 2.39. Drzewo dla zdania składowego *którego Jaś chce zaprosić* generowane przez analizator Świgra 2

ka *zaprosić*. Fraza ta brzmi jednak mniej naturalnie niż w przykładzie (255), w którym fraza *którego zaprosić* została rozbita składnikami zawierającego ją strukturalnie zdania z orzeczeniem *chce*.

Świdziński pisze (s. 68), że „gałęzie drzewa nie mogą się przecinać" (w domyśle: przy założonym formalizmie gramatycznym) i dlatego odrzuca drugą ze struktur przedstawionych na rysunku 2.38, mimo że jest ona „całkowicie zgodna z intuicją". W odniesieniu do trzeciej z pokazanych struktur pisze, że nie zdaje ona sprawy z tego, że *którego* jest podrzędnikiem *zaprosić*, w odróżnieniu od *Jaś*, który składniowo jest podrzędnikiem *chce*.

Typowe nieciągłości nie ograniczają się do wymienionych przypadków. Element zakłócający ciągłość frazy wymaganej nie musi być pytajny ani względny.

(256) *Marię chce zaprosić Jaś.*
(257) *O delfinach może mówić godzinami.*
(258) *Bracia Dawid i Damian umieszczeni zostali w domu małego dziecka.*

Ze wstępnej analizy korpusu Składnica wynika, że najczęstsza nieciągłość polega na wysunięciu pewnej frazy wymaganej na pozycję inicjalną zdania (jak w przykładach (255), (256) i (257)).

Przyjęto reprezentację takich zdań przedstawioną na rysunku 2.39. Fraza *którego* jest wprowadzona do struktury przy czasowniku *chce*, jednak jej szczególny status jest sygnalizowany poprzez obecność dodatkowego nieterminala oznaczonego literą ξ[10]. Oznaczenie to zostało wprowadzone przede wszystkim po to, żeby zwrócić uwagę, że dany składnik nie jest po prostu frazą wymaganą o takim samym statusie jak pozostałe dzieci danego wierzchołka. Odpowiedni układ warunków, omówiony w punkcie 4.5.5, zapewnia sprawdzenie,

[10] Oznaczenie to ma się kojarzyć z greckim *ksenos* (ξένος) i sygnalizować, że ten element w istocie nie przynależy do tego miejsca struktury, jest „obcy".

że jedyny składnik jednostki ξ jest dopuszczalnym i niezrealizowanym w jej obrębie podrzędnikiem następującej jako współskładnik frazy bezokolicznikowej. Fakt ten jest odnotowywany w opisie argumentów wypełnionych przy danych predykatach. W omawianym zdaniu czasownik *chce* ma wypełnioną pozycję podmiotu nominalnego *Jaś* i frazy wymaganej bezokolicznikowej *zaprosić*, natomiast czasownik *zaprosić* – pozycję wymaganej frazy nominalnej biernikowej *którego*. Warto przy tym zauważyć, że w tej konstrukcji obecność składnika względnego *którego* powoduje, że względne staje się zdanie z czasownikiem *chce*, mimo że jest on składnikiem bezpośrednim zdania z czasownikiem *zaprosić*.

Motywacja dla przyjęcia takiego rozwiązania jest techniczna: zostało ono wprowadzone późno, a wypracowane wcześniej narzędzia budowy korpusu nie umożliwiają wizualizacji struktur z przecinającymi się krawędziami. Dlatego anotatorzy Składnicy pracują na strukturach, które system jest w stanie wyświetlić. Warto jednak podkreślić, że koncepcyjnie jest to struktura nieciągła i po ujednoznacznieniu drzewa mogą zostać skonwertowane do postaci z przecinającymi się krawędziami (por. rys. 4.6 w p. 4.5.5).

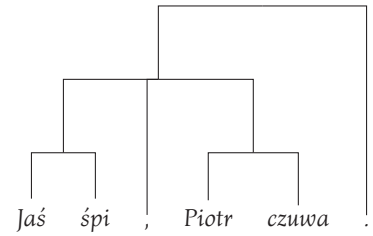
Inne rozstrzygnięcie przyjęto w odniesieniu do konstrukcji, w których źródłem nieciągłości jest fraza luźna. Dopuszczono mianowicie, aby frazy luźne zostawały uznane za podrzędnik na wyższym poziomie, niż by to wynikało z analizy składnikowej. Częściowym uzasadnieniem dla takiej decyzji jest to, że frazy luźne nie są konotowane (por. Świdziński 1992, s. 65), więc ich związek z konkretnym nadrzędnikiem ma w istocie naturę czysto semantyczną.

2.15. Interpunkcja

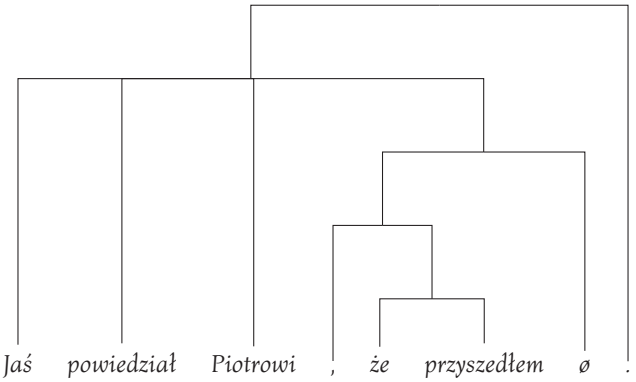
Świdziński uwzględnia w opisie znaki interpunkcyjne i traktuje je jako składniki konstrukcji składniowych (1992, s. 68). Podobnie Saloni i Świdziński (2001, §6.3, s. 66) argumentują, że niektóre przecinki pełnią funkcję spójnikową, a w związku z tym sensownie jest wyróżniać znaki interpunkcyjne jako samodzielne składniki przy analizie składnikowej.

W konsekwencji w taki sam sposób traktowane są znaki interpunkcyjne w prezentowanej tu gramatyce. Końcowym elementem wypowiedzenia jest jednostka **znakkońca** realizowana przez znak interpunkcyjny (kropkę, pytajnik, wykrzyknik i ich kombinacje). Znak końca jest charakteryzowany ze względu na pytajność i cecha ta musi zgadzać się z następującym wcześniej zdaniem. Uwzględniane są także myślniki wprowadzające kwestie dialogowe i wydzielające mowę niezależną, nawiasy i cudzysłowy wydzielające fragmenty wypowiedzenia.

Trudniejsze do interpretacji są przecinki. Niektóre przecinki pełnią funkcję elementu spajającego konstrukcję, a więc spójnika równorzędnego jak na



Rysunek 2.40. Postulowana przez Świdzińskiego struktura zdania z przecinkiem w roli spójnika współrzędnego (1992, adaptacja drzewa (24) na s. 70; por. także regułę (spój17) na s. 422)



Rysunek 2.41. Postulowana przez Świdzińskiego struktura z przecinkami ortograficznymi (1992, drzewo (25) na s. 70)

rysunku 2.40. Inne przecinki są wymagane jedynie regułami ortografii, jak przecinki okalające frazę zdaniową wprowadzaną spójnikiem ŻE na rysunku 2.41. Przy tym przecinek zamykający tę frazę ma realizację pustą ze względu na sąsiedztwo następującej dalej kropki kończącej zdanie. Opisanie tego zjawiska w gramatyce wymagało wprowadzenia dość skomplikowanego mechanizmu opisanego w punkcie 4.4.5. Na ilustracji 2.41 widać też echa binarnej strukturyzacji fraz w GFJP: każdy z przecinków jest wprowadzany na osobnym poziomie struktury drzewa. W drzewie wygenerowanym przez analizator Świgra 2 oba przecinki byłyby współskładnikami spójnika *że*.

2.16. Podsumowanie

Bieżący rozdział prezentuje gramatykę Świgra 2 niejako na tle GFJP, często poprzez konfrontowanie z nią. Warto może w sposób syntetyczny zestawić różnice między tymi dwiema gramatykami. Cechy charakterystyczne prezentowanej gramatyki to:

- konsekwentne odejście od strukturyzacji binarnych we wszystkich typach fraz opisywanych w gramatyce, w szczególności zmiana strukturyzacji zdania z centrum finitywnym na bardziej naturalną;
- znaczące zmniejszenie liczby kategorii składniowych, którymi operuje gramatyka, dzięki likwidacji rozbudowanych hierarchii jednostek;
- w konsekwencji znaczące uproszczenie drzew składniowych: drzewa są rozgałęzione, ale niskie;
- systematyczne wprowadzenie konstrukcji współrzędnych we wszystkich typach fraz, gdzie ma to sens;
- uwzględnienie nieredukowalnych konstrukcji nominalnych z koordynacją;
- dopuszczenie możliwości koordynowania fraz wymaganych różnych typów (zgodnie ze słownikiem walencyjnym);
- uwzględnienie możliwości podzielania podrzędników przez wiele fraz;
- opis konstrukcji apozycyjnych w obrębie fraz nominalnych;
- wprowadzenie liczebników jako możliwego składnika fraz nominalnych;
- dopuszczenie modyfikacji partykułą nadrzędników różnych typów;
- opis częstych konstrukcji nieciągłych, w tym analitycznych form czasowników oraz konstrukcji, w których fraza wymagana jest oddzielana od swojego nadrzędnika innym składnikiem;
- uwzględnienie konstrukcji zdaniopodobnych bez centrum finitywnego.

Choć gramatyka Świgra 2 w oczywisty sposób wywodzi się z GFJP, to przyjęty w niej opis różni się w sposób istotny od pierwowzoru. Prezentowane struktury są bardziej przejrzyste, w większym stopniu odpowiadają intuicjom językowym i są lepiej przystosowane do potrzeb przetwarzania komputerowego. Wyraźniejsze są też analogie w opisie fraz różnych typów. Dzieje się tak dzięki wyróżnieniu schematów konstrukcji (zwłaszcza współrzędnych) i konsekwentnemu ich zastosowaniu. Dodane typy konstrukcji zauważalnie zwiększają liczbę wypowiedzeń polskich akceptowanych przez analizator (por. p. 6.6).

3

Słownik walencyjny Walenty

Zasobem, który jest niezmiernie istotny dla jakości wyników analizy składniowej, jest słownik walencyjny. Jest to opis tego, z jakimi innymi jednostkami mogą wiązać się dane jednostki językowe. Poszczególne jednostki językowe „wybierają" towarzystwo, w jakim występują w tekstach, co łatwo dostrzec na przykładzie czasowników. Następujące przykłady pokazują, że własność konotowania frazy zdaniowej ze spójnikiem ŻE i frazy bezokolicznikowej jest indywidualną cechą poszczególnych czasowników. Na przykład czasownik OBIECAĆ konotuje oba typy fraz, czasownik POMYŚLEĆ – tylko frazę zdaniową, a czasownik ZDOŁAĆ – tylko frazę bezokolicznikową:

(1) *Marszałek obiecał, że połączenia kolejowe zostaną przywrócone.*
(2) *Marszałek obiecał przywrócić połączenia kolejowe.*
(3) *Marszałek pomyślał, że połączenia kolejowe zostaną przywrócone.*
(4) **Marszałek pomyślał przywrócić połączenia kolejowe.*
(5) **Marszałek zdołał, że połączenia kolejowe zostaną przywrócone.*
(6) *Marszałek zdołał przywrócić połączenia kolejowe.*

Jest też bardzo wiele czasowników (np. PRAĆ, SPACEROWAĆ), przy których nie może wystąpić żadna z tych fraz.

Tego rodzaju zależności warto zapisać w słowniku. Należy przy tym zauważyć, że frekwencja różnych typów podrzędników przy czasownikach bardzo się różni. Najbardziej typowymi argumentami są uzgadniająca fraza nominalna w mianowniku (podmiot) i fraza nominalna w bierniku. Jednak i one mogą pojawić się nie przy każdym czasowniku. Frazy wymienione w przykładach (1)– (6) są dopuszczalne tylko przy niektórych czasownikach. Natomiast podrzędnik w postaci frazy przysłówkowej (np. wyrażający sposób lub czas wykonania czynności) może wystąpić praktycznie przy każdym czasowniku, nie warto więc notować go w słowniku. Kwestia ustalenia, które z podrzędników powinny być notowane w słowniku, jest trudna, wiele pisze o niej Przepiórkowski (np. 2016, 2017). W niniejszej pracy przyjęto pragmatycznie, że jako argumenty będą traktowane te podrzędniki, które za takie zostały uznane w słowniku Walenty.

W implementacji GFJP (Woliński 2004) używany był słownik walencyjny tworzony *ad hoc* na potrzeby analizy poszczególnych przykładów, które stanowiły korpus testowy. Słownik ten liczył kilkaset pozycji.

Następnie, w ramach prac nad budową korpusu składniowego Składnica (rozdz. 6), wdrożono zaadaptowaną wersję słownika walencyjnego opracowanego przez Świdzińskiego (1994). Słownik ten został rozbudowany do ok. 1400 jednostek, co odpowiadało opisaniu ok. 75% form czasowników występujących w milionowym zrównoważonym podkorpusie NKJP.

Potrzeba dysponowania obszernym słownikiem walencyjnym języka polskiego była istotna również dla innych prac prowadzonych w Zespole Inżynierii Lingwistycznej IPI PAN, dlatego podjęto także próby zastosowania metod automatycznych do generowania słownika (Dębowski i Woliński 2007; Dębowski 2009; Hajnicz 2011). W końcu jednak zapadła decyzja o budowie metodą tradycyjną słownika walencyjnego Walenty (http://zil.ipipan.waw.pl/Walenty).

Słownik Walenty był od początku tworzony z myślą o zastosowaniach komputerowych, sprawą kluczową było więc opracowanie adekwatnego formalizmu zapisu słownika (Przepiórkowski *et al.* 2014c). Jednocześnie ma to być słownik czytelny dla człowieka, czemu służą różne postaci prezentacyjne. Interfejs pozwalający na przeglądanie słownika przez WWW jest dostępny pod adresem http://walenty.ipipan.waw.pl (Hajnicz *et al.* 2015). Przy budowie słownika przyjęto zasadę, że wszystkie reprezentowane zjawiska muszą mieć dokumentację w postaci cytatów korpusowych, które prezentowane są w interfejsie słownika włącznie ze wskazaniem, które typy argumentów są realizowane w danym przykładzie.

Zaczątkiem słownika Walenty stał się wspomniany słownik analizatora Świgra 2[1], w ramach prac został on jednak bardzo istotnie rozbudowany zarówno pod względem liczby opisywanych jednostek, jak i wielu innowacyjnych elementów zwiększających szczegółowość opisu – zostaną one omówione w następnych punktach tego rozdziału. Dotyczy to przede wszystkim opisu koordynacji argumentów, zjawiska kontroli składniowej, argumentów zleksykalizowanych (czasownikowych związków frazeologicznych). Bardzo istotne jest także to, że oprócz warstwy składniowej tworzona jest paralelna warstwa semantyczna wiążąca z argumentami role semantyczne (zob. p. 3.2). O ile słownik wyjściowy zawierał informację wyłącznie o czasownikach, w Walentym uwzględniono jako źródła oddziaływań czasowniki, rzeczowniki, przymiotniki i przysłówki[2]. Dokumentację formatu słownika i rozwoju prezentowanego opisu można znaleźć w pracach: Przepiórkowski *et al.* (2014c,b), Hajnicz *et al.*

[1] Udział autora niniejszej książki w pracach nad Walentym polegał na dostarczeniu owego zaczątka i uczestnictwie w pracach koncepcyjnych. Pracami leksykograficznymi kieruje Elżbieta Hajnicz.

[2] Warto przy tym zaznaczyć, że w wypadku nieczasowników hasła słownikowe tworzone są jedynie dla jednostek konotujących frazy nietypowe dla danej klasy gramatycznej (por. p. 2.8.1). Tak więc rzeczownik WĄTPLIWOŚĆ wymaga opisu ze względu na możliwą przy nim obecność frazy zdaniowej (*wątpliwość, czy to zrobić,*), podczas gdy obecność przy nadrzędniku rzeczownikowym podrzędnika przymiotnikowego nie byłaby wystarczającym powodem do uwzględnienia danego leksemu w słowniku.

(2016b), Przepiórkowski *et al.* (2014a), Hajnicz *et al.* (2016a), Patejuk i Przepiórkowski (2015, 2014) i Przepiórkowski *et al.* (2017).

Walenty jest obecnie największym i najbardziej szczegółowym publicznie dostępnym słownikiem walencyjnym języka polskiego. Bieżąca wersja zawiera około 13 000 czasowników, 3000 rzeczowników, 1000 przymiotników i 200 przysłówków. Jeżeli uwzględnić frekwencję występowania czasowników w tekstach, odpowiada to opisaniu walencji dla 99,8% form czasownikowych (oszacowanie na podstawie informacji frekwencyjnej z automatycznie tagowanego zrównoważonego korpusu NKJP300). Słownik jest udostępniany na otwartej licencji, może to być w zależności od preferencji użytkownika licencja Creative Commons Attribution Share-Alike 4.0 lub GNU General Public License wersja 3.

3.1. Warstwa składniowa

Hasła słownika Walenty odpowiadają leksemom analizatora fleksyjnego Morfeusz. W szczególności hasła nie zawierają nigdy cząstki *się*, a więc schematy dla „użyć zwrotnych" czasowników są zapisywane w obrębie tej samej jednostki słownika, co dla użyć bez *się*[3]. Hasła czasownikowe są dzielone na podhasła ze względu na zwrotność, czyli użycie z partykułą SIĘ (przy czym chodzi o *się* inherentne, jak w *bać się*, *zwracać (się)*, zob. p. 3.1.6), zanegowanie, czyli użycie z partykułą NIE, i aspekt. Hasła przymiotnikowe i przysłówkowe są dzielone ze względu na predykatywność (zob. p. 3.1.11, s. 157).

Każde podhasło jest zbiorem schematów składniowych. Schemat składniowy jest zbiorem pozycji składniowych, które mogą być wypełniane przez frazy wymagane określonych typów. W warstwie składniowej słownika frazy wymagane są identyfikowane powierzchniowoskładniowo, a więc operuje się typami fraz oraz przysługującymi im cechami gramatycznymi. W wypadku wymagań zleksykalizowanych podawany jest lemat formy stanowiącej centrum frazy.

Dla przykładu hasło słownika Walenty opisujące czasownik OBIECAĆ składa się z jednego podhasła o cechach: bez *się*, niezależny od negacji, dokonany. Zawarty w tym podhaśle schemat składniowy użyty w zdaniu (2) składa się z trzech pozycji składniowych. Są to: pozycja podmiotu dopuszczająca frazę nominalną w mianowniku, pozycja dopuszczająca frazę werbalną bezokolicznikową w dowolnym aspekcie i pozycja dopuszczająca frazę nominalną w celowniku (*obiecać komuś*, niezrealizowana w zdaniu (2)). Formalny sposób zapisu schematów zostanie objaśniony w kolejnych punktach.

[3] W obecnej wersji słownika nie są również używane zapisy dezambiguujące lematy, co uniemożliwia rozdzielenie schematów dla czasowników z homonicznymi formami bezokolicznika (np. SŁAĆ – *ściele* albo *śle*). Wprowadzenie lematów ujednoznacznionych jest planowane w jednej z przyszłych wersji.

3.1.1. Które podrzędniki notować w słowniku?

W języku polskim podrzędniki czasownika co do zasady nie są obligatoryjne. Jest to wyraźna różnica w stosunku do języka angielskiego, w którym brak dopełnienia przy czasowniku przechodnim powoduje niegramatyczność zdania. W języku polskim takie zdania pozostają gramatyczne, pojawia się jedynie trudny do precyzyjnego określenia poziom eliptyczności (por. przykłady (7) i (8)). Zwykle daje się skonstruować kontekst wypowiedzi, który by uwiarygadniał pojawienie się danej konstrukcji w tekście.

(7) *Who saw Mary? *John saw.*
(8) *Kto zobaczył Marysię? Jan zobaczył.*

Schematy notowane w słowniku należy więc rozumieć jako maksymalne: poszczególne ich elementy mogą, ale nie muszą zostać wspólnie zrealizowane w zdaniu. Wyjątek stanowią elementy nazwane zleksykalizowanymi służące do realizacji schematów frazeologicznych (zob. 3.1.13) oraz wymaganie sie (które można uznać za specyficzne wymaganie zleksykalizowane).

Z pragmatycznego punktu widzenia konieczne jest notowanie tych typów podrzędników, które mogą wystąpić nie przy wszystkich czasownikach. Są typy podrzędników, które ewidentnie noszą tę cechę, na przykład wspomniane frazy bezokolicznikowe. Jednak niektóre typy podrzędników, choć bardzo charakterystyczne dla pewnych czasowników, mogą też pojawić się przy innych czasownikach jako luźne. Na przykład frazy określające czas (*wczoraj*, *często*, *kiedy byłem mały*) są w słowniku przewidziane jako wymagane przy czasownikach takich jak: ZACZĄĆ, SKOŃCZYĆ, DEBIUTOWAĆ. Jest jednak wiele czasowników wyrażających czynność, która z natury jest osadzona w czasie, a mimo tego określenie czasu nie jest przy nich wymagane, np. czasownik CZYTAĆ. W wielu wypadkach decyzja, czy uwzględnić taki podrzędnik w schemacie składniowym czasownika, jest niejasna. W słowniku Walenty kwestia ta nie została poddana wyraźnym kryteriom, w pewnym stopniu pozostawiono ją intuicji leksykografów. Przyjęto, że w wypadkach wątpliwych lepiej jest uznać podrzędnik za argument i opisać go w słowniku.

3.1.2. Typy fraz

W GFJP typy fraz wymaganych są oznaczane za pomocą heterogenicznego zestawu symboli zawierającego oznaczenia przypadków, oznaczenia przypadków ze znakiem prim, oznaczenia przyimków i spójników oraz dodatkowe symbole, na przykład pz na oznaczenie frazy pytajnozależnej; obecność korelatu (zob. p. 2.5) jest notowana osobno (Świdziński 1992, §6.4.2–§6.4.8). Ponieważ zestaw ten jest dość nieczytelny, w analizatorze Świgra 1, czyli już na etapie implementacji GFJP, wprowadzono oznaczenia reprezentujące poszczególne informacje: typ frazy i jej dalszą charakterystykę zależną od typu (Woliński 2004, s. 78). Symbole typów fraz przyjęły postać termów złożonych: fraza bezoko-

licznikowa z czasownikiem niedokonanym to infp(nd), fraza nominalna w celowniku to np(cel), fraza przymiotnikowa w celowniku to adjp(cel), a fraza przyimkowo-nominalna z przyimkiem NA to prepnp(na, bier) lub prepnp(na, miej), w zależności od wymaganego przypadka. W implementacji GFJP i tworzonym dla niej słowniku walencyjnym oznaczenie korelatu stało się integralną częścią symbolu typu wymaganej frazy zdaniowej (zob. także Bień 2009, p. 5.4).

Ten system oznaczeń został przejęty przez słownik Walenty ze zmianą polegającą na zastąpieniu symboli bardziej spójnym zestawem inspirowanym łacińskimi i angielskimi terminami gramatycznymi (w szczególności wartości kategorii gramatycznych są zgodne z łacińskimi symbolami używanymi w znacznikach fleksyjnych Morfeusza).

Lista typów fraz w słowniku Walenty przedstawia się następująco:

- np(*Przypadek*) – fraza nominalna w przypadku *Przypadek* (oznaczenia wprowadzone w p. 1.6.2), np. np(dat);
- adjp(*Przypadek*) – fraza przymiotnikowa w przypadku *Przypadek*, np. adjp(dat);
- prepnp(*Przyimek*, *Przypadek*) – fraza przyimkowo-nominalna z przyimkiem *Przyimek* (lemat ze słownika) wymagającym przypadka *Przypadek*, np. prepnp(na,acc);
- prepadjp(*Przyimek*, *Przypadek*) – fraza przyimkowo-przymiotnikowa, charakterystyka jw.;
- infp(*Aspekt*) – fraza werbalna w bezokoliczniku o zadanej wartości *Aspektu* (wartość _ sygnalizuje dopuszczalność dowolnego aspektu, co dotyczy większości schematów; zdarzają się jednak czasowniki dopuszczające jedynie frazy o ustalonym aspekcie, np. ZACZĄĆ – tylko niedokonany imperf);
- cp(*Tfz*) – fraza zdaniowa typu *Tfz* (oznaczeniem typu frazy zdaniowej może być lemat spójnika wprowadzającego taką frazę, wartość int oznaczająca frazę pytajnozależną oraz wartość rel specyfikująca frazę względną);
- ncp(*Przypadek*, *Tfz*) – fraza zdaniowa typu *Tfz* z korelatem w postaci formy leksemu TO w zadanym przypadku;
- prepncp(*Przyimek*, *Przypadek*, *Tfz*) – fraza zdaniowa typu *Tfz* z korelatem w postaci przyimka *Przyimek* i formy leksemu TO w zadanym przypadku;
- xp(*Sem*) – fraza motywowana semantycznie danego typu – zob. p. 3.1.5;
- advp(*Sem*) – fraza przysłówkowa danego typu – zob. p. 3.1.5;
- or – mowa niezależna (*oratio recta*, przytoczenie);
- comprepnp(*Ozn*) – fraza z przyimkiem złożonym o oznaczeniu *Ozn* – zob. p. 3.1.8;
- compar(*Typ*) – fraza porównawcza, zob. p. 3.1.9;
- distrp – fraza dystrybutywna, zob. p. 3.1.7;
- nonch – fraza „niechromatyczna", zob. p. 3.1.7;
- refl – partykuła *się* realizująca zwrotne użycie czasownika, zob. p. 3.1.6;
- recip – partykuła *się* realizująca wzajemne użycie czasownika, zob. p. 3.1.6;
- E – znacznik wskazujący możliwość podniesienia podmiotu, zob. p. 3.1.12.

Następujące dwa typy zostały wprowadzone w celu reprezentacji fraz zleksykalizowanych, zob. p. 3.1.13:

– lex(*Tfw*, *Cechy_gramatyczne*, *Lemat*, *Modyfikatory*) – zleksykalizowana fraza typu *Tfw*, której głową semantyczną jest forma leksemu o lemacie *Lemat*;
– fixed(*Tfw*, *Napis*) – frazeologizm o ustalonym brzmieniu pełniący funkcję składniową frazy typu *Tfw*.

Na potrzeby opisu fraz zleksykalizowanych konieczne było także uszczegółowienie pewnych typów fraz:

– nump(*Przypadek*) – fraza nominalna, której centrum jest liczebnik (o zadanym lemacie);
– ppasp(*Przypadek*) – fraza przymiotnikowa, której centrum jest forma konkretnego imiesłowu biernego;
– prepnump(*Przyimek*, *Przypadek*) – fraza przyimkowo-nominalna, której komponent nominalny ma jako centrum liczebnik;
– prepgerp(*Przyimek*, *Przypadek*) – jw. z odsłownikiem;
– prepppasp(*Przyimek*, *Przypadek*) – fraza przyimkowo-przymiotnikowa, której komponent przymiotnikowy ma jako centrum imiesłów bierny;
– qub – wymagana konkretna partykuła (kublik w terminologii NKJP).

3.1.3. Pozycje składniowe

Schematy składniowe w Walentym są definiowane jako zbiory pozycji składniowych. Dla każdej pozycji określa się typy fraz, które mogą ją realizować. Pozycje są wyróżniane na podstawie testu koordynacji (Szupryczyńska 1996), a więc jeżeli pewne typy argumentów mogą przy danym nadrzędniku zostać skoordynowane, zostają uznane za należące do jednej pozycji. Pozycje są reprezentowane w postaci zbiorów specyfikacji typów fraz wymaganych, które mogą daną pozycję wypełniać.

Na przykład zdania (9) i (10) pokazują, że przy czasowniku PYTAĆ możliwe jest skoordynowanie frazy przyimkowej prepnp(o, acc) i frazy zdaniowej pytajnozależnej cp(int). W zdaniu (11) nastąpiła koordynacja frazy przyimkowej prepnp(o, acc) z frazą zdaniową pytajnozależną z korelatem przyimkowym prepncp(o, acc, int).

(9) *Ja pytam* prepnp(o, acc) *o rodzaj znajomości* i cp(int) *czy te kontakty przekładały się na relacje bliższej znajomości, towarzyskiej, służbowej, biznesowej*? [NKJP300]

(10) *Zdarzało się, że czasem brał na kolana i pytał* prepnp(o, acc) *o wyniki w szkole* lub cp(int) *czy dziecko było grzeczne*. [NKJP300]

(11) *Surma uprzedził ją w kilku słowach, że pytają* prepnp(o, acc) *o śmierć stróża* i prepncp(o, acc, int) *o to, gdzie przebywa obecnie dziedzic*. [NKJP300]

Przykłady potwierdzają więc słuszność tego, że w Walentym przy czasowniku PYTAĆ wyróżniono pozycję dopuszczającą trzy typy fraz wymaganych:

(12)

prepnp(o, acc)
cp(int)
prepncp(o, acc, int)

Pozostałe dwie pozycje w tym schemacie to podmiot nominalny i fraza nominalna w dopełniaczu np(gen), obie nie dopuszczające koordynacji z innymi typami fraz. Łącznie można je zapisać następująco:

(13)

np(nom)	np(gen)	prepnp(o, acc) cp(int) prepncp(o, acc, int)

Może się wydawać, że jeżeli dopuszczalna jest fraza zdaniowa pytajno-zależna z korelatem prepncp(o, acc, int), to dopuszczalna jest również fraza zdaniowa bez korelatu oraz fraza przyimkowa – tak dzieje się w grupie czasowników oznaczających komunikowanie. Jednak kombinacje fraz, które mogą zostać połączone w konstrukcji współrzędnej, zależą od czasownika. Na przykład możliwa jest konstrukcja *chodzi mi o to, że...*, ale nie *chodzi mi, że...* Dlatego kombinacje fraz podlegających koordynacji zawsze notowane są jawnie.

W schematach Walentego dwie pozycje składniowe są opatrywane wyróżniającymi etykietami. Etykieta subj wskazuje pozycję podmiotu, rozumianą prymarnie jako mianownikowa pozycja argumentu nominalnego, który uzgadnia się co do osoby, liczby i rodzaju z finitywną formą czasownika. Jawne oznaczenie podmiotu jest potrzebne, ponieważ istnieją nieuzgadniające frazy nominalne w mianowniku. Na przykład czasownik NAZYWAĆ SIĘ wymaga dwóch fraz mianownikowych, z których tylko jedna uzgadnia się z czasownikiem.

Koordynacja argumentów może zdarzyć się również na pozycji podmiotowej. Na przykład w zdaniu (14) na pozycji podmiotu stoi fraza współrzędna zawierająca frazę nominalną np(nom) i frazę zdaniową z ŻE – cp(że). Naturalnie brzmiące przykłady tego rodzaju wydają się rzadkie, ale dużo mniej kontrowersyjna jest koordynacja np(nom) z frazą zdaniową z korelatem mianownikowym ncp(nom, że) i ncp(nom, int), por. zdanie (15).

(14) *– Cieszy* subj np(nom) *wygrana w meczu derbowym oraz* cp(że) *że zrewanżowaliśmy się rywalowi za jesienną przegraną – powiedział po meczu trener sanoczan Piotr Kot.* [NKJP1800]

(15) *Największe emocje budziło we mnie* subj np(nom) *karmienie dziecka piersią i* ncp(nom,int) *to, jak ono bardzo pragnie pokarmu, ciepła i miłości, jak się uspokaja i jest szczęśliwe – a ja mogę mu to wszystko dać*. [NKJP1800]

W związku z tym odpowiednie frazy zdaniowe są uznawane za podmiotowe, również gdy nie są skoordynowane z frazą nominalną:

(16) *I cieszy mnie* subj *, że skończył się czas wszystkich tych miłosnych uniesień*. [NKJP300]

W zdaniach realizujących takie schematy czasownik występuje w formie równokształtnej z formą 3 osoby liczby pojedynczej w rodzaju nijakim. Przy przyjętej interpretacji frazy zdaniowej jako podmiotu wypada przyjąć, że ma ona taką właśnie charakterystykę, którą narzuca czasownikowi. Tym samym fraza taka wypełnia rozluźnioną definicję podmiotu (por. p. 2.1.7) – uzgadnia się z formą czasownika.

Etykieta obj oznacza pozycję, która stanie się podmiotem w stronie biernej (por. p. 3.1.10). W wypadku większości czasowników jest ona realizowana przez frazę biernikową (np. dla CZYTAĆ, ROZGROMIĆ, ZAPROWADZIĆ), jednak na tej pozycji mogą także wystąpić frazy nominalne w narzędniku (np. dla KIEROWAĆ), w dopełniaczu (np. dla WYSŁUCHAĆ), a także pewne frazy zdaniowe (np. cp(że) przy czasowniku DOWIEŚĆ).

(17) *[...] usłużny ober zaprowadził ich do loży...*

(18) *Mock i jego dwaj towarzysze zostali zaprowadzeni do służbowej loży przez usłużnego obera, który całą pensję wydawał chyba na pomadę do wąsów.* [NKJP300]

(19) *Kosztowało nas to majątek, ale mąż kierował firmą poligraficzną i byliśmy ludźmi zamożnymi.* [NKJP300]

(20) *Firma poligraficzna jest formalnie kierowana przez męża, ale w rzeczywistości działa niezależnie.*

(21) *Podczas procesu nie zostało dowiedzione , że doszło do rozboju.*

Obecność w schemacie etykiety obj wskazuje w szczególności, że schemat ten może być użyty w konstrukcji biernej (takie schematy nazywa się pasywizowalnymi).

Oto zapis schematu dla czasownika KIEROWAĆ użytego w zdaniach (19) oraz (20):

(22)

subj	obj
np(nom)	np(inst) ncp(inst, int)

3.1.4. Przypadek strukturalny

Istotnym zjawiskiem, wymagającym uwzględnienia w słowniku walencyjnym, jest tzw. dopełniacz negacji. Jest to zjawisko polegające na tym, że przypadek frazy nominalnej na pewnej pozycji zależy od obecności w zdaniu negacji, mianowicie w zdaniu bez negacji jest to biernik, a przy obecności negacji – dopełniacz:

(23) *Piotr czyta książkę.*

(24) *Piotr nie czyta książki.*

Zjawisko to jest w tradycyjnych opisach polonistycznych interpretowane jako zmiana przypadka frazy pod wpływem negacji. Frazy takie są w słowniku notowane jako frazy nominalne w bierniku (np. Polański 1980–1992). Bardziej eleganckim rozwiązaniem jest zaproponowane w teoriach generatywistycznych uznanie, że przypadek frazy realizującej daną pozycję jest strukturalny, przez co rozumie się, że pozycja może być realizowana przez frazy w różnych przypadkach w zależności od pewnych uwarunkowań. Rozwiązanie to przyjęto w słowniku Walenty (Przepiórkowski *et al.* 2014c, s. 160) oznaczając takie frazy np(str).

Komplikację w interpretacji tego symbolu powoduje to, że w Walentym przyjęto za Przepiórkowskim (2004b), że fraza z centrum liczebnikowym na pozycji podmiotu jest w bierniku, czyli że przypadek frazy nominalnej na pozycji podmiotu również jest strukturalny: jest mianownikiem dla fraz z centrum rzeczownikowym i biernikiem – dla fraz z centrum liczebnikowym. Ponieważ przyjęcie tej koncepcji ma niejasne konsekwencje teoretyczne, co najmniej wymaga zdefiniowania na nowo pojęcia przypadka w języku polskim (zob. Saloni 2005), w niniejszej pracy przyjęto, że przypadek frazy nominalnej na pozycji podmiotowej jest z definicji leksykalnym mianownikiem oznaczanym np(nom)[4].

W związku z tym w niniejszej pracy za strukturalny uznawany jest jedynie przypadek frazy podlegającej zjawisku dopełniacza negacji. Dla podkreślenia zakresu niejednoznaczności używane jest oznaczenie np(accgen) w schematach składniowych i drzewach generowanych przez analizator Świgra 2.

Warto wspomnieć, że system przyjęty w słowniku Walenty nie umożliwia opisania nielicznych czasowników, które oprócz podmiotu dopuszczają nieuzgadniającą frazę mianownikową (Przepiórkowski *et al.* 2014c, s. 159), np.:

(25) *Moja bohaterka* nazywa się *Pyszny, Aleksandra Pyszny*. [Walenty/NKJP1800]

(26) *Zagadnięty o Wałęsę też krzyczy: – Wałęsa jest debil!* [Walenty/NKJP1800]

(27) *Ja mam na nazwisko Stefaniak*. [Walenty/NKJP300]

(28) *Pani uważa, że śmierć Dunina to było morderstwo...?* [NKJP300]

Według zasad Walentego nieuzgadniająca fraza mianownikowa powinna być bowiem oznaczona np(str) ze względu na możliwość realizacji liczebnikowej. Jednak ten symbol na pozycji niepodmiotowej oznacza frazę, która może mieć jako wartość przypadka biernik lub dopełniacz, a nie mianownik. Poprawna interpretacja tych przykładów byłaby możliwa, gdyby uznać, że wpływ konkretnego czasownika należy do czynników ograniczających możliwe realizacje

[4] W słowniku Walenty przypadkiem strukturalnym ma być wyjaśniana również realizacja tej frazy przez np(gen), gdy centrum konstrukcji nie jest forma finitywna, ale odsłownik (Hajnicz *et al.* 2016b, s. 74). Jest to jednak wyjaśnienie niesatysfakcjonujące, bowiem w tym wypadku pozycja może być realizowana przez np(gen) albo prepnp(przez, acc). Tej drugiej możliwości nie da się wyjaśnić przez przypadek strukturalny, jako że nie jest to jedynie wybór wartości przypadka. Potrzebny jest więc bardziej ogólny mechanizm, aby wyjaśnić obie możliwe realizacje argumentu przy gerundium.

przypadka strukturalnego, a więc, że istnieją dwa różne konteksty dopuszczające przypadek strukturalny wymagające rozróżnienia słownikowego. Można by je oznaczyć strnom dla kontekstu typowo realizowanego przez mianownik i stracc dla kontekstu typowo realizowanego przez biernik. Zwiększyłoby to też czytelność słownika. Problem ten nie dotyczy prezentowanego tu opisu, w którym frazy w omawianych zdaniach są oznaczone np(nom), w odróżnieniu od fraz np(accgen).

3.1.5. Wymagane frazy motywowane semantycznie

W analizatorze Świgra przed wdrożeniem słownika Walenty przyjmowano, że wymaganie frazy przysłówkowej advp może zostać również zrealizowane przez frazy przyimkowe. Uzasadnieniem tej koncepcji jest fakt, że argumenty wyrażane przysłówkami, np. określenia miejsca, często mogą być również wyrażone pewnymi frazami przyimkowo-nominalnymi:

(29) *Jan pochodził stąd / znikąd / z Warszawy / spod Cieszyna / spoza województwa.*
(30) *Poszedł dokądś / tam / do miasta / nad Świder / pod most.*
(31) *Posiedzenie trwało długo / przez cały dzień / do wieczora.*

Mechanizm ten był nieprecyzyjny. Przede wszystkim wymaganie frazy przysłówkowej jest zbyt ogólne, ponieważ zwykle na danej pozycji dopuszczalna jest tylko ograniczona grupa przysłówków. Ponadto ograniczone są typy dopuszczalnych fraz przyimkowych, które korelują z typami dopuszczalnych przysłówków. Wyrazistymi przykładami są frazy nazwane w Walentym adlatywnymi (xp(adl)) i ablatywnymi (xp(abl)). Pierwsze określają punkt docelowy ruchu (por. (30)), a drugie – punkt startowy (por. (29)). Dzięki tym oznaczeniom można wyrazić, że tzw. czasowniki ruchu otwierają pozycje dla dwóch fraz „przysłówkowych", ale jedna z nich jest ablatywna, a druga adlatywna.

Rozróżnienie to wyrażono w słowniku Walenty za pomocą typu fraz xp z podtypami. Dla każdego podtypu zdefiniowano, jakie przysłówki i jakie frazy przyimkowo-nominalne mogą go realizować. Niektóre podtypy mogą także być realizowane przez pewne frazy zdaniowe, nominalne i przyimki złożone. Na przykład typ xp(instr), wyrażający narzędzie, może być realizowany przez frazę nominalną w narzędniku np(inst) lub przez frazy z przyimkami złożonymi za pomocą, przy użyciu, przy pomocy.

Możliwe realizacje fraz danego typu mogą też być koordynowane, na przykład dla typu xp(adl):

(32) *Poszedł* *do domu lub na dworzec*.

W obecnej wersji słownika wyróżniono następujące typy fraz xp (za Przepiórkowski *et al.* 2014c):

– locat – miejsce, np. *być* *w sklepie*, *znajdować się* *przy drzwiach*, *pod mostem*, *nad rzeką*;

– abl – punkt początkowy, np. *wyprowadzić* *ze strefy wojny*;
– adl – punkt końcowy, np. *przywieźć* *do Zabrza/pod most/nad rzekę*;
– perl – trasa, np. *biec* *przez wieś*;
– temp – czas, np. *mieć miejsce* *wczoraj*;
– dur – długość trwania, np. *trwać* *dwie godziny*;
– mod – sposób, np. *traktować* *źle*, *zachowywać się* *jak dziecko*;
– caus – przyczyna, np. *drętwieć* *z przerażenia*/*na myśl o egzaminie*;
– dest – cel abstrakcyjny, np. *dążyć* *do sukcesu*/*na szczyt*;
– instr – narzędzie, np. *malować* *kredkami/za pomocą pędzla*.

Listy możliwych rozwinięć poszczególnych typów xp są częścią słownika, ale są zadane osobno od schematów. Uzupełnianie tych list wpływa na wszystkie schematy, które z nich korzystają. Można w tym widzieć zarówno zaletę – brak redundancji, jak i wadę – wszystkie wystąpienia muszą dopuszczać wszystkie realizacje. Dlatego w poszczególnych schematach można ograniczyć frazę tylko do jednej lub kilku wymienionych dopuszczalnych realizacji.

Oznaczenie advp (z podklasami) jest używane w Walentym, jeżeli wymagana fraza może być realizowana wyłącznie przez przysłówki.

3.1.6. Zwrotność

W słowniku Walenty rozróżnia się trzy funkcje składniowe partykuły *się*:
– *się* inherentne, np. BAĆ SIĘ, jego obecność jest notowana jako cecha podhasła;
– *się* zwrotne, które zwykle jest wymienne na biernikowe *siebie samego*, np. MYĆ (SIĘ) oznaczane refl (Przepiórkowski *et al.* 2014c, §7.1);
– *się* wzajemnościowe recip, które może współwystępować z formami *nawzajem* i *wzajemnie* (Hajnicz *et al.* 2016b, s. 73).

Oto przykłady ilustrujące każdą z wymienionych funkcji:

(33) *Julek boi się wygłupić, jego młodociana ambicja jest bardzo wrażliwa.* [NKJP300]
(34) *Anielka jakby nigdy nic ciągle czesała się [=siebie samą] wielkim czerwonym grzebieniem.* [NKJP300]
(35) *Nie darzyli się też wzajemnie ani sympatią, ani zaufaniem.* [NKJP300]

Te trzy *się* różnią się przypisanymi im rolami semantycznymi (w szczególności *się* inherentne nie ma osobnej roli), jednak w warstwie składniowej analizatora Świgra 2 wszystkie są traktowane tak samo, jako fraza wymagana typu sie realizowana przez partykułę *się*. W ten sposób wszystkie rodzaje *się* mają swobodną pozycję wystąpienia w zdaniu względem czasownika (również *się* inherentne, np. *On się tego bardzo bał.*).

Argumenty typu refl i recip można widzieć jako zleksykalizowane podtypy wymagania biernikowego np(accgen) – sygnalizują one, przy jakich czasowni-

kach to wymaganie może być realizowane przez *się*. W sposób zbliżony przy niektórych czasownikach są używane formy *sobie* i *siebie*. Takie wymagania są zapisywane w Walentym jako zleksykalizowane frazy nominalne w odpowiednim przypadku z zaimkiem SIEBIE (por. p. 1.7.4). Co do zasady notowane są w ten sposób użycia analogiczne do *się* inherentnego (np. *podchmielić sobie*, *wyobrażać sobie*).

Partykuła *się* może też pełnić w zdaniu funkcję podmiotu nominalnego:

(36) *O planach grupy biznesmenów z Łeby mówi się coraz głośniej.* [Skł.]

(37) *Jak się ją trenuje?* [Skł.]

Ten fakt nie musi jednak być notowany w słowniku, zakłada się, że każdy podmiot np(nom) może być realizowany przez *się*.

3.1.7. Specjalne wartości typów fraz

np(part)

Na miejscu przypadka frazy nominalnej może wystąpić oznaczenie part. Jest to sygnał, że w danym schemacie możliwy jest dopełniacz cząstkowy (*genetivus partitivus*). Oznacza to, że na danej pozycji dopuszczalna jest fraza nominalna w bierniku, może ona jednak być substytuowana frazą w dopełniaczu, jak w następujących przykładach (Saloni 2005, s. 44):

(38) *Weźcie wino.*

(39) *Weźcie wina.*

Obie wyróżnione frazy realizują tę samą pozycję składniową, drugi sposób realizacji wnosi naddatek znaczeniowy *trochę*. Możliwe są zdania, w których oba typy realizacji są skoordynowane:

(40) *Dajcie wina i całą świnię.* [Przepiórkowski 1999]

Zjawisko to jest notowane *explicite* w słowniku, ponieważ nie każdy czasownik dopuszczający frazę nominalną w bierniku dopuszcza dopełniacz cząstkowy.

Podobnie jak w wypadku frazy np(accgen) możliwe realizacje są zależne od negacji: przy obecności negacji możliwa jest tylko realizacja dopełniaczowa; przy jej braku – zarówno biernikowa, jak i dopełniaczowa.

adjp(pred)

Symbol pred wskazuje na nietypowe uwarunkowania przypadkowe fraz przymiotnikowych na pozycji predykatywnej, czyli wymaganych przez czasowniki takie jak: BYĆ, ZOSTAĆ, CZUĆ SIĘ, WYDAWAĆ SIĘ, PAMIĘTAĆ (Przepiórkowski *et al.* 2014c, s. 169). Przypadek takiej frazy albo uzgadnia się z innym argumentem, wskazanym opisaną dalej relacją kontroli składniowej, albo jest równy

narzędnikowi. W obu wypadkach oba argumenty uzgadniają się co do rodzaju i liczby.

Na przykład dla czasownika OKAZAĆ SIĘ argument adjp(pred) jest powiązany z podmiotem, więc może wystąpić w mianowniku lub w narzędniku – (41). W wypadku czasownika WIDYWAĆ powiązanie dotyczy argumentu np(accgen), więc możliwy jest biernik/dopełniacz (zależnie od negacji) lub narzędnik – (42) i (43).

(41) *Kierownik okazał się wredny/wrednym.*
(42) *Kierownika widywał pijanego/pijanym.*
(43) *Kierowniczki nie widywał pijanej/pijaną.*

Dodatkowa komplikacja występuje, jeżeli fraza nominalna ma jako nadrzędnik rządzącą rec formę liczebnika (zob. p. 2.8.3). W takiej sytuacji fraza przymiotnikowa może uzgodnić się zarówno ze składnikiem liczebnikowym, jak i z jego podrzędnikiem nominalnym:

(44) *Pięć osób okazało się wredne/wrednych/wrednymi.*
(45) *Pięć kobiet widywał pijane/pijanych/pijanymi.*

distrp

Typ frazy distrp został wprowadzony na oznaczenie fraz dystrybutywnych z przyimkiem PO. Oznaczenie to jest stosowane przy nielicznych czasownikach, np. SKŁADAĆ SIĘ, które wymagają frazy dystrybutywnej niewymiennej na frazę nominalną w bierniku; fraza dystrybutywna może bowiem typowo być realizacją frazy nominalnej np(accgen) (Hajnicz *et al.* 2016b, s. 77).

(46) *Mieszkańcy składają się po 2–3 tysiące złotych.* [NKJP300]
(47) *Zamierzają kupić 10 tys. gumowych kaczek w Chinach po 20 gr sztuka i odsprzedawać je po 3 zł chętnym do puszczenia własnej kaczki w wyścigu.*

Wymaganie tego typu może być realizowane przez frazę przyimkowo-nominalną ze składnikiem liczebnikowym w bierniku (*po dwa jabłka*); frazę z centrum nominalnym w miejscowniku liczby pojedynczej (*po jabłku*) lub też mnogiej w wypadku rzeczowników *plurale tantum* (*po skrzypcach dla każdego*).

prepadjp(po, postp)

Symbol postp stojący na miejscu przypadka w specyfikacji frazy przyimkowo-przymiotnikowej z PO sygnalizuje, że składnikiem przymiotnikowym tej frazy musi być forma poprzyimkowa adjp:dat przymiotnika typu *polsku, niemiecku*. Frazy takie są wymagane przez czasowniki typu MÓWIĆ, CZYTAĆ, POROZUMIEWAĆ SIĘ.

cp(żeby2)

Symbol żeby2 oznacza typ frazy zdaniowej, której realizacja jest zależna od obecności w zdaniu negacji. Mianowicie fraza ta może być zawsze realizowana przez frazę zdaniową wprowadzoną spójnikiem ŻE, a w zdaniu z negacją – również spójnikiem ŻEBY, np.:

(48) *Sądzili, że wrócą późno w nocy.* [NKJP300]
(49) *Chyba nie sądzisz, że wstąpię do paczki, w której dziewczyna będzie szefem?* [NKJP300]
(50) *Nie sądzę, żeby ktokolwiek mógł się na to zgodzić.* [NKJP300]

Symbol żeby2 występuje również we frazach z korelatem.

advp(pron)

Oznaczenie to sygnalizuje możliwość realizacji argumentu przez przysłówki TAK i JAK niewymienne na inne frazy przysłówkowe. Argument taki zwykle „zastępuje" frazę zdaniową występującą w innych schematach dla danego czasownika (argument advp(pron) jest notowany w osobnym schemacie, bo nie może się koordynować z innymi frazami). Dzieje się tak dla czasowników takich jak DECYDOWAĆ, SĄDZIĆ, TWIERDZIĆ, UWAŻAĆ:

(51) *Uważam, że za szybko pozbywamy się takich parowozów.* [NKJP1800]
(52) *Tak uważam.*
(53) *Jak uważasz?*

nonch

Argument oznaczany nonch, nazwany frazą niechromatyczną (Przepiórkowski *et al.* 2014c, s. 168), może być realizowany przez wyrażenia: *co*, *coś*, *nic*, *to*, *to samo*, *dużo*, *wiele*, *niewiele*, ale nie przez bardziej rozbudowane frazy nominalne. Jest notowany na przykład przy czasownikach TWIERDZIĆ, BZDURZYĆ, PRZEBĄKIWAĆ:

(54) *Nieśmiało przebąkują coś, że być może oznacza to późną i mroźną zimę.* [NKJP1800]
(55) *Jeśli ktoś z tu obecnych będzie kiedykolwiek twierdził coś innego, nie słuchaj, to tylko i wyłącznie zazdrość w czystej postaci.* [NKJP300]
(56) *Mój mąż do tej pory niczego nie twierdził o Bogu.* [NKJP300]

Argument ten podlega dopełniaczowi negacji (por. (56)).

3.1.8. Przyimki złożone

Pojęcie przyimków złożonych zostało w Walentym wprowadzone do opisu konstrukcji takich jak *w stronę*. Konstrukcje takie zachowują się podobnie

do przyimków w tym sensie, że wymagają frazy nominalnej w konkretnym przypadku (dopełniaczu). Gdy to wymaganie zostanie spełnione, całość może realizować wymaganie innej jednostki. Na przykład konstrukcja z przyimkiem złożonym W STRONĘ może realizować wymaganie adlatywne xp(adl), podobnie jak konstrukcja z przyimkiem KU (z bardzo podobną interpretacją semantyczną). Samo wyrażenie *w stronę* jest niepełne. Typowo pozycja otwierana przez przyimek złożony może być wypełniona przez podrzędnik rzeczownikowy w dopełniaczu (57) lub przez uzgadniający podrzędnik przymiotnikowy (58).

(57) *Poszedł w stronę domu / ku domowi.*
(58) *Poszedł w tamtą stronę.*
(59) **Poszedł w stronę.*

Przyimki złożone najczęściej składają się z przyimka i formy rzeczownikowej dopuszczającej dalszy podrzędnik nominalny (DO SPRAW, NA GRUNCIE, POD KIERUNKIEM), w niektórych jednak wymaganym podrzędnikiem jest kolejna fraza przyimkowo-nominalna (BEZ WZGLĘDU NA, W OPARCIU O). Niektóre przyimki złożone dopuszczają podrzędniki innych typów, w szczególności frazy zdaniowe, również z korelatem; przykładem może być przyimek złożony W SPRAWIE:

(60) *Podobno w sprawie, co dalej robić, trwają głębokie namysły z udziałem UPR.*
(61) *Nie mam nic do powiedzenia w sprawie, czy ktoś tu przenocuje czy nie.*
(62) *Podkreślił, że wnioski w sprawie tego, że taka ustawa powinna podlegać notyfikacji Komisji Europejskiej, były składane już za rządów Jarosława Kaczyńskiego.* [NKJP300]

Wymaganie przyimka złożonego ma w słowniku postać comprepnp(*Ozn*). Jedynym argumentem jest umowne oznaczenie (np. w sprawie). Podrzędniki wymagane przez dany przyimek złożony są dokładnie wyspecyfikowane za pomocą tego samego mechanizmu, który służy do opisu wymagań zleksykalizowanych (zob. p. 3.1.13). Przyimki złożone można więc uważać za skrót notacyjny dla specyfikacji zleksykalizowanych wymagań stanowiących ich definicje.

Konstrukcje z przyimkami złożonymi są traktowane jako regularne, tak więc na przykład *w stronę domu* to fraza przyimkowa z przyimkiem *w* i frazą nominalną *stronę domu*, złożoną z nadrzędnika *stronę* i podrzędnika *domu* w dopełniaczu. Odpowiedni mechanizm gramatyki wykrywa jednak, że taka konstrukcja jest zgodna z definicją przyimka złożonego i może w związku z tym stanowić realizację wymagania comprepnp(w stronę).

W tym miejscu warto jeszcze wspomnieć, że konstrukcja CO DO jest traktowana jako zwykły (choć dwuczłonowy) przyimek konstytuujący typ frazy prepnp(co do,gen). Dokładniejsze zasady uznawania konstrukcji za przyimki złożone można znaleźć w artykule Hajnicz *et al.* (2016b, §3.3, s. 77).

3.1.9. Konstrukcje porównawcze

Symbol compar(*Typ*) jest w słowniku Walenty stosowany na oznaczenie konstrukcji porównawczych (Hajnicz *et al.* 2016b, §3.4, s. 79). *Typ* może mieć wartość jak, oznaczającą konstrukcje ze spójnikiem JAK lub NICZYM; niż – oznaczającą konstrukcje z NIŻ, NIŹLI lub ANIŻELI; wreszcie jako – z JAKO, choć te ostatnie w istocie nie mają charakteru porównania. Frazy tego typu doczekały się ostatnio opracowań analitycznych (Wróblewska i Wieczorek 2018; Patejuk 2017).

Omawiane konstrukcje mogą być realizowane w zdaniu w różny sposób w zależności od struktury zdania, w którym wystąpiły:

(63) *Wolę czytać niż pisać.*
(64) *Wolę czytać z Piotrem niż w samotności.*
(65) *Wolę kawę z Piotrem niż czytać w samotności.*
(66) *Wolę czytać książki w wannie niż gazety przy śniadaniu.*
(67) *Jan traktuje siostrę jak lekarz pacjentów.*

Ciekawe są szczególnie przykłady (66) i (67), w których fragment wprowadzany przez NIŻ lub JAK jest ciągiem fraz, a nie pojedynczą frazą. Można twierdzić, że jest to ciąg podrzędników (wymaganych i luźnych) właściwych dla czasownika, przy którym wystąpiła dana konstrukcja porównawcza.

Frazy porównawcze z niż są możliwe przy każdym przymiotniku w stopniu wyższym, Walenty notuje zaś kilka przymiotników w stopniu równym dopuszczających taki podrzędnik, np. PRZECIWNY i ODWROTNY.

3.1.10. Realizacje niefinitywne schematów czasownikowych

W wypadku czasowników schematy podawane są dla form finitywnych. Zakłada się, że sposób ich realizacji dla form niefinitywnych (imiesłowów przysłówkowych i przymiotnikowych, odsłownika) jest jednoznacznie określony niezależnie od czasownika.

Przy realizacji schematu jako wymagań formy imiesłowu biernego (a więc w szczególności w stronie biernej) pozycja opisana obj nie może być realizowana, natomiast fraza o typie np(nom) jest na pozycji subj realizowana jako prepnp(przez, acc), a ncp(nom, *Typ*) – prepncp(przez, nom, *Typ*). Frazy zdaniowe bez korelatu cp(*Typ*) można również uznać za realizowane przez prepncp(przez, nom, *Typ*), jednak w obecnej wersji Walentego nie warto tego robić, jako że podmiotowe cp(*Typ*) w schematach pasywizowalnych zawsze współwystępują z ncp(nom, *Typ*).

W stronie biernej specyfikacja pozycji obj staje się specyfikacją podmiotu dla nadrzędnego czasownika BYĆ/ZOSTAĆ. Na pozycji obj najczęściej stoi fraza nominalna w przypadku strukturalnym np(accgen), ale możliwe są też frazy nominalne w narzędniku np(inst), przykłady (68) i (69); dopełniaczu; partytyw-

ne; frazy niechromatyczne, (70) i (71); frazy zdaniowe z korelatem w wymienionych przypadkach; niektóre frazy zdaniowe bez korelatu, (72) i (73), (74) i (75), (76) i (77) oraz frazy bezokolicznikowe, (78) i (79). Przy realizacji schematu w stronie biernej pozycja oznaczona jako obj nie jest realizowana przy imiesłowie biernym, jednak specyfikacja typu frazy w niej określona determinuje postać podmiotu nadrzędnego czasownika w formie osobowej. Wszelkie frazy nominalne w tej pozycji odpowiadają podmiotowi mianownikowemu, frazy zdaniowe bez korelatu zachowują się bez zmiany, frazy zdaniowe z korelatem przechodzą we frazę zdaniową z korelatem mianownikowym.

(68) *Matusiński administrował* obj(np(inst)) *kamieniczką bezinteresownie [...].* [NKJP300]

(69) subj(np(nom)) *Węzeł z przyległymi liniami był znów administrowany przez DOKP Warszawa.* [NKJP300]

(70) *Czy wspominał* obj(nonch) *coś o dyktafonie?* [NKJP300]

(71) *– Było* subj(nonch) *coś wspominane o progu, ale ja nie wiedziałam, co to jest ten próg [...].* [NKJP300]

(72) *Podczas dyskusji akcentowali* obj(cp(int)) *, jak wielką rolę mają do odegrania właśnie młodzieżowi liderzy.* [NKJP1800/Walenty]

(73) *Podczas dyskusji zostało zaakcentowane* subj(cp(int)) *, jak wielką rolę mają do odegrania.*

(74) *Pyta Pani, czy akceptuję* obj(cp(że)) *, że jestem kobietą?*

(75) *Ogólnie akceptowane jest* subj(cp(że)) *, że Broceach była chrześcijanką, zaś Dutbach został prawdopodobnie nawrócony w późniejszych latach.* [Walenty]

(76) *Bóg nakazał Jozuemu* obj(cp(żeby)) *, żeby ich wytępił i przejął ich ziemie.* [NKJP1800]

(77) *Jozuemu zostało nakazane* subj(cp(żeby)) *, żeby ich wytępił.*

(78) *Sąd nakazał mi* obj(infp) *zapłacić tysiąc złotych – mówi Janina Szyguła.* [NKJP1800]

(79) *Zostało im nakazane* subj(infp) *zapłacić.*

Można argumentować, że w przytoczonych zdaniach w stronie biernej frazy zdaniowe i bezokolicznikowe stoją na pozycji podmiotu, ponieważ ich obecność wpływa na dopuszczalną postać czasownika ZOSTAĆ: ogranicza ją do formy nijakiej pojedynczej.

Warto zaznaczyć, że frazy zdaniowe w pozycji obj, choć postulowane chętnie w słowniku Walenty, są trudne do znalezienia w tekstach i rzadko obecne w przykładach w słowniku Walenty[5]. Dlatego w analizatorze Świgra 2 nie uwzględniono możliwości realizacji fraz zdaniowych w pozycji obj.

[5] Opinia ta nie jest wynikiem badania frekwencyjnego. Jednak do znalezienia wymienionych przykładów nie wystarczył 300-milionowy NKJP, konieczne było odwołanie się do korpusu 2-miliardowego, a i w nim trafiało się na jeden lub dwa przykłady.

Jeżeli pozycja obj ma realizację zleksykalizowaną, to ograniczenie również przechodzi na podmiot czasownika nadrzędnego. Na przykład Walenty zawiera schemat dla wyrażenia *zwrócić komuś uwagę, że...*, który przewiduje na pozycji obj frazę z rzeczownikiem UWAGA. W związku z tym w konstrukcji biernej podmiotem czasownika nadrzędnego musi być UWAGA, co więcej, wynika z tego, że realizująca ten schemat fraza z imiesłowem od czasownika ZWRÓCIĆ nie może być podrzędnikiem dla innych rzeczowników niż UWAGA.

(80) *Możesz to oczywiście robić, jeżeli sobie tego życzysz, ale licz się z tym, proszę, że zostanie Ci zwrócona uwaga, że takie zachowanie jest postrzegane jako nieeleganckie.*

Fraza nominalna na pozycji subj jest w konstrukcji biernej realizowana przez frazę typu prepnp(przez,acc). W wypadku fraz zdaniowych z korelatem mechanizm jest analogiczny: fraza z korelatem staje się podrzędnikiem przyimka PRZEZ. Na przykład w zdaniu (81) podmiot jest realizowany przez frazę typu ncp(nom,int), czyli frazę pytajnozależną z korelatem mianownikowym. W konstrukcji biernej (82) przypadek korelatu zmienia się na biernik w wyniku oddziaływania przyimka. Warto zauważyć, że bardziej naturalnie brzmiące zdanie (83) stanowi bierną realizację innego schematu czasownika (z ncp(inst,int)). W bieżącej wersji słownika Walenty notowane są cztery czasowniki ze schematami pasywizowalnymi z frazą zdaniową bez korelatu w pozycji subj: SPRAWIAĆ, SPRAWIĆ (subj(cp(że))), ZASKAKIWAĆ, ZASKOCZYĆ (subj(cp(int);cp(że))). Nie jest jasne, czy taki podmiot może wystąpić w konstrukcji biernej, jeżeli tak to zapewne z korelatem (por. (84) i (85)).

(81) *Do współpracy zniechęciło ją to, jak ją potraktował.*

(82) *Do współpracy została skutecznie zniechęcona przez to, jak ją potraktował.*

(83) *Do współpracy została skutecznie zniechęcona tym, jak ją potraktował.*

(84) *Zaskoczyło mnie, że zna moje imię.* [NKJP300]

(85) ?*Zostałem zaskoczony przez to, że zna moje imię.*

Subtelności te nie są uwzględnione w bieżącej wersji analizatora Świgra 2.

W wypadku realizacji schematu przy odsłowniku fraza np(nom) na pozycji subj może być zrealizowana przez np(gen) albo prepnp(przez, acc):

(86) *Bardzo mnie męczy Piotra ciągłe słuchanie muzyki.*

(87) *Bardzo mnie męczy ciągłe słuchanie muzyki przez Piotra.*

Nie przy każdym odsłowniku obie możliwości mogą być realizowane, co jednak jest trudne do precyzyjnego ujęcia, obecna wersja Walentego nie notuje takiej informacji. Frazy biernikowe są realizowane przy gerundium w dopełniaczu, tak więc np(accgen) może być realizowane jedynie przez np(gen); fraza z korelatem ncp(accgen, *Typ*) – przez ncp(gen, *Typ*); fraza partytywna np(part) – jedynie przez np(gen).

Jeżeli schemat jest realizowany przy formie bezokolicznikowej, bezosobniku, imiesłowie przysłówkowym lub imiesłowie przymiotnikowym czynnym, pozycja oznaczona subj nie może zostać zrealizowana w wypowiedzeniu. Jej specyfikacja może być jednak istotna w wypadku podnoszenia podmiotu, zob. p. 3.1.12.

3.1.11. Hasła nieczasownikowe w słowniku

W słowniku Walenty oprócz czasowników uwzględniono również rzeczowniki, przymiotniki i przysłówki posiadające wymagania walencyjne. Następujące zdania zawierają przykłady form rzeczownikowych wiążących się z podrzędnikami nietypowymi dla większości rzeczowników, mianowicie z frazami zdaniowymi:

(88) *Ich twarda miłość była dla mnie bodźcem, żeby to rzucić.* [NKJP300]

(89) *Obowiązujący aksjomat, jakoby demokracja była celem samym w sobie, sprawia właśnie, że głosować idzie byle kto na byle kogo.* [Walenty]

(90) *Wasza argumentacja, że wegetarianizm to choroba psychiczna, to żenada.* [NKJP300]

(91) *Moje aspiracje, by grać w wyższej lidze nie przeminęły.* [NKJP1800]

Rzeczowniki są najliczniejszą klasą gramatyczną (ok. 80 000 leksemów w SGJP). Dlatego nie było możliwe uwzględnienie ich w słowniku na zasadzie frekwencyjnej. Przyjęto inną zasadę (por. Hajnicz *et al.* 2016b, §4.1, s. 81): opisywane są rzeczowniki, których walencja nie jest zbieżna z typowymi modyfikatorami rzeczowników, czyli które wymagają czegoś oprócz fraz nominalnych w dopełniaczu i uzgodnionych fraz przymiotnikowych. Szczególną uwagę poświęcono rzeczownikom powiązanym derywacyjnie z czasownikami (np. INFORMACJA od INFORMOWAĆ). Można powiedzieć, że walencja jest opisywana dla tych rzeczowników, dla których jest to konieczne, aby uzyskać rozbiór składniowy.

Podobne zasady dotyczyły doboru haseł przymiotnikowych. Uwzględniono te, które wymagają jakiegoś argumentu niebędącego frazą przysłówkową lub przyimkowo-nominalną, ze szczególnym uwzględnieniem derywowanych od czasowników, a więc tzw. imiesłowów przeszłych typu OSIADŁY, quasi imiesłowów biernych ZADBANY i przymiotników zakończonych na *-alny* MIERZALNY, OSIĄGALNY (Hajnicz *et al.* 2016b, s. 83).

Opisano wszystkie przysłówki, dla których udało się zaobserwować jakieś wymagania.

Oznaczenie schematu przymiotnikowego lub przysłówkowego jako predykatywnego oznacza, że schemat taki może być użyty, tylko jeżeli opisywana jednostka stoi w zdaniu przy czasowniku (w pozycji predykatywnej), a nie jest podrzędnikiem rzeczownika lub przymiotnika (pozycja atrybutywna) (Hajnicz *et al.* 2016b, s. 83). Przykładem może być schemat z frazą zdaniową dla przymiotnika OBOJĘTNY:

(92) *Jest państwu obojętne, czy ktoś popełni samobójstwo.* [za Hajnicz *et al.* 2016b]

Opis wymagań jednostek nieczasownikowych wymagał wprowadzenia kilku dodatkowych oznaczeń. Symbolem agr na miejscu jakiejś cechy gramatycznej (np. rodzaju lub przypadka) sygnalizowane jest uzgodnienie tej cechy z nadrzędnikiem. Tak więc wymagana fraza przymiotnikowa przy rzeczowniku opisywana jest adjp(agr), jako że uzgadnia ona przypadek z rzeczownikiem.

Typ frazy possp oznacza frazę dzierżawczą (Hajnicz *et al.* 2016b, s. 82). Może być ona realizowana przez tzw. zaimki dzierżawcze (*mój*, *czyjś*, wymienione w słowniku) i frazy nominalne w dopełniaczu (*nauczyciela*, *Marysi*). Frazy typu possp są wymagane przez rzeczowniki oraz czasownik BYĆ:

(93) *Europa jest nasza.* [za Hajnicz *et al.* 2016b]
(94) *Dom był dziadka.* [za Hajnicz *et al.* 2016b]

3.1.12. Kontrola składniowa

W słowniku Walenty uwzględniono zjawisko kontroli składniowej (Przepiórkowski *et al.* 2014c, §4). Klasycznym przykładem ilustrującym to zjawisko jest różnica między czasownikami OBIECAĆ i KAZAĆ.

(95) *Jan obiecał Marii czytać.*
(96) *Jan kazał Marii czytać.*

W wypadku zdania (95) osobą, która ma czytać, jest jednoznacznie Jan, w wypadku zdania (96) – Maria. Powstaje pytanie, do jakiego poziomu opisu należy to zjawisko. Z pewnością konieczne jest uwzględnienie go w opisie semantycznym (odpowiednie argumenty semantyczne obu predykatów są tożsame). Kontrola ma jednak również skutki składniowe w postaci uzgodnienia rodzaju i liczby. Jeżeli bowiem fraza bezokolicznikowa ma podrzędnik przymiotnikowy, który uzgadnia się z podmiotem, jak dla czasownika UDAWAĆ, zob. (97), to w kontekście kontroli bezokolicznika rodzaj gramatyczny i liczba podrzędnika przymiotnikowego muszą być tożsame z przysługującymi odpowiedniemu argumentowi czasownika nadrzędnego, por. (98) i (99).

(97) *Piotr udaje miłego/*miłą/*miłe.*
(98) *Jan obiecał Marii udawać miłego/*miłą/*miłych.*
(99) *Jan kazał Marii udawać miłą/*miłego/*miłych.*

Według teorii generatywistycznych niewyrażony podmiot bezokolicznika jest tożsamy (dzielony) z odpowiednim argumentem czasownika nadrzędnego. Takie objaśnienie wyjaśnia tożsamość wartości rodzaju i liczby, powoduje jednak, że w strukturze reprezentującej frazę bezokolicznikową pojawia się niewyrażony na powierzchni argument podmiotowy wyposażony w komplet swoich cech, a więc w szczególności przypadek. Na przykład w strukturze przypisywanej przez gramatykę POLFIE zdaniu (98) predykat UDAWAĆ ma podmiot

niosący wartość przypadka równą celownikowi. W prezentowanym tu opisie nie uznaje się istnienia podmiotu dla frazy bezokolicznikowej, jednak zjawisko kontroli daje się opisać w zakresie jego konsekwencji co do uzgodnień poprzez wymaganie, aby rodzaj i liczba gramatyczna przypisane frazie bezokolicznikowej były równe atrybutom odpowiedniego argumentu czasownika nadrzędnego.

Kontrola jest wykorzystywana nie tylko w odniesieniu do argumentów bezokolicznikowych. Na przykład występujące w przykładzie (97) uzgodnienie między podmiotem nominalnym a argumentem przymiotnikowym jest sygnalizowane za pomocą relacji kontroli między tymi argumentami.

W słowniku Walenty kontrola składniowa jest stosowana do oznaczania:

– który argument czasownika kontroluje „podmiot" frazy bezokolicznikowej (powoduje uzgodnienie liczby i rodzaju);
– istnienia powiązania (koreferencji) między argumentem nominalnym a argumentem przymiotnikowym jak w zdaniu (97) (powoduje uzgodnienie liczby i rodzaju odpowiednich argumentów);
– istnienia powiązania frazy przyimkowo-przymiotnikowej z innym argumentem (gdy kontrolerem jest fraza nominalna, następuje uzgodnienie liczby i rodzaju; gdy kontrolerem jest fraza bezokolicznikowa – uzgodnienie domyślne do rodzaju nijakiego pojedynczego):

(100) *Prokurator uznał kwestię za zamkniętą/*zamknięty/*zamknięte.*
(101) *Prokuratura uznała za stosowne/*stosowny/*stosowną zająć się tym problemem.*

UZNAĆ jest przykładem czasownika z dwiema relacjami kontroli: podmiot kontroluje frazę bezokolicznikową, a ona kontroluje frazę przyimkowo-przymiotnikową (por. Przepiórkowski *et al.* 2014c, s. 163);

– istnienia analogicznego powiązania dotyczącego argumentu przymiotnikowego oznaczonego jako predykatywny (adjp(pred)). W tym wypadku następuje uzgodnienie liczby i rodzaju, natomiast przypadek frazy przymiotnikowej zależy od postaci frazy kontrolującej (realizującej pozycję subj albo obj):

(102) *Chłopiec urodził się owłosiony.*
(103) *Pięciu chłopców urodziło się owłosionych.*
(104) *Jan Marię zastał nieubraną.*
(105) *Jan Marii nie zastał nieubranej.*

Relacja kontroli jest oznaczana w słowniku Walenty za pomocą atrybutów controller i controllee (w schemacie może być jedna lub dwie pary takich atrybutów). Na przykład schemat dla czasownika OBIECAĆ ma postać następującą:

(106)

subj, controller	controllee	
np(nom)	infp(_)	np(dat)

Atrybuty te są przypisywane całym pozycjom składniowym, należy jednak rozumieć, że odnoszą się do tych realizacji pozycji, dla których mają sens.

Kontrola jest również wykorzystywana do opisu zjawiska tzw. podnoszenia podmiotu: czasowniki takie jak ZACZĄĆ i WYDAWAĆ SIĘ w schematach zawierających frazę bezokolicznikową dziedziczą specyfikację podmiotu od czasownika w pozycji bezokolicznika (w odróżnieniu na przykład od czasownika UWIELBIAĆ, który ma podmiot standardowy):

(107) *Jan zaczął czytać.*
(108) *Zaczęło dnieć.*
(109) **Jan zaczął dnieć.*
(110) *Że jest przepracowany, zaczęło męczyć Jana.*
(111) **Że jest przepracowany, zaczęło adorować Jana.*

W Walentym zjawisko to wyrażone jest w ten sposób, że podmiot czasownika ZACZĄĆ jest zasygnalizowany symbolem E, oznaczającym, że podmiot może zostać zrealizowany, tylko jeżeli zostanie „podniesiony" z argumentu oznaczonego jako controllee:

(112)

subj, controller	controllee
E	infp(imperf)

Mechanizm ten działa też w drugą stronę: frazy bezokolicznikowe z czasownikiem nie dopuszczającym podmiotu lub dopuszczającym tylko podmiot nienominalny mogą być podrzędnikiem jedynie czasowników podnoszących podmiot. Na przykład:

(113) *Zaczęło zmierzchać.*
(114) **Uwielbiało zmierzchać.*
(115) *Zaczęło mnie bawić, że nie umie zaparkować.*
(116) **Uwielbiało mnie bawić, że nie umie zaparkować.*

Podnoszony podmiot może być frazą zleksykalizowaną (por. 3.1.13):

(117) *Serce zaczyna mu krwawić z powodu tego zdarzenia.*
(118) **Serce uwielbia mu krwawić z powodu tego zdarzenia.*

3.1.13. Argumenty zleksykalizowane

Początkowo zapis elementów zleksykalizowanych w schematach składniowych został wprowadzony, aby uchwycić uwarunkowania polegające na tym, że obecność pewnego elementu leksykalnego uprawomocnia obecność innej pozycji schematu, czyli sytuacje, gdy ogólny schemat dla danego czasownika nie obejmował zdań z danym elementem leksykalnym. Przykładem może być konstrukcja *bić się z myślami*:

(119) *Joanna biła się z myślami, czy przyjść.*

(120) **Joanna biła się, czy przyjść.*
(121) **Joanna biła się z Piotrem, czy przyjść.*
(122) *Jan dostał rozkaz czytać.*
(123) **Jan dostał czytać.*

W przykładzie (119) obecność frazy *z myślami* uprawomocnia obecność frazy zdaniowej pytajnozależnej. Musi to być fraza przyimkowo-nominalna z rzeczownikiem MYŚL w liczbie mnogiej; frazy z innymi rzeczownikami nie pozwalają na obecność bezokolicznika, por. (121). Gdyby dopuścić realizację schematu bez tej frazy, możliwa byłaby analiza przykładu (120). Dlatego argumenty oznaczone jako zleksykalizowane są obowiązkowe.

W toku prac zasady uwzględniania elementów zleksykalizowanych ewoluowały. Obecna wersja zawiera schematy dodane ze względu na różnicę znaczeniową konstrukcji z elementem zleksykalizowanym, co pozwala poprawnie przypisać role semantyczne w warstwie semantycznej Walentego.

Sposób zapisu elementów zleksykalizowanych w schematach też jest wynikiem ewolucji (Przepiórkowski *et al.* 2014a). Początkowo, jako najczęstsze, dopuszczano tylko zleksykalizowane frazy nominalne i przyimkowo-nominalne z prostą sygnalizacją możliwości modyfikacji. Z czasem przyjęto jednak, że leksykalizacja może dotyczyć każdego typu frazy i znacząco rozbudowano sposób sygnalizowania możliwych podrzędników elementów zleksykalizowanych. Zapis elementu zleksykalizowanego ma postać: lex(*Tfw*, *Cechy_gramatyczne*, *Lemat*, *Modyfikatory*), gdzie *Tfw* jest dowolnym typem wymagania, a *Lemat* jest lematem formy stanowiącej centrum realizującej frazy. Podawane cechy gramatyczne zależą od typu frazy, na przykład dla fraz nominalnych i przyimkowo-nominalnych jest to liczba (w przykładzie (119) liczba musi mieć wartość pl, w odpowiednim znaczeniu można *bić się z myślami*, ale nie *z myślą*).

W najprostszym wypadku element *Modyfikatory* jest jednym z następujących symboli wskazujących, czy centrum frazy zleksykalizowanej może mieć podrzędniki: natr – podrzędniki niemożliwe, atr – podrzędniki możliwe, ratr – podrzędniki koniecznie obecne, atr1 i ratr1 – jeden podrzędnik dopuszczalny lub konieczny. W przykładzie (119) forma rzeczownika MYŚL nie może mieć podrzędników, więc pełna specyfikacja tego argumentu ma postać lex(prepnp(z, inst), pl, 'myśl', natr).

Prawdziwa siła mechanizmu leksykalizacji polega na tym, że wszystkie specyfikacje podrzędników z wyjątkiem natr mogą mieć argument w nawiasach o takiej samej postaci jak pełny schemat składniowy. Specyfikacja ta ogranicza dopuszczalne podrzędniki we frazie zleksykalizowanej. Na przykład schemat dla zdania (122) zawiera specyfikację lex(np(str),sg,'rozkaz',atr(adjp(agr))), która określa, że argument *rozkaz* może mieć podrzędnik w postaci uzgadniającej się frazy przymiotnikowej adjp(agr), ale żadnych innych. Możliwe jest więc *Dostał nieoczekiwany rozkaz czytać.*, ale nie **Dostał rozkaz do ataku czytać.*

W specyfikacji podrzędników mogą wystąpić kolejne frazy zleksykalizowane, mechanizm jest więc rekurencyjny. Oznacza to, że specyfikacje argumen-

tów zleksykalizowanych w Walentym mogą nakładać warunki na dowolnie zagłębione fragmenty poddrzewa składniowego. Na przykład dla frazeologizmu *witać z otwartymi ramionami* wprowadzono następujący opis frazy zleksykalizowanej:

(124) lex(prepnp(z, inst), pl, 'ramię',
ratr1(lex(adjp(agr), agr, agr, 'otwarty',
atr1(lex(advp(misc), 'szeroko', natr))))).

Oznacza on, że do zrealizowania tego schemat wymagana jest fraza przyimkowo-nominalna z centrum w postaci formy rzeczownika RAMIĘ w liczbie mnogiej. Podrzędnikiem tej formy musi (ratr1) być fraza przymiotnikowa o centrum OTWARTY. Ta z kolei forma może, ale nie musi (atr1), być modyfikowana formą przysłówka SZEROKO, który już nie ma podrzędników (natr).

Notacja wymagań w Walentym zawiera jeszcze kilka możliwości takich jak podawanie semantycznego typu frazy xp, ale ograniczonego do konkretnej realizacji, a dla fraz zleksykalizowanych podawanie alternatywnych *Lematów* w postaci listy. Szczegóły dotyczące tego zagadnienia można znaleźć w artykule Hajnicz *et al.* (2016b).

Niekiedy jednak wszystkie te mechanizmy nie wystarczają do zapisania nietypowego elementu zleksykalizowanego. Stosowana jest wtedy notacja fixed(*Tfw*, *Napis*), która sygnalizuje, że ustalony *Napis* może pełnić funkcję składniową *Tfw*. W taki sposób opisano na przykład wyrażenie *stanąć dęba*: element frazeologiczny jest tu oddany jako fixed(np(gen), 'dęba'), ponieważ segment *dęba* nie jest żadną współczesną formą rzeczownika.

Argumenty zleksykalizowane są rozpoznawane przez analizator Świgra 2 w sposób uproszczony: badana jest zgodność typu frazy oraz cech charakterystycznych zadanych dla tego argumentu, nie jest badana zgodność modyfikatorów ze specyfikacją.

Przyjęty sposób rozumienia argumentów zleksykalizowanych nastręcza pewnych problemów, gdy w schemacie jest więcej niż jedna fraza zleksykalizowana. Na przykład w schemacie dla wyrażenia *czuć do kogoś miętę przez rumianek* zarówno *miętę*, jak i *przez rumianek* oznaczone są jako zleksykalizowane. Oznacza to, że według przyjętych zasad oba argumenty są niepomijalne. Tymczasem adekwatny opis powinien podawać, że *miętę* jest niepomijalnym elementem konstytuującym frazeologizm, podczas gdy *przez rumianek* jest elementem opcjonalnym, choć o konkretnym wypełnieniu leksykalnym.

3.2. Warstwa semantyczna

Warstwa semantyczna słownika Walenty korzysta z szeroko rozumianego paradygmatu semantyki ram (ang. *frame semantics*). Jej celem jest możliwość budowy struktury predykatowo-argumentowej dla wypowiedzenia. Przegląd

innych słowników walencyjnych uwzględniających semantykę można znaleźć w pracy Hajnicz *et al.* (2016a).

Na poziomie semantycznym hasła słownika są dzielone na znaczenia, reprezentowane jednostką leksykalną Słowosieci lub zbiorem takich jednostek.

Słowosieć (Piasecki *et al.* 2009) jest wordnetem, czyli dużą leksykalną siecią semantyczną reprezentującą znaczenia poprzez wskazanie ich wzajemnych powiązań. Nośnikiem znaczenia w Słowosieci jest jednostka leksykalna. Jednostki leksykalne są powiązane z leksemami w sensie SGJP (stanowią ich możliwe znaczenia) i z wyrażeniami wieloczłonowymi, jak na przykład DAĆ NOGĘ, BIAŁY WALC.

Najważniejszymi typami uwzględnianych powiązań jest synonimia, zbierająca jednostki leksykalne w grupy o tym samym znaczeniu, nazywane synsetami (z założenia: używalne wymiennie w dowolnym kontekście) oraz hiperonimia/hiponimia łącząca z synsetem o bardziej ogólnym/bardziej szczegółowym znaczeniu. Inne relacje to antonimia (przeciwstawność), meronimia (bycie częścią), holonimia (składanie się z części) i inne włącznie z fuzzynimią, za pomocą której notowane jest intuicyjne poczucie, że między pojęciami zachodzi zależność, ale trudno określić jej charakter.

Z każdym znaczeniem jest w Walentym związana dokładnie jedna rama semantyczna. Rama semantyczna opisuje sytuację (determinowaną przez predykat, czyli powiązaną jednostkę Słowosieci), a jej elementy (argumenty semantyczne) reprezentują byty uczestniczące w sytuacji. Rama ma określać, w jaki sposób argumenty uczestniczą w sytuacji (określają to role semantyczne) oraz jakie jednostki mogą w danej sytuacji brać udział (preferencje selekcyjne). Tak więc każdy argument semantyczny jest parą ⟨rola semantyczna, preferencje selekcyjne⟩.

Istotną regułą opisu semantycznego w Walentym ma być niezależność od składniowych sposobów wyrażenia ról. Tak więc różne składniowo sposoby wyrażenia tej samej treści powinny być opatrywane tą samą reprezentacją semantyczną.

Układ powiązań danej ramy semantycznej ze schematami składniowymi można uważać za charakterystyczny dla danego znaczenia hasła. Wypowiedzenia zgodne ze schematem powiązanym z daną ramą semantyczną powinny dać się sparafrazować do postaci zgodnej z każdym innym schematem powiązanym z tą ramą. Jeżeli tak się nie dzieje, konieczny jest drobniejszy podział na znaczenia. Podział taki jest wprowadzany, nawet jeżeli oznacza to konieczność wprowadzenia jednostek leksykalnych niewystępujących (jeszcze) w Słowosieci.

Jednocześnie wiele jednostek leksykalnych może być powiązanych z tą samą ramą. Dzieje się tak na przykład dla jednostek leksykalnych reprezentujących użycia zwrotne i niezwrotne danego leksemu czasownikowego (ZBIĆ i ZBIĆ SIĘ). Przyjęto jednak, że wspólnie opisywane będą tylko jednostki leksykalne powiązane z tym samym leksemem w sensie SGJP lub z leksemami powiązanymi derywacyjnie (CHĘĆ – CHCIEĆ). Nie próbuje się więc wyszukać

wszystkich jednostek leksykalnych, które mogłyby zostać powiązane z tą samą ramą. Uznano, że próba unifikowania ram semantycznych dla zbliżonych predykatów mogłaby na tym etapie rozwoju zasobu zbyt łatwo prowadzić do przekłamań w wyniku łączenia ram „na siłę". Przy przyjętej strategii późniejsza analiza powiązań może pozwolić na wykrycie ewentualnych błędów, gdy z analizy układów ról wynika, że dane jednostki nie mogą być synonimami albo zgodność ram sugeruje możliwość rozszerzenia synonimii.

W chwili pisania tych słów warstwa semantyczna Walentego nie jest jeszcze wykonana w całości. Ramy semantyczne ma przypisane około 66% haseł (81% haseł czasownikowych i 37% haseł rzeczownikowych).

3.2.1. Role semantyczne

W literaturze zaproponowano wiele różnych zestawów ról – od wyróżniania ról bardzo specyficznych dla poszczególnych predykatów (typu *czytelnik czyta czytadło*) do bardzo generycznych typu arg0, arg1, ..., argn. Po analizie istniejących systemów ról, zwłaszcza FrameNet, VerbNet, Vallex, autorki warstwy semantycznej Walentego przyjęły własny zestaw ról, najbardziej zbliżony do VerbNetu (Hajnicz i Andrzejczuk 2018).

Przyjęto, że zestaw ról musi spełniać następujące warunki: role powinny być przypisywane wszystkim podrzędnikom predykatu, wymaganym i luźnym; dany podrzędnik może odgrywać tylko jedną rolę; zestaw ról jest jak najmniejszy; powinno być możliwe przypisanie tych samych ról w wypowiedzeniach stanowiących parafrazy.

Role podzielono na dwie grupy: główne, reprezentujące uczestników sytuacji (*Initiator*, *Theme*, *Stimulus*, *Experiencer*, *Instrument*, *Factor*, *Recipient*, *Result*) oraz role poboczne reprezentujące okoliczności (*Condition*, *Attribute*, *Manner*, *Location*, *Path*, *Time*, *Duration*, *Measure*, *Purpose*), zob. tabelę 3.1. Role poboczne częściej dotyczą niewymaganych podrzędników predykatów, choć zdarzają się wyjątki, na przykład *Manner* dla wymaganego podrzędnika xp(mod) przy czasowniku TRAKTOWAĆ lub *Duration* przy TRWAĆ. Jak piszą Hajnicz i Andrzejczuk (2018), w Walentym notowane są nie tyle podrzędniki wymagane, co charakterystyczne.

Role są również klasyfikowane z innego punktu widzenia na trzy typy. Role inicjujące obejmują tych uczestników sytuacji, którzy sprawiają, że się ona dzieje. Role towarzyszące są przypisywane komuś (czemuś), z kim (z czym) coś się dzieje lub też kto (co) podlega czynności, umożliwia jej realizację lub znajduje się w stanie wyrażonym przez predykat. Role zamykające opisują wyniki działania (są przydzielane komuś lub czemuś, co zrodziło się lub ukształtowało podczas czynności) lub charakteryzują sytuację po wykonaniu akcji.

Taki system nie wystarcza jednak jeszcze do odróżnienia argumentów przy niektórych czasownikach. Dlatego przyjęto, że reprezentacja ról będzie dwupoziomowa: w razie potrzeby argumenty z przypisaną tą samą rolą są odróż-

Tabela 3.1. System ról semantycznych w Walentym

	inicjujące	towarzyszące	zamykające
role główne	*Initiator* *Stimulus*	*Theme* *Experiencer* *Factor* *Instrument*	*Recipient* *Result*
role poboczne	*Condition*	*Attribute* *Manner* *Location* *Path* *Time* *Duration* *Measure*	*Purpose*
atrybuty	*Source*	*Foreground* *Background*	*Goal*

niane za pomocą czterech atrybutów, występujących w postaci dwóch par: *Foreground* i *Background* oraz *Source* i *Goal*.

Atrybuty *Foreground* i *Background* są dodawane do nazwy roli, gdy dwa argumenty w zdaniu różnią się jedynie kwestią fokusu, a więc gdy zamiana argumentów nie zmienia istotnie zdarzenia. Na przykład w zdaniu (125) obu osobom przypisana jest rola *Initiator* z atrybutem *Foreground* lub *Background* – nie jest istotne, kto z kim się zamienił, efekt jest ten sam. W zdaniu (126) jest para argumentów *Initiator* i para *Theme* z analogicznymi atrybutami.

(125) *Maria*$_{\textit{Initiator Foreground}}$ *zamieniła się z Anną*$_{\textit{Initiator Background}}$ *na kurtki.*
(126) *Stryjek*$_{\textit{Initiator Foreground}}$ *zamienił z Markiem*$_{\textit{Initiator Background}}$ *siekierkę*$_{\textit{Theme Foreground}}$ *na kijek*$_{\textit{Theme Background}}$*.*

Atrybuty *Source* i *Goal* różnicują sytuacje, gdy argumenty różnią się szeroko rozumianym kierunkiem działania. Rola *Location Source* odpowiada xp(abl) na poziomie składniowym, a *Location Goal* – xp(adl). Możliwe jest nawet trzykrotne użycie roli *Location*:

(127) *Przyszedł z domu*$_{\textit{Location Source}}$ *do pracy*$_{\textit{Location Goal}}$*.*
(128) *W Polsce*$_{\textit{Location}}$ *latał z Krakowa*$_{\textit{Location Source}}$ *do Warszawy*$_{\textit{Location Goal}}$*, w Stanach*$_{\textit{Location}}$ *– z Nowego Jorku*$_{\textit{Location Source}}$ *do Chicago*$_{\textit{Location Goal}}$*.*

Atrybuty te są także używane m.in. przy rolach dla czasowników związanych z kupowaniem i sprzedawaniem.

3.2.2. Preferencje selekcyjne

Argumenty są w ramach opatrywane informacją o preferencjach selekcyjnych. Preferencje selekcyjne (ang. *selectional preferences*) to ograniczenia możliwości zastosowania danego predykatu do określonych argumentów (Hajnicz 2011). Na przykład, żeby odgrywać rolę *Initiator* predykatu PIĆ-1, trzeba być osobą lub inną istotą żywą, a rolę *Theme* musi odgrywać ciecz. Preferencje określają typowe sposoby realizacji danych ról, możliwe są więc zdania wykraczające poza preferencje selekcyjne. W wyniku tego zdanie może stać się bezsensowne, przenośne, poetyckie.

Preferencje selekcyjne mogą być wyrażane prawdziwościowo – mają stanowić kategoryczną odpowiedź na pytanie, czy dane wypowiedzenie narusza zasady selekcyjne, w takim ujęciu nazywa się je ograniczeniami selekcyjnymi. Preferencje wyrażone statystycznie określają, w jakim stopniu dane wypowiedzenie je narusza. Preferencje zapisywane w Walentym mają wskazywać przeważającą większość możliwych wypełnień argumentów, a nie tylko wskazywać jednostki najbardziej typowe (Hajnicz i Andrzejczuk 2018). Tak więc interpretacje wykraczające poza preferencje selekcyjne powinny być uznawane za mniej prawdopodobne, ale nie niemożliwe.

W literaturze proponowano różne sposoby zapisywania preferencji selekcyjnych (por. Hajnicz 2011, rozdz. 6). Na przykład słownik Polańskiego (1980–1992) określa preferencje za pomocą predefiniowanych cech takich jak: osobowość +/–Hum, żywotność +/–Anim, abstrakcyjność +/–Abstr. Możliwości tego systemu są dość ograniczone, dlatego redaktorzy posługują się też wolnymi glosami, które nie nadają się do zastosowań formalnych.

W Walentym przyjęto bardzo precyzyjny sposób określania preferencji selekcyjnych poprzez wskazanie zbioru jednostek leksykalnych Słowosieci, które mogą być centrami fraz realizujących argument.

Najprostszą postacią tego mechanizmu jest wskazanie jednego lub kilku synsetów Słowosieci. Oznacza ono, że dany argument może być realizowany przez elementy tego synsetu i ich hiponimy (jeżeli więc SZCZEKAĆ może PIES, to należy rozumieć, że szczeka też JAMNIK i BULDOG). Ze względu na strukturę Słowosieci należy także dorozumiewać, że preferencję wypełniają jednostki powiązane pewnymi innymi relacjami, np. deminutywy (PIESEK) i augmentatywy (PSISKO), niekiedy może być też konieczne odwołanie do żeńskich odpowiedników.

Ponieważ pewne listy synsetów często się powtarzają, wprowadzono dla nich specjalne oznaczenia, np. symbol JADŁO obejmujący POKARM-1 i NAPÓJ-1, LUDZIE – OSOBA-1, GRUPA LUDZI-1, ISTOTY – OSOBA-1, ISTOTA ŻYWA-1, GRUPA ISTOT-1 itd.

Symbol ALL oznacza dowolną realizację argumentu, a więc w zasadzie brak możliwości określenia preferencji selekcyjnych.

Wprowadzono także możliwość określania preferencji poprzez wskazanie relacji argumentu względem innego argumentu. Pozwala to na przykład wy-

specyfikować, że jeżeli urządzenie *miele*, to robi to za pomocą swoich części (meronimia).

Niekiedy używane jest ogólniejsze określenie RELAT oznaczające dowolną bliską relację w Słowosieci. Na przykład rola *Theme Foreground* predykatu SKŁADAĆ SIĘ-1 może mieć dowolną realizację, ale realizacja roli *Theme Background* musi być jakoś blisko powiązana z tą pierwszą:

(129) *Obiad składał się z drugiego dania i kompotu.*
(130) *Jury składało się ze znanych dziennikarzy.*
(131) *Stado składało się z byków, krów, kóz i owiec.*

Ostatni sposób określenia preferencji selekcyjnej polega na wskazaniu relacji z konkretnym synsetem. Na przykład preferencja dla czasownika PISAĆ obejmuje m.in. DŁUGOPIS, PIÓRO, OŁÓWEK. Jednak bezpośrednim hiperonimem tych pojęć w Słowosieci jest ARTYKUŁ PAPIERNICZY, który jest pojęciem zbyt szerokim. Lepszy zakres preferencji uzyskano, specyfikując, że argument ten musi być powiązany relacją holonimii z jednostką Słowosieci PRZYBORY DO PISANIA, która reprezentuje kolekcję narzędzi pisarskich (choć niestety zawiera też jednostkę EKIERKA).

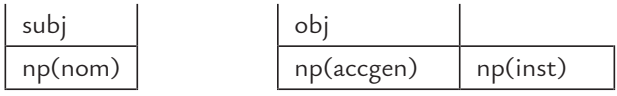
Jak z tego wynika, niekiedy precyzyjne określenie preferencji jest problematyczne. Tak dzieje się na przykład dla argumentu *Theme Source* czasownika POSYPYWAĆ (tego, czym się posypuje). Posypywać można czymś, co składa się z cząstek odpowiednio drobnych w stosunku do posypywanego obiektu. Własność tę mają bardzo różne substancje, których nie wyróżnia wspólny hiperonim ani przynależność do kolekcji. W takiej sytuacji konieczne jest wypisanie w Walentym listy lub przyjęcie zbyt ogólnego ograniczenia (w najgorszym razie – ALL).

3.2.3. Powiązanie warstwy składniowej z semantyczną

Ramy semantyczne Walentego są powiązane ze schematami składniowymi, co pokazuje, w jaki sposób składniowy może być realizowana dana interpretacja semantyczna. Powiązanie to ma charakter wiele-do-wielu. Dana rama semantyczna może być powiązana z wieloma schematami składniowymi – oznacza to, że argumenty semantyczne realizujące dane role mogą być składniowo wyrażane w różny sposób (przykłady (132) i (133)[6]). Powiązanie danego schematu z wieloma ramami oznacza, że dany predykat ma wiele znaczeń (przykłady (134) i (135), Hajnicz i Andrzejczuk 2018). Powiązanie ramy ze schematem wymaga wskazania, które role semantyczne odpowiadają którym pozycjom składniowym.

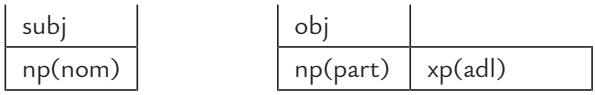
[6] Pokazano schemat składniowy przypasowany do składników zdania, a poniżej role semantyczne z powiązanej ramy semantycznej i jednostkę Słowosieci stanowiącą predykat. Nie pokazano zapisu preferencji selekcyjnych.

(132) *Chłopcy napchali kieszenie pieniędzmi.*



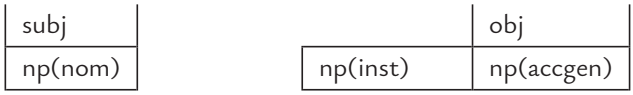
Initiator napchać-1 Location Goal Theme

(133) *Chłopcy napchali pieniędzy do kieszeni.*



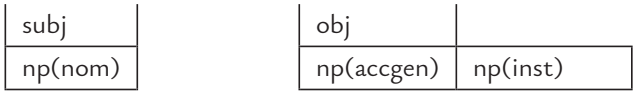
Initiator napchać-1 Theme Location Goal

(134) *Piotr przejechał samochodem psa.*



Initiator przejechać-4 Instrument Theme

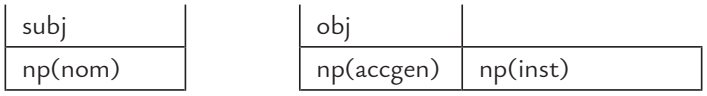
(135) *Piotr przejechał ten dystans samochodem.*



Initiator przejechać-1 Measure Instrument

Zdarza się także, że ten sam schemat może być powiązany z daną ramą na więcej niż jeden sposób. W Walentym zjawisko takie jest nazywane autoalternacją. Przykładem (wg. Hajnicz *et al.* 2016a) mogą być następujące dwa zdania realizujące ten sam schemat składniowy zawierający podmiot np(nom), dopełnienie np(accgen) i pozycję np(inst). Przypisanie ról do pozycji jest jednak w obu zdaniach różne, w szczególności drugi schemat wyklucza przyporządkowanie roli *Attribute*:

(136) *Trawa porasta zbocze zielonym dywanem.*



Theme Source Theme Goal Attribute

(137) *Zbocze porasta trawą.*

Theme Goal Theme Source

Ze względu na to, że poziom składniowy abstrahuje od semantyki, a więc że dany schemat może być wykorzystywany dla różnych znaczeń predykatu i z różnymi ramami, może się zdarzyć, że nie wszystkie typy argumentów przypisane do danej pozycji mogą odgrywać daną rolę semantyczną. Dlatego zawsze wskazywane jest, jaki podzbiór typów fraz składających się na daną pozycję składniową może odgrywać daną rolę. Można to też widzieć w ten sposób,

że powiązane są poszczególne typy fraz z rolami, ale frazy dzielące wspólną pozycję muszą być powiązane z tą samą rolą.

Również nie wszystkie pozycje w schemacie składniowym muszą pasować do przypisanej ramy. Jednak do opisu danego wystąpienia predykatu może posłużyć jedynie rama, która ma interpretacje dla wszystkich pozycji, które zostały wypełnione w danym wypowiedzeniu. Na przykład w powiązanej ze schematem (138) ramie opisującej czasownik ruchu NOSIĆ-1 pozycje xp(abl) i xp(adl) są powiązane, odpowiednio, z rolami *Location Source* i *Location Goal*. W ramie dla czasownika stanowego NOSIĆ-2 (charakteryzować się, cechować się) tych argumentów nie ma. Dlatego zdanie (139), w którym jest zrealizowana pozycja xp(adl), nie może być interpretowane jako realizacja ramy dla NOSIĆ-2. Na podstawie samej analizy ról zdanie (140) mogłoby być przypisane do obu ram, odrzucenie pierwszej powinny zapewnić preferencje selekcyjne.

(138)

subj	obj				
np(nom)	np(accgen)	np(dat)	xp(abl)	xp(adl)	xp(locat)

(139) subj(np(nom)) *Dwaj mężczyźni* *nosili* xp(adl) *gdzieś* np(accgen) *nowoczesne kaloryfery*. [NKJP300]

(140) *[...]* subj(np(nom)) *utwory te* *noszą* np(accgen) *cechy wczesno-pozytywistycznej tendencji i nadmiernego dydaktyzmu*. [NKJP300]

W wypadku schematów z elementami zleksykalizowanymi możliwe są dwie sytuacje. W pierwszej element leksykalny jest modyfikatorem predykatu i ma w związku z tym przypisaną rolę semantyczną (którąś z ról typowych dla luźnych modyfikatorów, np. *Manner*). Dzieje się tak w wypadku określenia *upijać się na umór*, w którym *na umór* odgrywa rolę *Manner* przy predykacie PIĆ.

Druga możliwość polega na tym, że element zleksykalizowany sprawia, że całość nabiera wyraźnie innego znaczenia. Wówczas wprowadzana jest wieloczłonowa jednostka leksykalna, fraza zleksykalizowana zaś nie ma przypisywanej roli (technicznie jest jej przypisywana rola *Lemma*). Na przykład wyrażenie *drzeć koty*, choć składniowo stanowiące realizację schematu czasownika DRZEĆ, jest powiązane z jednostką Słowosieci DRZEĆ KOTY-1. Fraza *koty* składniowo jest zleksykalizowanym podrzędnikiem np(accgen), który w ramie semantycznej nie ma roli – staje się częścią lematu wieloczłonowego.

W drzewach składniowych generowanych przez analizator Świgra 2 notowane są zrealizowane w danej konstrukcji fragmenty schematu składniowego. Dzięki powiązaniu warstwy składniowej i semantycznej w słowniku Walenty jest to informacja wystarczająca do generowania struktury predykatowo-argumentowej odpowiadającej danej konstrukcji składniowej (Bartosiak 2017). Otwiera to możliwość rozbudowy przedstawionego w niniejszej pracy opisu składniowego w kierunku reprezentacji semantyki wypowiedzeń.

4

Implementacja gramatyki w analizatorze Świgra 2

Celem tego rozdziału jest przedstawienie sposobu realizacji gramatyki w analizatorze Świgra 2, który pozwala generować struktury opisane w rozdziale 2. Nie jest to pełna prezentacja wszystkich reguł gramatyki, taka byłaby bowiem zbyt szczegółowa, a przez to trudna do prześledzenia i raczej niepotrzebna. Gramatyka jest publicznie dostępna w postaci źródłowej i działającego programu komputerowego pod adresem http://swigra.nlp.ipipan.waw.pl/. Można więc się odwołać zarówno do jej treści, jak i do wyników działania analizatora.

Zaprezentowany zostanie formalizm i sposoby zastosowania go do realizacji przedstawianej gramatyki[1]. Oprócz ogólnej charakterystyki zostaną pokazane elementy stanowiące rozszerzenia formalizmu DCG, dzięki którym możliwe jest generowanie drzew w postulowanym kształcie na podstawie zwartych reguł gramatyki (p. 4.2). W punkcie 4.4 omówiono funkcję, jaką pełnią poszczególne parametry jednostek nieterminalnych, oraz ciekawsze z mechanizmów przetwarzania parametrów. W punkcie 4.5 opisano fundamentalny dla gramatyki mechanizm realizacji wymagań składniowych, a więc dopuszczania fraz wymaganych poszczególnych typów w zależności od zapisów w słowniku walencyjnym.

4.1. Zastosowany formalizm gramatyczny

Przyjęty formalizm ma charakter składnikowy, opisuje więc budowę jednostek poprzez wskazanie ich dopuszczalnych zestawów składników (w szczególności dopuszczalnej kolejności ich występowania). Dane wypowiedzenie jest akceptowane (zgodne z gramatyką), jeżeli da się wywieść z symbolu początkowego gramatyki za pomocą jej reguł. Gramatyka nie jest więc zestawem ograniczeń, które musi spełniać wypowiedzenie, ale przepisem na budowę wypowiedzenia i jednoczesną budowę reprezentującej je struktury.

[1] Pełne zrozumienie szczegółów przedstawianych przykładów wymaga znajomości języka programowania Prolog. Dalsze przypisy w tym rozdziale służą przybliżeniu znaczenia zapisów w Prologu osobom niezaznajomionym z tym językiem.

Świgra 2 wywodzi się z gramatyki Świdzińskiego, która jest zapisana w formalizmie gramatyk metamorficznych (Colmerauer 1978), później lepiej znanych jako Definite Clause Grammars (Pereira i Warren 1980). Formalizm ten można widzieć jako gramatykę unifikacyjną rozszerzającą gramatyki bezkontekstowe.

Gramatyka taka stanowi czwórkę uporządkowaną złożoną ze zbioru symboli terminalnych, zbioru symboli nieterminalnych, zbioru produkcji i symbolu początkowego gramatyki. Symbole terminalne to elementy przetwarzanego tekstu. W prostym wypadku byłyby to słowa lub segmenty, a w wypadku analizatora Świgra 2 symbolami terminalnymi są formy fleksyjne. Symbole nieterminalne to jednostki składniowe reprezentujące składniki wypowiedzenia. Reguła gramatyki bezkontekstowej składa się z lewej strony będącej jednym symbolem nieterminalnym i prawej strony będącej ciągiem symboli nieterminalnych i terminalnych. Reguła wyraża to, że składnikami bezpośrednimi jednostki stanowiącej lewą stronę reguły mogą być jednostki wymienione po prawej stronie (w zadanej kolejności). Na przykład następująca reguła mogłaby opisywać wewnętrzną budowę frazy nominalnej *mały biały domek na wzgórzu* przy założeniu, że odcinki *mały* i *biały* są analizowane jako frazy przymiotnikowe **fpt**, *domek* stanowi prostszą frazę nominalną **fno**, a odcinek *na wzgórzu* – frazę przyimkowo-nominalną **fpm**:

fno ⟶ **fpt**, **fpt**, **fno**, **fpm**.

W przytoczonym przykładzie jednostki nieterminalne są atomami języka Prolog, formalizm DCG dopuszcza jednak, żeby rolę jednostek nieterminalnych pełniły dowolne termy[2]. Można więc wyobrazić sobie na przykład wpro-

[2] Na potrzeby tej prezentacji można przyjąć, że atom to nazwa złożona z liter, cyfr i znaków podkreślenia, zaczynająca się od małej litery, np. kotu, kot, m3, dat, sg. Atomami są też liczby naturalne. Jeżeli atom ma zawierać inne znaki, jego zapis ujmuje się w apostrofy, np. 'Psem'. Atomy stanowią zwartą reprezentację dla danych, które w językach imperatywnych byłyby reprezentowane w postaci stałych napisów. W analizatorze Świgra 2 w postaci atomów reprezentowane są segmenty, lematy, wartości kategorii gramatycznych itp.

Zmienne w Prologu mają nazwy zaczynające się wielką literą lub podkreśleniem, np. *Rodzaj*. Zmienne oznaczają wartości nieznane. Wykonanie programu w Prologu jest próbą znalezienia wartości dla zmiennych, spełniających zadane przez program warunki. Zmienne mogą być w trakcie wykonywania programu ukonkretniane (otrzymywać jako wartość pewien term). Raz nadana wartość nie zmienia się w trakcie dalszego wykonania programu.

Termy definiuje się rekurencyjnie. Atomy i zmienne są termami. Ponadto termem jest wyrażenie złożone z atomu (nazywanego funktorem termu), po którym w nawiasach okrągłych stoi pewna liczba termów rozdzielonych przecinkami (te termy nazywa się argumentami, a ich liczbę – arnością termu), np. fno(dat, sg, *Rodzaj*). W niniejszej pracy funktory termów stanowiących symbole nieterminalne gramatyki są wyróżniane pogrubieniem.

Dla przejrzystości zapisu język Prolog daje możliwość stosowania operatorów infiksowych (np. +, −, *, /). Są one skrótowym zapisem termów o arności 2. Tak więc wyrażenie 1 + 2 jest tym samym co '+'(1, 2).

wadzenie argumentów jednostek **fno** i **fpt** wyrażających przypadek gramatyczny *P*, rodzaj *R* i liczbę *L*. Za ich pomocą można wyrazić zachodzenie uzgodnienia w zakresie tych kategorii między frazami przymiotnikowymi i nadrzędną frazą nominalną:

fno(*P*, *R*, *L***)** ⟶ **fpt(***P*, *R*, *L***)**, **fpt(***P*, *R*, *L***)**, **fno(***P*, *R*, *L***)**, **fpm**.

Wystąpienie tego samego symbolu zmiennej jako atrybutu wielu jednostek oznacza, że wartości odpowiedniego atrybutu poszczególnych jednostek muszą dać się zunifikować[3]. Powtórzenie tych samych symboli po lewej stronie reguły oznacza, że odpowiednie wartości atrybutów staną się charakterystyką frazy nominalnej jako całości.

Jeżeli zbiór wartości atrybutów jednostek nieterminalnych jest skończony, to możliwość użycia termów złożonych nie zwiększa siły wyrazu względem gramatyk bezkontekstowych, pozwala jedynie na elegancką generalizację: zamiast wypisywać reguły realizacji fraz nominalnych w poszczególnych przypadkach, podaje się jedną regułę sparametryzowaną wartością przypadka gramatycznego.

Jednak w formalizmie DCG atrybutami jednostek mogą być dowolnie złożone termy, na przykład struktury listowe, których można użyć do realizacji nieograniczonych liczników. To zaś oznacza istotne rozszerzenie mocy formalizmu, do poziomu tzw. gramatyk indeksowanych, które pozwalają m.in. zapisać gramatykę języka $\{a^n b^n c^n : n \geq 0\}$, który nie jest bezkontekstowy.

Kolejnym elementem DCG, który wykracza poza gramatyki bezkontekstowe, jest możliwość wprowadzenia do reguł warunków, które są wywołaniami dowolnych predykatów zdefiniowanych w Prologu. Jeżeli obliczenie danego predykatu zawodzi, to znaczy, że dla danych wartości argumentów reguła nie może zostać zastosowana[4]. W następującym przykładzie pokazano zwarty sposób realizacji dopełniacza negacji:

fw(np(accgen), *Neg***)** ⟶ **fno(***P*, *R*, *L***)**, **{** member(*P*/*Neg*, [acc/tak, gen/nie]) **}**.

Termy nie mają żadnej z góry ustalonej semantyki, tak więc wyrażenie 1 + 2 nie jest obliczane do liczby 3, reprezentuje ono tylko dwuargumentowy term z funktorem +. Wyjątek stanowią wbudowane predykaty arytmetyczne, o których mowa w przypisie 8.

[3] Unifikacja jest fundamentalną operacją programowania w logice zdefiniowaną rekurencyjnie. Zmienną można zunifikować z dowolnym termem (w szczególności z inną zmienną). Atom unifikuje się tylko ze sobą samym lub zmienną. Dwa termy złożone można zunifikować, jeżeli mają ten sam funktor, tę samą arność i kolejne ich argumenty dają się parami zunifikować ze sobą.

[4] Procedury definiowane w języku Prolog, nazywane predykatami przez nawiązanie do logiki pierwszego rzędu, implementują relacje logiczne. Wykonanie programu w języku Prolog jest próbą sprawdzenia, czy zdefiniowana relacja zachodzi dla danych argumentów. Jeżeli niektóre argumenty są zmiennymi, obliczenie prowadzi do wyznaczenia wartości tych zmiennych spełniających relację. Jeżeli takie wartości istnieją, predykat odnosi sukces, w przeciwnym razie ponosi porażkę (zawodzi).

Fraza wymagana typu np(accgen) (a więc realizowana przez frazę nominalną w przypadku strukturalnym, por. p. 3.1.4) występująca w kontekście wartości *Neg* negacji w zdaniu może być realizowana przez frazę nominalną **fno** o wartości przypadka *P*. Wartości rodzaju *R* i liczby *L* nie są w tej regule istotne. Warunek zawarty w regule wymaga, żeby term *P*/*Neg* był elementem listy [acc/tak, gen/nie][5]. Oznacza to, że reguła może zostać zastosowana, jeżeli składnik **fno** ma wartość przypadka acc, a wartością negacji w zdaniu jest tak albo przypadek ma wartość gen, a negacja – nie. Inne kombinacje wartości sprawią, że warunek zawiedzie.

Dzięki przekazywaniu informacji w atrybutach jednostek oraz obliczeniom w warunkach formalizm DCG jest równoważny maszynie Turinga co do siły wyrazu. W związku z tym oczywiście można zapisać gramatyki odpowiadające rozwiązaniu problemów NP zupełnych, a w związku z tym dla odpowiednio napisanej gramatyki czas analizy musi być wykładniczy (Francez i Wintner 2011). Problem ten nie występuje jednak w typowych gramatykach języka naturalnego, które mogą zachować wielomianowy czas analizy.

Drzewa przypisywane wypowiedzeniom są drzewami wywodu (drzewami analizy) dla gramatyk DCG. Każdy wierzchołek drzewa odpowiada zastosowaniu jednej reguły gramatyki. W wierzchołku umieszcza się nieterminal stojący po lewej stronie reguły, a jego dzieci reprezentują symbole nieterminalne i terminalne występujące po prawej stronie reguły.

W analizatorze Świgra 2 nie jest stosowana implementacja DCG wbudowana w interpreter Prologu, ale niezależna implementacja stworzona od podstaw, stosująca strategię wstępującą i zapamiętywanie wyników pośrednich (Woliński 2004). Przyjęta strategia powoduje, że niemożliwe są reguły o pustej prawej stronie, co nie stanowi istotnego ograniczenia dla twórcy gramatyki. Warto także podkreślić, że analizator może pracować na nieujednoznacznionych grafach fleksyjnych generowanych przez analizator fleksyjny (por. p. 1.11), a generuje upakowane lasy drzew składniowych (*shared parse forests*, por. p. 6.3).

4.2. Rozszerzenia formalizmu DCG

Drzewa wywodu odpowiadające zastosowaniu reguł formalizmu składnikowego o sztywnym formacie reguł mają arność wierzchołków (liczbę wierzchołków potomnych) wynikającą ze składu poszczególnych reguł. Jak wspomniano wcześniej (s. 92), powoduje to trudność opisania zmiennej liczby pod-

[5] Zapis [*A*, *B*, *C*, ...] oznacza w Prologu listę złożoną z elementów *A*, *B*, *C*, ... Predefiniowany predykat member(*E*, *L*) wyraża fakt, że term *E* jest elementem listy *L*. Operator / został użyty do zestawienia wartości w pary (jak wspomniano wcześniej, nie ma on semantyki dzielenia). Dzięki temu zabiegowi sprawdzana jest jednocześnie równość dwóch wartości.

rzędników, zwłaszcza przy czasownikach, odpowiadającej elastycznie budowanym schematom składniowym w słowniku Walenty.

Aby uporać się z tym problemem, zaproponowano rozszerzenie formalizmu DCG umożliwiające ujęcie w jednej regule konstrukcji, które mają być reprezentowane w drzewie składnikowym wierzchołkami o różnej arności. Wprowadzono w tym celu operatory wyrażające pomijalność lub powtarzalność wskazanych składników i ciągów składników prawej strony reguły. Bartosiak i Woliński (2015) analizują problemy efektywności analizy z tym związane.

Warto może jeszcze wspomnieć, że Świdziński stosował w oryginalnym zapisie swojej gramatyki rozszerzenie oznaczające permutację składników (co umożliwiało zdanie sprawy ze swobodnego szyku, ale nie ze zmiennej liczby składników), natomiast pierwotna implementacja GFJP (Świgra 1) została wyposażona w rozszerzenia służące realizacji wymagań czasownikowych. Były to operatory **wymagania** i **wymagane** (Woliński 2004, §5.2.4, s. 84). W obecnej realizacji rozszerzono formalizm w sposób bardziej ogólny, niezwiązany tak silnie z konkretną gramatyką.

4.2.1. Elementy opcjonalne reguł

Prostszym z wprowadzonych rozszerzeń jest operator **optional**, który pozwala wyrazić to, że wystąpienie danego składnika jest opcjonalne. Operator ma trzy argumenty: nieterminal, warunki obliczane, jeśli nieterminal wystąpił, i warunki obliczane w przeciwnym wypadku. Warunki zapisywane są w nawiasach klamrowych, pusta para nawiasów sygnalizuje brak warunków.

Następujący przykład ilustruje typowe użycie w analizatorze Świgra 2. Opisywana jest realizacja frazy luźnej w postaci frazy przyimkowej **fpm** ujętej w przecinki. Przecinek po prawej stronie jest pomijany, jeżeli fraza ta występuje na przykład na końcu wypowiedzenia.

fl(*A*, *C*, *Rl*, *O*, *Neg*, *Dest*, *I*, *Pk*, *Sub***)** ⟶
  przec(*Pk0*, po**)**,
  fpm(*Pm*, *H*, *P*, *Kl*, *Zap*, *Neg*, *Dest*, *I*, *Pk1*, na**)**,
  optional(przec(*Pk2*, po**)**, **{** oblpk(*Pk*, [*Pk0*, *Pk1*, *Pk2*]) **}**, **{** oblpk(*Pk*, [*Pk0*, *Pk1*, p/bp]) **})**.

Warunki oblpk zostały użyte do sprawdzenia zgodności ze sobą wartości „przecinkowości" *Pk* poszczególnych składników. Ten mechanizm zostanie dokładniej objaśniony w punkcie 4.4.5. W tym miejscu warto jedynie zauważyć, że jeżeli jednostka **przec(***Pk2*, po**)** wystąpiła, to w obliczeniu „przecinkowości" frazy luźnej biorą udział wartości *Pk0*,*Pk1*,*Pk2* pochodzące od poszczególnych składników. Jeżeli jednostka nie wystąpiła, w obliczeniu używana jest zamiast ostatniego składnika ustalona wartość p/bp.

Dla uproszczenia implementacji przyjęto, że operatora **optional** nie można użyć jako pierwszego składnika prawej strony reguły.

Drugim rozszerzeniem w stosunku do DCG jest operator pozwalający wyrazić, że pewne składniki mogą wystąpić w danym miejscu zero lub więcej razy.

W najprostszej postaci operatora **sequence_of** używa się z argumentem będącym listą nieterminali gramatyki. Każdy z nich może występować w dowolnej liczbie i kolejności. Na przykład reguła

zdanie ⟶ **ff**, **sequence_of([fw, fl])**.

pozwala akceptować ciągi złożone z jednostki **ff**, po której nastepuje dowolny ciąg wystąpień jednostek **fw** i **fl**, w szczególności pusty. W utworzonym drzewie składniowym zastosowaniu tej reguły będzie odpowiadał wierzchołek **zdanie** o arności zależnej od liczby zaakceptowanych jednostek **fw** i **fl**.

Użycia operatora **sequence_of** komplikują się, gdy uwzględnić możliwość nakładania warunków na wartości argumentów nieterminali. Zmienne występujące w argumentach jednostek nieterminalnych są domyślnie dzielone z otaczającą regułą na mocy zasad zasięgu zmiennych w Prologu. Tak więc w następującym przykładzie pierwszy argument jednostki **ff** musi być unifikowalny z pierwszym argumentem każdego wystąpienia jednostki **fw** w sekwencji (jeżeli takie są):

zdanie ⟶ **ff(***A***)**, **sequence_of([fw(***A***)])**.

Ważką i nie zawsze pożądaną konsekwencją tej konwencji jest to, że domyślnie wszystkie wystąpienia jednostki w sekwencji dzielą się wartościami swoich argumentów. Aby uzyskać odpowiednią siłę wyrazu formalizmu, konieczne było wprowadzenie mechanizmu pozwalającego sterować sposobem unifikowania się zmiennych. Jest to potrzebne na przykład, aby opisać sekwencje fraz wymaganych **fw** różnych typów, skoro typ jest jednym z argumentów.

Wprowadzono trzy mechanizmy: warunki obliczane osobno dla poszczególnych elementów sekwencji, warunki iterowane przekazujące częściowe wyniki obliczeń do kolejnych wystąpień w sekwencji, wreszcie blokowanie uzgadniania zmiennych ze środowiskiem zewnętrznym.

Warunki wiązane z poszczególnymi wystąpieniami elementu sekwencji

Pierwszą możliwość ilustruje następujący przykład (nieistotne dla bieżących rozważań zmienne i fragmenty faktycznej reguły zastąpiono wielokropkiem):

```
fno(P, ...) ⟶
    fno(P, ...),
    sequence_of([
        ...,
```

```
    fno(gen, ..., Kl4, ...)
    ^[sprawdz_kl, Kl4],
    ...
    ]).
sprawdz_kl(Kl) :- Kl \= kto.
```

Notacja z operatorem ^ jest inspirowana standardowym predykatem bagof/3 Prologu, w wypadku którego podobny zapis jest stosowany, aby wskazać, że pewne zmienne mają być potraktowane w sposób szczególny (nie być unifikowane ze środowiskiem).

Zapis ^[sprawdz_kl, *Kl4*] wskazuje warunek sprawdzany dla poszczególnych wystąpień iterowanego nieterminala **fno** (w przykładzie – frazy nominalnej w dopełniaczu stanowiącej podrzędnik dla nadrzędnika rzeczownikowego w przypadku *P*). Warunek jest zapisywany w postaci listy, której pierwszy element powinien być funktorem głównym jednoargumentowego predykatu, a drugi – zmienną, która stanie się argumentem. Dla każdego znalezionego wystąpienia nieterminala **fno(**gen, ..., *Kl4*, ...**)** zostanie wywołany predykat sprawdz_kl(*Kl4*), przy czym zmienna *Kl4* jest lokalna dla tego wystąpienia nieterminala[6].

Podanie po operatorze ^ dwuelementowej listy wskazuje więc dwie rzeczy: że ma być sprawdzony warunek oraz że wskazana zmienna ma nie być uzgadniana z innymi instancjami danego nieterminala w sekwencji. Taki „rozczłonkowany" zapis warunku ma zwracać uwagę na to, że nie jest to tylko wywołanie predykatu prologowego.

W związku z przyjętym zapisem warunki sprawdzane dla poszczególnych elementów sekwencji muszą być formalnie jednoargumentowe. Jeżeli zachodzi potrzeba odwołania się do wartości większej liczby atrybutów nieterminala, można jako argument warunku przekazać term złożony, na przykład zbudowany za pomocą funktora / podobnie jak w poprzednich przykładach. Wszystkie zmienne występujące w tym termie nie są unifikowane pomiędzy wystąpieniami nieterminali.

W następującym przykładzie obliczany warunek odwołuje się do wartości dwóch zmiennych: *Kl* i *Kl4*:

```
fno(P, ..., Kl, ...) ⟶
    fno(P, ..., Kl, ...),
    sequence_of([
        ...,
```

[6] Predykat sprawdz_kl jest zdefiniowany za pomocą zapisu podanego pod regułą. Symbol :– można czytać jako strzałkę implikacji skierowaną w lewą stronę. Tak więc sprawdz_kl(*Kl*) zachodzi, jeżeli zachodzi *Kl* \= kto. Operator \= przywołuje wbudowany predykat Prologu sprawdzający niemożność unifikacji jego argumentów. W tym wypadku sprawdza się, czy wartość zmiennej *Kl* jest różna od atomu kto. Tak więc przykładowy warunek dopuszcza frazy nominalne klasy różnej od kto.

```
        fno(gen, ..., Kl4, ...)
        ^[sprawdz_kl, Kl/Kl4],
        ...
        ]).
sprawdz_kl(Kl/Kl4) :- Kl \= os,
    Kl4 \= kto.
```

Warto zwrócić uwagę, że mechanizm przetwarzający zmienne warunków wyłącza unifikację jedynie między poszczególnymi wystąpieniami w sekwencji. W przykładzie, skoro zmienna *Kl* występuje też poza sekwencją, jest mianowicie atrybutem frazy **fno** po lewej stronie reguły i nadrzędnika, to te wystąpienia unifikują się z *Kl* w warunku, a zatem wartość jest ta sama we wszystkich elementach sekwencji. O to właśnie chodziło w przedstawionym wywołaniu – aby warunek sprawdz_kl był zależny od zmiennej *Kl* ze środowiska zewnętrznego sekwencji i zmiennej *Kl4* charakteryzującej konkretne wystąpienie nieterminala.

Warunki iterowane po sekwencji

Warunki iterowane są nieco bardziej skomplikowane. Można za ich pomocą wyrazić warunek obliczany w krokach dla sekwencji o dowolnej długości. Ma to znaczenie dla efektywności analizy, ponieważ jeśli warunek zawiedzie w danym kroku, nie jest konieczne konstruowanie dalszej części sekwencji.

W następującym przykładzie przedstawiono wykorzystanie warunków iterowanych w uproszczonej regule opisującej realizację zdania w postaci frazy wymaganej, po której następuje sekwencja fraz wymaganych i luźnych, a fraza finitywna stoi na końcu:

```
zdanie(...) ⟶
    fw(Tfw1, ...),
    sequence_of([
        fw(Tfw2, ...)
        ^[lista_iter, [Tfw1], Tfw2, Wymagane],
        fl(...)
        ^[najwyżej, 3, _, _]
        ]),
    ff(..., Schematy, ...),
    { dopuszczalne_wymagania(Wymagane, Schematy) }.

lista_iter(Lista, Elem, [Elem | Lista]).
najwyżej(N, _, N1) :- N > 0,
    N1 is N – 1.
```

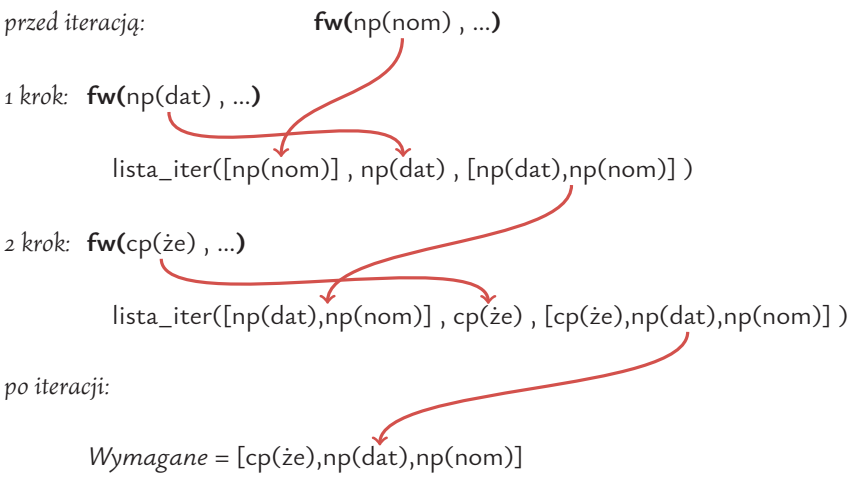
Formalizm warunków iterowanych wymaga zapisania specyfikacji wywołania danego warunku w postaci czteroelementowej listy. Jej elementy to: funk-

tor wywoływanego predykatu; wartość początkowa obliczenia; wartość bieżąca, wiążąca obliczany warunek z wartościami argumentów iterowanego nieterminala, i wartość końcowa. Zmienne występujące w wartości bieżącej podlegają tym samym zasadom co dla warunków prostych, czyli są nowymi zmiennymi w każdym kolejnym wystąpieniu nieterminala.

Sam warunek musi być zdefiniowany jako predykat trójargumentowy. Dla każdego rozpoznanego wystąpienia nieterminala opatrzonego warunkiem iterowanym wywoływany jest predykat warunku. W pierwszym wywołaniu pierwszy argument jest wartością początkową ze specyfikacji. Drugi argument wywołania jest zawsze wartością bieżącą (czyli jest ukonkretniany odpowiednimi wartościami charakteryzującymi dane wystąpienie nieterminala). Trzeci argument danego wywołania stanie się pierwszym argumentem następnego. W wypadku ostatniego wywołania w sekwencji trzeci argument jest unifikowany z wartością końcową ze specyfikacji.

Pierwszy z warunków iterowanych występujących w przykładzie, lista_iter, służy do zestawienia elementów w listę[7]. Ponieważ w regule tej fraza finitywna z czasownikiem determinującym możliwe do zastosowania schematy składniowe występuje na końcu reguły, nie da się sprawdzać zgodności typów fraz wymaganych **fw** z czasownikiem na bieżąco. Zamiast tego typy fraz wymaganych *Tfw2* będą zestawione w listę, która zostanie skonfrontowana z dostępnymi schematami składniowymi, co zamarkowano wywołaniem predykatu dopuszczalne_wymagania.

Jeżeli omawiana reguła zostałaby zastosowana do analizy zdania, w którym można rozpoznać kolejno frazy wymagane typów np(nom), np(dat) i cp(że), obliczenie wykorzystujące przytoczoną definicję predykatu lista_iter/3 wyglądałoby następująco:



[7] Występujący w nim zapis [*G* | *O*] oznacza listę złożoną z pierwszego elementu (głowy) *G* i listy pozostałych elementów (ogona) *O*. Gdyby term ten zunifikować z listą [a, b, c, d], zmienna *G* uzyskałaby wartość a, natomiast zmienna *O* – wartość [b, c, d].

Jak widać, elementy zostały zestawione w listę w kolejności odwrotnej do wystąpień. Można temu zaradzić, stosując następującą wersję definicji predykatu:

```
lista_iter([Elem | Lista], Elem, Lista).
```

i mniej intuicyjną specyfikację warunku: ^[lista_iter, *Wymagane*, *Tfw2*, []].

Drugi przykładowy warunek, najwyżej, służy do ograniczenia liczby wystąpień danego nieterminala[8]. W omawianej regule ograniczono za jego pomocą liczbę możliwych do rozpoznania fraz luźnych fl do trzech. Przykład ten jest interesujący, gdyż dla warunku najwyżej nie jest istotna żadna wartość związana z danym wystąpieniem nieterminala, a jedynie jego sama obecność. Nie jest też istotna wartość wynikowa obliczenia. Dlatego w specyfikacji odpowiednie wartości zostały zamarkowane zmiennymi anonimowymi. Między kolejnymi wywołaniami są przekazywane liczby malejące o 1. Gdy wartość spadnie do 0, warunek zawodzi.

Warunki wspólne

Mechanizm sekwencji umożliwia także nałożenie warunku iterowanego na wszystkie elementy sekwencji. Specyfikacje takich warunków podaje się po nawiasie zamykającym listę iterowanych nieterminali:

```
zdanie(..., Dest, ...) ⟶
    ff(..., Dest1, ...),
    sequence_of([
        fw(..., Dest2, ...),
        fl(..., Dest2, ...)
        ]
         ^[obldest_iter, Dest1, Dest2, Dest]).
```

W tym przykładzie warunek obldest_iter jest wywoływany dla każdego elementu sekwencji, niezależnie od tego, czy jest to fraza wymagana **fw** czy luźna **fl**. Oba te typy jednostek mają atrybut nazywany predestynacją (zob. p. 4.4.1), reprezentowany przez zmienną *Dest2*. Mechanizm obsługi sekwencji zapewnia, że jest to nowa zmienna w każdym elemencie sekwencji dzieląca wartość tylko z odpowiednim warunkiem w tym kroku iteracji.

---

[8] W definicji warunku występują wbudowane predykaty arytmetyczne Prologu, które, nietypowo, przypisują termom pewną semantykę. Predykat is interpretuje swój drugi argument jako wyrażenie arytmetyczne, oblicza jego wartość i unifikuje z pierwszym argumentem. Predykat > odnosi sukces, jeżeli jego argumenty interpretowane jako liczby są w odpowiedniej relacji arytmetycznej.

Zmienne wolne

Ostatnia konwencja notacyjna związana z operatorem **sequence_of** pozwala wskazać, że pewna zmienna występująca w nieterminalu w sekwencji ma się nie uzgadniać z otoczeniem, co oznacza, że dla każdego wystąpienia dopuszczalna jest inna wartość, na którą nie nakłada się żadnych ograniczeń.

Następujący przykład przedstawia regułę z ciągiem dowolnych fraz wymaganych, które mogą różnić się typami:

```
zdanie ⟶ ff,
    sequence_of([
        fw(Tfω)
        ^[Tfω]
        ]).
```

Sekwencje sekwencji

W regułach gramatyki zachodzi czasem potrzeba zapisania powtarzalności ciągu kilku elementów. Na przykład w konstrukcji szeregowej typu I (zob. s. 111) elementem powtarzalnym jest trójka elementów złożona z przecinka, spójnika szeregowego i frazy koordynowanej. Innym przykładem może być występowanie w szeregu frazy jakiegoś typu ujętej w przecinki. Tego rodzaju konfiguracje można również zapisać za pomocą **sequence_of**. Na przykład reguła

```
test ⟶ sequence_of([[a, b], [c, d, e], f]).
```

opisuje konstrukcje złożone z wystąpień ciągu **a**, **b** (a więc jeżeli wystąpiło **a**, musi po nim wystąpić **b**) oraz ciągu **c**, **d**, **e** oraz jednostki **f**.

W wypadku tego zapisu warunki dotyczące poszczególnych nieterminali podaje się po zamykającym nawiasie obejmującym daną podsekwencję:

```
test(K) ⟶ sequence_of([
        [a(X), b(Y)]
        ^[warunek(X)]
        ^[warunek_iterowany, 0, Y, K],
        [c, d, e],
        f
        ]).
```

Przykład reguły

Oto przykład reguły gramatycznej faktycznie wykorzystywanej przez analizator Świgra 2 zapisanej za pomocą sekwencji i warunków iterowanych. Reguła opisuje zdanie złożone z frazy wymaganej, po której następuje sekwencja

fraz wymaganych, luźnych i ewentualne wystąpienie jednostki **posiłk**, po której następuje obowiązkowa fraza finitywna **ff** i druga sekwencja o podobnym składzie.

```
zdanie(Wf, A, C, T, Rl, O, Neg, Dest, I, Pk, Sub) ⟶
    fw(W1, H1, A, C, Rl, O, pre, Neg, Dest1, I1, Pk1, po),
    sequence_of([
        fw(W2, H2, A, C, Rl, O, pre, Neg, Dest2, I2, Pk2, po)
         ^[lista_iter, [W1/H1], W2/H2, PreFfWym],
         fl(A, C, Rl, O, Neg, Dest2, I2, Pk2, po)
         ^[najwyżej, 3, _, IleFl],
         posiłk(KlPosiłk2, Rl, Dest2, I2, Pk2, po)
        ]
         ^[obldest_iter, Dest1, Dest2, DestPreF]
         ^[oblink_iter, I1, I2, IPreF]
         ^[oblpk_iter, Pk1, Pk2, PkPreF]
         ^[sprawdź_posiłk, bezpo, KlPosiłk2, TP4]),
    ff(Wf4, A4, C4, T4, Rl, O4, Wym, Neg4, Dest4, I4, Pk4, na),
    { obldest_iter(DestPreF, Dest4, DestPostF),
    oblink_iter(IPreF, I4, IPostF),
    oblpk_iter(PkPreF, Pk4, PkPostF),
    wyjmijl(PreFfWym, Wym, PostFfWym) },
    sequence_of([
        fw(W5, H5, A, C, Rl, O, post, Neg, Dest5, I5, Pk5, po)
         ^[oblwym_iter, PostFfWym, W5/H5, ReszteWym],
         fl(A, C, Rl, O, Neg, Dest5, I5, Pk5, po)
         ^[najwyżej, IleFl, _, _],
         posiłk(KlPosiłk5, Rl, Dest5, I5, Pk5, po)
        ]
         ^[obldest_iter, DestPostF, Dest5, Dest]
         ^[oblink_iter, IPostF, I5, I]
         ^[oblpk_iter, PkPostF, Pk5, Pk]
         ^[sprawdź_post_posiłk, pierw(TP4), KlPosiłk5, TP]),
    { oblposiłk(TP, Wf4/A4/C4/T4/O4/Neg4, Wf/A/C/T/O/Neg),
    wymagania_zamknij(ReszteWym),
    wyklucz_podmiot_bezokolicznika(Wf, Wym) }.
```

Interesującym aspektem jest przekazywanie pewnych wartości z pierwszej sekwencji do drugiej. Prezentowany poprzednio warunek najwyżej został użyty w obu sekwencjach i końcowa wartość z pierwszej jest przekazywana jako początkowa wartość drugiej. W ten sposób jest ograniczana liczba fraz luźnych występujących w sumie w obu sekwencjach.

Jednostka **posiłk** reprezentuje fragment nieciągłej formy analitycznej czasownika, który może wystąpić zarówno przed czasownikiem (*będzie to czytać, niech to czyta*), jak i po nim (*czytać to będzie*). Warunki sprawdź_posiłk

i sprawdź_post_posiłk są zdefiniowane tak, aby jednostka **posiłk** mogła w danym zdaniu wystąpić tylko raz, a jako partykuła trybu rozkazującego – jedynie przed czasownikiem. Ponadto obecność tej partykuły może zmienić charakterystykę czasową, trybową i wyróżnik fleksyjny frazy czasownikowej. Fraza bezokolicznikowa zestawiona z jednostką **posiłk** wnoszącą czas przyszły otrzymuje charakterystykę frazy finitywnej. Dlatego sprawdzenie bezokolicznikowości następuje dopiero na końcu reguły (wyklucz_podmiot_bezokolicznika).

W sekwencji poprzedzającej frazę finitywną gromadzona jest lista zaobserwowanych fraz wymaganych (lista_iter). Ściślej, gromadzone są typy fraz wymaganych i wartości argumentu *H* reprezentującego informację o centrum leksykalnym frazy, potrzebną w wypadku dopasowywania do zleksykalizowanych opisów wymagań z Walentego (element ten był pominięty w poprzednim uproszczonym przykładzie). Po rozpoznaniu frazy finitywnej zgromadzona lista *PreFFWym* jest konfrontowana z rekcją czasownika (więcej o tym procesie w p. 4.5). Reprezentacja pozostałych do wypełnienia części schematów *PostFfWym* zostaje przekazana do warunku iterowanego oblwym_iter i dopuszczalność fraz wymaganych w drugiej sekwencji jest sprawdzana na bieżąco. Wreszcie na zakończenie opis niewypełnionych wymagań *ReszteWym* trafia do warunku wymagania_zamknij, który może sprawdzić, czy dopełnione zostały ewentualne warunki na (nie)eliptyczność.

Warunki obldest_iter, oblink_iter i oblpk_iter służą do obliczania wartości parametrów, odpowiednio: predestynacji, inkorporacji i przecinkowości zdania. Przysługują one wszystkim składnikom, w szczególności frazie finitywnej, muszą więc być użyte dodatkowo pomiędzy sekwencjami poprzez jawne wywołanie odpowiadających im predykatów.

4.2.3. Zalety i wady mechanizmu sekwencji

Zasadniczym celem wprowadzenia mechanizmów sekwencji i opcjonalności było umożliwienie zapisywania reguł o zmiennej liczbie składników. Z użyciem tych mechanizmów zapisano obszerną gramatykę i wydają się one dobrze odpowiadać potrzebie formułowania reguł w sposób naturalny, unikając namnażania ich liczby. Zastosowanie pokazanych mechanizmów pozwala zachować elegancką odpowiedniość reguł gramatyki użytych do wyprowadzenia danego wypowiedzenia i wierzchołków drzewa składniowego. Ceną jest jednak zauważalna komplikacja zapisu reguł.

Jak wspomniano wcześniej, dla uproszczenia implementacji przyjęto, że operator **sequence_of** ani **optional** nie może wystąpić jako pierwszy składnik reguły. Ograniczenie to nie jest dotkliwe, ponieważ pierwszy linearnie składnik często pełni funkcję specjalną, wymaga swoistego przetwarzania atrybutów i w związku z tym są inne powody do rozpisania osobnych reguł ze względu na pierwszy składnik. Dopuszczenie sekwencji jako pierwszego składnika być może zmniejszyłoby nieco liczbę reguł kosztem skomplikowania warunków obliczających wartości atrybutów.

Pewnych problemów nastręcza konwencja dotycząca zasięgu zmiennych występujących w sekwencjach. Obowiązująca w Prologu reguła widoczności zmiennych w obrębie jednej klauzuli programu sprawia, że wymienienie zmiennej tylko w obrębie sekwencji oznacza, że ma się ona zunifikować we wszystkich wystąpieniach iterowanego elementu. Rzadko jest to zgodne z intencją twórcy reguł (choć zdarzają się takie wypadki, więc nie można tej możliwości całkiem wyeliminować). Częstszy jest wypadek, że wartość pewnego atrybutu jest nieistotna dla tworzonej jednostki, co w przyjętym mechanizmie wymaga specjalnego jej oznaczenia.

4.3. Styl pisania reguł

W GFJP Świdziński przyjął zasadę sztywnego klasyfikowania reguł gramatyki. Na przykład dla zdania elementarnego podane są osobno reguły dla realizacji pytajnych, niezależnych (neutralnych), pytajnozależnych, względnych, zależnych aglutynacyjnych, wreszcie innych realizowanych przez jednostkę wyższego poziomu (Świdziński 1992, s. 356–361). Pierwsze cztery typy mają te same składniki (kombinacje fraz wymaganych i luźnych i jedną frazę finitywną). Różnią się jedynie charakterystykami składników: aby konstrukcja była pytajnozależna lub względna, jej inicjalny składnik musi być odpowiednio pytajnozależny lub względny. Konstrukcja jest pytajna, jeżeli jakiś jej składnik jest pytajny. Zestawy reguł wymienionych w tych czterech grupach są poza tym identyczne. Analogicznie są klasyfikowane frazy werbalne, nominalne, przymiotnikowe i przysłówkowe.

Podobne powtarzające się grupy reguł pojawiają się dla pozostałych jednostek zdaniowych i fraz zdaniowych klasyfikowanych jako pytajne, niezależne i zależne różnych typów (zależność oznacza tu występowanie w specyficznym kontekście, np. pewnych spójników podrzędnych).

Poszczególnym grupom reguł towarzyszą przykłady ilustrujące dane zjawisko (np. typy zależności), co służy poglądowości wykładu.

Z inżynierskiego punktu widzenia taka organizacja gramatyki jest jednak niekorzystna. Powtarzające się reguły o bardzo podobnych składnikach nie służą efektywności automatycznej analizy składniowej. Sprawiają także, że reguły są trudniejsze w utrzymaniu, bo ewentualne zmiany trzeba wprowadzać w wielu miejscach analogicznie. Dlatego w prezentowanej tu gramatyce przyjęto, że nie powinny się powtarzać reguły złożone z tych samych jednostek składniowych (a więc różniące się tylko parametrami). Grupy reguł Świdzińskiego odpowiadają jednej regule Świgry 2 zawierającej odpowiednio rozbudowane warunki obliczające wartości parametrów (w szczególności uwzględniające kwestię pytajności i względności, por. p. 4.4.1). Dzięki takiej organizacji liczba reguł opisujących odpowiednie jednostki zmniejszyła się w stosunku do GFJP około czterokrotnie.

4.4. Parametry jednostek nieterminalnych

Repertuar parametrów jednostek składniowych został w analizatorze Świgra 2 w większości przejęty z GFJP. Część parametrów odpowiada kategoriom gramatycznym wprowadzonym na poziomie fleksyjnym. Są to: przypadek, rodzaj, liczba, osoba, stopień i aspekt. Wartości tych kategorii są na poziomie składniowym notowane zawsze jawnie, tak więc na przykład wszystkie frazy nominalne mają parametr osoby sygnalizujący, z jaką formą czasownika fraza taka uzgodni się na pozycji podmiotu. Dla form rzeczownikowych parametr ten jest ustalony na osobę trzecią (co na poziomie fleksyjnym nie jest notowane jako niezmienne dla wszystkich rzeczowników), dla zaimków osobowych ma wartości zgodne ze znacznikami fleksyjnymi. Podobnie wszystkim składniowym formom czasownikowym i frazom werbalnym jest przyznawana wartość rodzaju, mimo że dla form nieprzeszłych jest ona pomijana w znacznikach jako nieustalona (ukonkretnienie parametru składniowego nastąpi w wypadku uzgodnienia z podmiotem).

Na poziomie składniowym wprowadzane są kategorie czasu i trybu (są to parametry jednostki **formaczas**), które w znacznikach fleksyjnych reprezentowane są pośrednio. Tak więc formom nieprzeszłym fin (*czyta/przeczyta*) w zależności od aspektu przypisywany jest czas teraźniejszy ter lub przyszły przy (por. s. 44). Analityczne formy czasu przyszłego (*będziemy czytać/czytali*) oraz traktowane w przyjętym systemie znaczników jako analityczne formy trybu warunkowego (*czytał·by·ś*) otrzymują odpowiednie wartości tych cech.

Rozróżnienie między finitywnymi a niefinitywnymi formami czasowników jest kodowane za pomocą parametru wyróżnika fleksyjnego (Świdziński 1992, s. 85). Przyjmuje on następujące wartości: os – forma osobowa, bos – bezosobnik, bok – bezokolicznik, psu – forma imiesłowu przysłówkowego uprzedniego, psw – forma imiesłowu przysłówkowego współczesnego.

Wymienione cechy są określane na podstawie wartości przysługujących formom fleksyjnym, a następnie dziedziczone przez frazę od swojego reprezentanta. Z wartością wyróżnika fleksyjnego i czasu wiąże się komplikacja w wypadku nieciągłej realizacji czasu przyszłego (*będę to czytał/czytać*), trybu rozkazującego (*niech Piotr to czyta*) i trybu warunkowego (*byśmy to czytali*). W takiej sytuacji **zdanie** ma wśród swoich składników pomocniczą jednostkę **posiłk** stanowiącą formę posiłkową, czyli składnik formy nieciągłej (por. p. 2.13). Frazie finitywnej i frazie werbalnej w takim zdaniu przypisywany jest wyróżnik fleksyjny i czas wynikający z formy stojącej w obrębie tej frazy, a więc wartość osobowa i przeszła albo bezokolicznikowa i nieustalona. Obecność jednostki **posiłk** sprawia, że zdaniu jako całości przypisywane są inne wartości wynikające z rozpoznania formy analitycznej. Na przykład jeśli jednostka **posiłk** jest reprezentowana przez formę *będzie*, wartość wyróżnika fleksyjnego zdania zostanie obliczona jako osobowa os, a wartość czasu – jako przyszła przy.

Najważniejsze parametry czysto składniowe to: predestynacja, negacja, klasa, inkorporacja oraz wymagania i typ frazy wymaganej (por. GFJP, s. 87 i n. oraz

s. 320 i n.). W ramach rozwoju gramatyki wprowadzono także parametry pozycji, nadrzędności i przecinkowości. Parametry związane z wymaganiami i ich wysycaniem zostaną omówione w p. 4.5, a pozostałe – w kolejnych punktach bieżącego podrozdziału.

4.4.1. Predestynacja

Wśród parametrów składniowych występujących w GFJP najbardziej skomplikowaną interpretację miał parametr zależności (Świdziński 1992, s. 89). Był on pomyślany „od zewnątrz": jego wartości modelują „rozmaite uwarunkowania zewnętrzne poszczególnych jednostek zdaniowych", odpowiadają „kontekstom składniowym, które w różny sposób ograniczają charakterystykę gramatyczną jednostek składniowych w nich się pojawiających" (Świdziński 1992, s. 102).

Na przykład wartość chociaż zależności oznacza, że dana jednostka jest zdaniem podrzędnym wprowadzanym spójnikiem CHOCIAŻ, mimo że spójnik jest elementem zewnętrznym względem tego zdania. Wartość pz zależności oznacza, że dane zdanie jest zdaniem podrzędnym pytajnozależnym, co dla odmiany wynika z jego budowy wewnętrznej. Wartość pyt może zarówno oznaczać, że dana jednostka zawiera składnik pytajny, jak i że została umieszczona w pytajnym kontekście, co niekoniecznie ma takie same konsekwencje składniowe (por. Woliński 2004, s. 88). Pełny repertuar wartości parametru zależności można znaleźć w książce Świdzińskiego (1992, §5.1.4, s. 103–106).

W nowej gramatyce parametr ten został zarzucony. Zamiast tego wprowadzono parametr predestynacji *Dest* (Świdziński i Woliński 2009, s. 152) odróżniający tylko cztery klasy konstrukcji: neutralne neut, pytajne pyt, pytajnozależne pz i względne wz (z podtypami w zależności od klasy względnika). Inaczej niż w wypadku zależności klasyfikacja ma miejsce wyłącznie ze względu na budowę wewnętrzną konstrukcji, a nie ze względu na pełnioną funkcję (por. Woliński 2004, s. 106). Na przykład w wypowiedzeniu

(1) *Piotr przyszedł do domu?*

żaden ze składników zdania nie zawiera jawnego elementu pytajnego, dlatego zarówno składniki, jak i samo zdanie mają wartość neutralną predestynacji. Dopiero wypowiedzenie jako całość staje się pytajne w wyniku skonfrontowania neutralnego zdania ze znakiem końca o predestynacji pytajnej.

Predestynacja jest przypisywana na zasadzie słownikowej. Dla większości jednostek jest to wartość neutralna, ale formy pewnych leksemów wnoszą pytajność (np. DLACZEGO, CZY, KTÓRY, JAK, KTO), a pewnych leksemów – względność (np. KTÓRY, JAKI, KTO). Wartość pytajnozależna jest możliwa dla tych samych form co wartość pytajna; formy o takiej charakterystyce mogą realizować jedynie inicjalny składnik fraz.

Wartości predestynacji jednostek złożonych nie są transmitowane od ich centrów, jak to się dzieje z parametrami fleksyjnymi, ale obliczane dla każdej

jednostki na podstawie predestynacji jej składników. Do obliczenia predestynacji zdania służy warunek iteracyjny obldest_iter zdefiniowany w postaci następującego predykatu:

```
obldest_iter(neut, neut, neut).
obldest_iter(neut, pyt, pyt).
obldest_iter(pyt, neut, pyt).
obldest_iter(pyt, pyt, pyt).
obldest_iter(wz(Tfz, RL), neut, wz(Tfz, RL)).
obldest_iter(pz, neut, pz).
obldest_iter(pz, pyt, pz).
```

Krok obliczenia konfrontuje wynik poprzednich obliczeń (pierwszy argument) z predestynacją kolejnego składnika (drugi argument), dając nową wartość wynikową (trzeci argument). Same wartości neutralne prowadzą do neutralnej wartości wynikowej (pierwsza klauzula). Pojawienie się wartości pytajnej (druga klauzula) powoduje zmianę wyniku na pytajną. Jeżeli dotychczas obliczona wartość jest pytajna, kolejne składniki mogą być neutralne lub pytajne (trzecia i czwarta klauzula), zachowując pytajny wynik.

Aby dana jednostka była względna lub pytajnozależna, takiż musi być jej pierwszy składnik i tylko on. Dlatego te wartości nie są dopuszczalne jako drugi argument. Mogą się one pojawić tylko jako wartość początkowa iteracji. Dla predestynacji względnej kolejne składniki mogą być tylko neutralne (piąta klauzula). Predestynacja pytajnozależna dopuszcza kolejne składniki neutralne i pytajne (ostatnie dwie klauzule definicji predykatu).

Wartość względna predestynacji jest termem z dwoma argumentami. Jego pierwszy argument jest lematem względnika występującego w danej frazie, przyjmuje w szczególności wartości który, co, jaki, co pozwala odróżnić nieco inaczej zachowujące się jednostki. Drugi argument ma wartość przys dla względników przysłówkowych, a dla fraz wprowadzanych zaimkiem przymiotnym jest wartością liczby i rodzaju, które muszą w odpowiedniej regule zostać uzgodnione z nadrzędnikiem nominalnym.

(2) *Idź* *, dokąd poszli tamci* .

(3) *Znam chłopca* *, który/*która przyszedł* .

(4) **Znam chłopca* *, którzy przyszli* .

Typowe użycie warunku obldest_iter można prześledzić w przytoczonej na stronie 182 regule opisującej jedną z realizacji zdania elementarnego. Parametr predestynacji przysługuje wszystkim składnikom prawej strony reguły. Predestynacja pierwszej frazy wymaganej *Dest1* jest przekazywana jako wartość początkowa do warunku iterowanego w następującej dalej sekwencji. Predestynacja składników sekwencji *Dest2* (wszelkich typów) jest przekazywana do instancji warunku. Wartość wynikowa jest zapamiętywana w zmiennej *DestPreF*, która jest konfrontowana z predestynacją frazy finitywnej *Dest4*

przez wywołanie predykatu warunku poza sekwencjami, a wynik *DestPostF* zostaje przekazany do drugiej sekwencji. Wszystkie składniki w drugiej sekwencji wnoszą swoją predestynację *Dest5* poprzez warunek iterowany, a końcowy wynik obliczenia jest unifikowany ze zmienną *Dest*, występującą również po lewej stronie reguły jako wartość predestynacji całego zdania.

Inaczej obliczana jest predestynacja w konstrukcjach skoordynowanych. W takiej sytuacji dozwolona jest dowolna wartość predestynacji składników, musi ona jednak być taka sama we wszystkich składnikach (ściślej: wszystkich składnikach niespójnikowych; spójniki są pod tym względem neutralne). Wartość predestynacji frazy skoordynowanej jest równa tej wspólnej wartości predestynacji składników. Wyjątkiem są wartości pytajne, które mogą współwystępować z neutralnymi, predestynacja całości jest wtedy pytajna. Warto też zwrócić uwagę, że frazy z inicjalnym spójnikiem (a więc układ równorzędny I, p. 2.9.1, i szeregowy I, p. 2.9.2) nie mogą być względne ani pytajnozależne, ponieważ ich inicjalnym elementem jest neutralny spójnik. Kwestia ta jest odpowiednio uwzględniana w regułach opisujących te układy.

(5) *Wiem* , *kto przyszedł* , *dlaczego się spóźnił* i *kto jest temu winny* .

(6) *Przyszedł chłopiec* , *którego lubię* i *który to docenia* .

(7) **Przyszedł chłopiec* , i *którego lubię* , i *który to docenia* .

Atrybut predestynacji jest również wykorzystywany do przekazania informacji o obecności w zdaniu inicjalnej jednostki **posiłk** (por. p. 2.13). Mianowicie w takim zdaniu jego wartość predestynacji jest ustawiana na wartość specjalną agl. Wartość ta jest wymagana od zdania składnikowego w definicji fraz zdaniowych wprowadzanych spójnikami CHOĆBY, CZYŻBY, GDYBY, JAKBY, JAKOBY, ŻEBY. Jednocześnie wartość ta nie jest dopuszczana w warunkach obliczających predestynację zdań skoordynowanych, co blokuje koordynację zdań z inicjalnym aglutynantem typu

(8) **Wiedzielibyśmy, gdyby* *śmy przyszli* i *śmy przeczytali* .

4.4.2. NEGACJA

Parametr negacji reprezentuje negację zdaniową związaną z obecnością partykuły NIE przy formie czasownika (Świdziński 1992, s. 88). Parametr ten służy przede wszystkim do opisania zjawisk rozważanych w artykule Przepiórkowskiego i Świdzińskiego (1997), a więc dopełniacza negacji oraz tzw. rządu negacji związanego z tzw. zaimkami negatywnymi.

Wartość parametru negacji jest ustalana w regule, w której forma czasownikowa **formaczas** zamienia się we frazę werbalną **fwe**. Jeżeli forma czasownikowa jest poprzedzona partykułą NIE, fraza werbalna otrzymuje wartość negacji nie. W przeciwnym wypadku wartość ta jest ustalana jako tak. Parametr negacji jest dziedziczony przez bardziej skomplikowane frazy werbalne, frazę finitywną i zdanie (elementarne) zawsze od swojego centrum składniowego.

Wartość negacji frazy finitywnej jest narzucana wszystkim frazom wymaganym i luźnym w obrębie zdania, wpływając na sposób realizacji niektórych z nich. W szczególności następująca reguła gramatyki opisuje realizację frazy wymaganej typu np(accgen) przez frazę nominalną w bierniku lub dopełniaczu w zależności od wartości negacji (por. p. 3.1.4). Jest to tzw. dopełniacz negacji (por. Świdziński 1992, p. 6.4.5).

fw(np(accgen), *Lex*, *A*, *C*, *Rl*, *O*, *Poz*, *Neg*, *Dest*, *I*, *Pk*, *Sub***)** ⟶
fno(*H*, *P1*, *Rl1*, *O1*, wym([], _, _), *Kl*, *Zap*, pre, *Neg*, *Dest*, *I*, *Pk*, na**)**,
{ member(*P1*/*Neg*, [acc/tak, gen/nie]),
lex_dla_np(*Lex*, *H*, *Rl1*) **}**.

Występujący w regule predefiniowany w Prologu predykat member/2 sprawdza, czy jego pierwszy argument jest elementem listy będącej drugim argumentem. W tym wypadku pozwala on, by wartość przypadka *P1* frazy nominalnej była równa biernikowi, gdy wartość negacji jest tak, oraz dopełniaczowi dla negacji nie.

W analogiczny sposób opisany jest wpływ negacji na realizację fraz partytywnych (typ np(part), por. p. 3.1.7).

Parametr negacji przysługuje również większości innych jednostek gramatyki jak fraza nominalna, przymiotnikowa, przysłówkowa i przyimkowo-nominalna, ponieważ mogą one zawierać pewne jednostki leksykalnie wymuszające negatywność:

(9) *Nikt jej nie pomógł.*
(10) **Nikt jej pomógł.*
(11) *Maria nie dała niczego Piotrowi.*
(12) **Maria dała nic Piotrowi.*
(13) *Biografia syna żadnego prezydenta mnie nie zainteresowała.*
[Przepiórkowski i Świdziński (1997)]
(14) **Biografia syna żadnego prezydenta mnie zainteresowała.*

Na przykład parametr negacji izolowanej frazy nominalnej jest nieustalony, chyba że zawiera ona formę leksemu NIKT lub NIC, powodującą ustalenie negacji na nie. Gdy taka fraza nominalna staje się podrzędnikiem centrum werbalnego, jej wartość negacji zostaje zunifikowana z wartością negacji frazy werbalnej. W ten sposób fraza nominalna zawierająca zaimek negatywny nie może wystąpić w zdaniu z czasownikiem niezanegowanym, bo jej wartość negacji nie nie zunifikuje się z wartością tak pochodzącą od czasownika. Frazy nominalne, które nie zawierają elementu negatywnego, mogą znaleźć się w zasięgu oddziaływania zarówno czasowników niezanegowanych, jak i zanegowanych. Ich parametr negacji zostanie ukonkretniony zgodnie z negacją czasownika (jest to przykład jednej z niewielu sytuacji w tej gramatyce, gdy wartość parametru zależy od kontekstu zewnętrznego, a nie tylko od składu wewnętrznego jednostki; jest to użyteczny aspekt gramatyk unifikacyjnych).

Takie zachowanie negacji we frazach składnikowych uzyskuje się poprzez unifikację jej wartości we wszystkich składnikach (wszystkie wartości negacji w regule są reprezentowane tą samą zmienną prologową).

Wymaganie negatywności przenika przez granice fraz: negatywność frazy przymiotnikowej *żadnego* przeniesie się na frazę nominalną *żadnego miasta*, w której jest ona podrzędnikiem. Przeniesie się także na frazę przyimkową *do żadnego miasta*. Konstrukcje takie można dalej zagnieżdżać *pociąg do żadnego miasta* i wymaganie negatywności pozostaje aktywne. Granicę dla niego stanowi zdanie: zdania z negacją i bez negacji mogą być swobodnie łączone podrzędnie i współrzędnie (ale zob. dalej o konstrukcjach ze spójnikiem ANI).

W bardziej skomplikowany sposób zachowuje się negacja w zbitkach czasowników (ang. *verb clusters* Przepiórkowski i Świdziński 1997). Chodzi o zdania z wymaganą frazą bezokolicznikową, której wymaganiem może być kolejna fraza bezokolicznikowa itd. W takich konstrukcjach zjawiska związane z negacją są realizowane niekoniecznie lokalnie:

(15) *Jan chce* *zbudować dom*.
(16) *Jan nie chce* *zbudować żadnego domu*.
(17) *Jan chce* *nie zbudować żadnego domu*.
(18) **Jan nie chce* *zbudować (żaden) dom*.
(19) **Jan chce* *zbudować (żadnego) domu*.
(20) *Jan może* *nie chcieć próbować obiecać zbudować żadnego domu*.

W przykładach obecność negatywnej formy *żadnego* oraz realizacja np(accgen) przez frazę w dopełniaczu współwystępuje z negacją przy dowolnym z czasowników w zbitce.

W analizatorze Świgra 2 nielokalna negacja została zrealizowana za pomocą warunku używanego w regułach wprowadzających podrzędniki do frazy werbalnej (w tym bezokolicznikowej). Oto przykład jednej z tych reguł:

```
fwe(Wf, A, C, T, Rl, O, Wym, Neg, Dest, I, Pk, Sub) ⟶
    fwe(Wf, A, C, T, Rl, O, Wym1, Neg1, Dest1, I, Pk1, na),
    { wymagania_czyste(_, Wym1) },
    { wyklucz_podmiot_jeśli_niefinitywna(Wf, Wym1, WymN) },
    sequence_of([
        fw(W2, H2, A, C, Rl, O, post, Neg2, Dest2, ni, Pk2, po)
         ^[oblwym_iter, WymN, W2/H2, ReszteWym],
         fl(A, C, Rl, O, Neg2, Dest2, ni, Pk2, po)
         ^[najwyżej, 3, _, _]
        ]
         ^[obldest_iter, Dest1, Dest2, Dest]
         ^[sprawdz_pk_iter, Pk1, Pk2, Pk]
         ^[conajmniej1, 0, _, 1]),
```

```
{ wymagania_oznacz_użyte(ReszteWym, Wym),
obl_neg_fwe(Wf, Neg, Neg1, Neg2) }.
```

W obliczeniu negacji biorą udział cztery argumenty warunku obl_neg_fwe: wyróżnik fleksyjny frazy werbalnej, negacja frazy jako całości, negacja centrum werbalnego, negacja podrzędników (zunifikowana dla wszystkich podrzędników). Ponieważ analiza odbywa się wstępująco, negacja całej frazy jest obliczana na podstawie pozostałych argumentów. Jeżeli fraza werbalna ma jako centrum formę inną niż bezokolicznik, wszystkie argumenty reprezentujące negację są ze sobą unifikowane:

```
obl_neg_fwe(Wf, Neg, Neg, Neg) :— member(Wf, [os, bos, psu, psw]).
```

W wypadku fraz bezokolicznikowych obliczenie zależy przede wszystkim od tego, czy wśród podrzędników jest jakiś wymuszający konkretną wartość negacji. Jeżeli *Neg2* jest nieustalone, to znaczy, że podrzędniki są obojętne na negację i „w górę" będzie przekazany parametr negacji centrum werbalnego. (Dla ustalenia uwagi wartością tą zostaną też oznaczone podrzędniki).

```
obl_neg_fwe(bok, Neg, Neg, Neg2) :— var(Neg2),
    !,
    Neg = Neg2.
```

Ustalona wartość *Neg2* może pochodzić od leksykalnego „negatora" lub brać się z faktu wystąpienia frazy typu np(accgen). W szczególności jeżeli była ona realizowana przez frazę nominalną w bierniku, to negacja podrzędników będzie miała ustaloną wartość tak.

Dalsze obliczenie zależy od tego, czy centrum frazy werbalnej jest zanegowane. Gdy lokalny czasownik jest zaprzeczony, zaprzeczone muszą być również podrzędniki. Jako wartość negacji całości jest przekazywana wartość nie, jak gdyby fraza była zaprzeczonym czasownikiem bez podrzędników:

```
obl_neg_fwe(bok, nie, nie, nie).
```

Najtrudniejszy jest wypadek, gdy lokalny czasownik jest niezaprzeczony, a podrzędniki wskazują na konkretną wartość negacji. W takim wypadku wartością negacji całości stanie się wartość specjalna req(*Neg*) – od kontekstu wymagana jest konkretna wartość negacji.

```
obl_neg_fwe(bok, req(Neg), tak, Neg).
```

W regule opisującej realizację frazy wymaganej typu infp(_) wykrywana jest wartość negacji podrzędnika o postaci req(*Neg*) i tylko w takim wypadku wartość *Neg* jest unifikowana z negacją frazy wymaganej (a więc nadrzędnego czasownika). Dzięki temu fraza bezokolicznikowa bez podrzędników wymuszających konkretną wartość negacji (np. *kupić od Piotra*) może wystąpić w kontekście zarówno zanegowanego, jak i niezanegowanego czasownika nadrzędnego. Jednak fraza zawierająca taki podrzędnik (*kupić kurczę* albo *kupić*

kurczęcia) może wystąpić tylko w kontekście zgodnym co do negacji (odpowiednio *Chcę* albo *Nie chcę*).

Dodatkowe komplikacje pojawiają się przy obecności w wypowiedzeniu spójnika szeregowego ANI. Gdy spójnik ten spaja frazy, wymusza zaprzeczenie jednostki werbalnej, której jest podrzędnikiem, niekoniecznie bezpośrednim (por. Świdziński 2001).

(21) *Nie lubię Piotra ani Marii.*
(22) *Nie lubię czytać ani pisać.*
(23) *Nie lubię, gdy pada ani gdy grzmi.*
(24) *Wiem, że nie umarł ani że nie usechł.*
(25) **Wiem, że nie umarł ani że usechł.*
(26) *Wiem, że nie umarł, że nie był umęczon ani że nie został pogrzebion.*

W zdaniu szeregowym ze spójnikiem ANI zaprzeczone muszą być wszystkie łączone składniki. Zdania niezaprzeczone są niedopuszczalne nawet w zagnieżdżonych zdaniach skoordynowanych innymi spójnikami (Świdziński 1992, p. 5.3.4).

(27) *Nie pisał, nie czytał ani nie spał.*
(28) **Nie pisał, czytał ani nie spał.*
(29) *Ani nie pisze oraz nie czyta, ani nie śpi.*
(30) **Ani nie pisze oraz czyta, ani nie śpi.*

Parametr negacji przyjmuje w GFJP trzy wartości (por. Woliński 2004, s. 97). Oprócz wartości tak i nie jest to wartość ani – „zaprzeczoność odspójnikowa narzucona przez kontekst spójnika szeregowego ANI" (Świdziński 1992, s. 127). Jest ona potrzebna, by przekazywać informację o obecności spójnika ANI między jednostkami opisującymi szeregi (a więc zdaniami szeregowymi i zdaniami jednorodnymi). W analizatorze Świgra 2 możliwe było prostsze zrealizowanie tej zależności, ponieważ konstrukcja szeregowa jest realizowana w obrębie jednej reguły. Dzięki temu informacja o typie spójnika jest dostępna bezpośrednio i można na jej podstawie wymóc zanegowanie wszystkich składników zdaniowych.

W opisie negacji uwzględniono także kilka uwikłań natury leksykalnej. Po pierwsze, nietypowo zachowuje się przyimek BEZ. Fraza przyimkowo-nominalna z tym przyimkiem może wystąpić w zdaniu bez negacji, nawet gdy podrzędna fraza nominalna zawiera element negatywny (Przepiórkowski i Świdziński 1997):

(31) *Przepłynął Atlantyk bez żadnej pomocy.*
(32) *Nie prosił o żadną pomoc.*
(33) **Prosił o żadną pomoc.*

Fakt ten uwzględnia się prosto poprzez dodanie warunku w regule opisującej frazę przyimkową, który dopuszcza dowolną wartość negacji frazy nominalnej w wypadku przyimka BEZ.

Ograniczenia związane z negacją wykazują także niektóre jednostki funkcyjne, na przykład w zdaniu wprowadzanym spójnikiem DOPÓKI wykluczony jest aspekt dokonany przy braku negacji (Świdziński 1992, s. 105).

(34) *Nie pozwoliłem niczego ruszać, dopóki nie przyjedziesz.* [NKJP300]
(35) **Nie pozwoliłem niczego ruszać, dopóki przyjedziesz.*

4.4.3. Inkorporacja

Parametr inkorporacji, odziedziczony po GFJP, służy do opisu zdań ze spójnikiem inkorporacyjnym (zob. układ III konstrukcji równorzędnych w p. 2.9.1 i układ IV zdań prostych w p. 2.10).

(36) *– Znamy metody, dlaczego* *więc* *umieralność jest tak wysoka?* [Skł.]
(37) *Dziś byłoby to raczej niemożliwe, są* *natomiast* *inne sposoby, o czym świadczy los biblioteki PAN.* [Skł.]

Konstrukcje te są w istocie nieciągłe: spójnik inkorporacyjny pełni funkcję łączącą i może być postrzegany (w wypadku spójników inkorporacyjnych równorzędnych) jako centrum składniowe dla spajanych zdań. Aby opisać tę konstrukcję za pomocą formalizmu, który nie pozwala na wyrażenie nieciągłości, Świdziński wprowadził w prawie wszystkich jednostkach swojej gramatyki parametr inkorporacyjności, który sygnalizuje, czy dana jednostka zawiera w sobie spójnik inkorporacyjny (por. Świdziński 1992, p. 4.6.3). Przyjął przy tym założenie, że spójnik inkorporacyjny przyłącza się do poprzedzającego wyrazu składniowego, wnosząc do jego opisu cechę inkorporacyjności. Wymagało to zdublowania praktycznie wszystkich reguł opisujących jednostki preterminalne gramatyki, aby dopuścić obecność po dowolnej z nich spójnika inkorporacyjnego.

Według GFJP spójnik inkorporacyjny musi stać po pierwszym, niekoniecznie bezpośrednim, składniku zdania inkorporacyjnego (jak w przykładach (36) i (37)). Badania korpusowe pokazują, że o ile faktycznie pozycja ta wydaje się preferowana, to możliwe są także inne realizacje, np.:

(38) *W ten sposób mógł zarobić nawet 3,5 mln zł, z ustaleń gazety wynika* *bowiem*, *że chodzi o około 1500 osób.* [Skł.]

Dlatego w analizatorze Świgra 2, o ile ogólna koncepcja interpretacji spójników inkorporacyjnych została zachowana, to przyjęto, że spójnik inkorporacyjny może stać po dowolnym składniku konstrukcji (czyli że wykluczona jest wyłącznie inicjalna pozycja takiego spójnika). Nie oznacza to jednak, że pozycja linearna spójnika inkorporacyjnego jest zupełnie dowolna, nie może on bowiem znaleźć się we wnętrzu fraz zdaniowych (co zostaje zapewnione przez

regułę wymagającą, aby zdanie stanowiące składnik frazy zdaniowej miało wartość ni inkorporacji).

Wartością inkorporacyjności może być albo ni – nieinkorporacyjna, albo term i(*TI*, *OI*) sygnalizujący obecność we frazie spójnika inkorporacyjnego typu *TI* (r – równorzędny, p – podrzędny) i oznaczeniu *OI* (natomiast i zaś dla równorzędnych, bowiem i więc dla podrzędnych).

Obliczenie inkorporacyjności frazy odbywa się na podstawie inkorporacyjności jej składników za pomocą następującego warunku iterowanego:

oblink_iter(ni, ni, ni).
oblink_iter(ni, i(*TI*, *OI*), i(*TI*, *OI*)).
oblink_iter(i(*TI*, *OI*), ni, i(*TI*, *OI*)).

Przypisuje on wynikową wartość nieinkorporacyjną, jeżeli oba agregowane składniki mają taką wartość. Wynikową wartość inkorporacyjną uzyskuje się, gdy dowolny, ale tylko jeden ze składników ma wartość inkorporacyjną o pewnej charakterystyce. Gdy oba składniki są inkorporacyjne, warunek zawodzi.

Przetwarzanie parametru inkorporacji wymaga obecności odpowiednich warunków w większości reguł gramatyki. Informacja o obecności spójnika jest transmitowana przez poziomy drzewa. Poziom komplikacji tego mechanizmu może wydawać się nieadekwatny w stosunku do potrzeby, którą realizuje: opisu prostego typu nieciągłości. Jest to cena za brak wsparcia dla konstrukcji nieciągłych w stosowanym formalizmie gramatycznym. Warto jednak zauważyć, że ewentualny mechanizm wprowadzania nieciągłości musiałby pozwalać na opisanie dopuszczalnych miejsc wystąpienia spójnika inkorporacyjnego, trudno więc powiedzieć, czy byłby dużo prostszy od tu przedstawionego.

4.4.4. Klasa

Parametr klasy (Świdziński 1992, s. 90) pełni w GFJP dwie funkcje. Pierwszą z nich jest sygnalizowanie obecności jako centrum frazy pewnych typów form wyrazowych, które nadają frazom specyficzne właściwości. Są to: zaimki typu co i kto, zaimki osobowe, zaimki względne. Centra tych klas ograniczają możliwe podrzędniki we frazie nominalnej. Wartości przym i rzecz klasy pozwalają odróżnić frazy nominalne z centrum przymiotnikowym i rzeczownikowym.

Drugą funkcją tego parametru jest sygnalizowanie (wartością klasy tk) obecności we frazie, niekoniecznie jako składnika bezpośredniego, formy przymiotnika TAKI lub przysłówka TAK. Fraza nominalna tej klasy może zawierać jako składnik frazę zdaniową typu jakby, jaki, że lub żeby (Świdziński 1992, §7.4.2.3).

W gramatyce Świgra 2 ze względu na konieczność uwzględnienia zleksykalizowanych schematów składniowych (por. p. 3.1.13) wprowadzono nowy parametr podający lemat formy stanowiącej centrum każdej frazy. Dostępność tego parametru pozwala zrealizować dokładniej pierwszą funkcję parametru klasy – wskazywanie obecności specyficznych zaimków w centrach fraz.

4.4.5. Przecinkowość

W GFJP znaki interpunkcyjne są traktowane jako pełnoprawne składniki wypowiedzenia. Przecinki w zdaniu polskim powodują trudność w ujęciu regułowym, ponieważ w wypadku zbiegu przecinków, które powinny się pojawić z różnych powodów, następuje „uwspólnienie" ich funkcji i przecinek występuje tylko raz. Przecinek jest też pomijany na początku wypowiedzenia, a na końcu wypowiedzenia jego funkcję przejmuje końcowy znak interpunkcyjny.

Na przykład elementy wtrącone lub luźne (w szczególności podrzędniki zdaniowe takie jak *ponieważ Jan przyszedł*) są wydzielane ze zdania głównego przecinkami. Jeżeli jednak taka fraza występuje na końcu zdania, znak interpunkcyjny kończący zdanie pełni jednocześnie funkcję przecinka zamykającego frazę. Podobnie dzieje się z przecinkiem otwierającym frazę luźną na początku wypowiedzenia. Wreszcie przecinek pełniący funkcję spójnika współrzędnego pochłania ewentualne przecinki otwierające lub zamykające wtrącenie w tym samym miejscu.

GFJP zawiera na okoliczność przecinków dwie reguły kontekstowe pozwalające rozpoznać ciąg pusty jako przecinek w kontekście innego znaku interpunkcyjnego lub początku wypowiedzenia. W pierwotnej implementacji GFJP zastosowano inną sztuczkę techniczną: struktura reprezentująca tekst do analizy była rozszerzana o alternatywę fleksyjną zawierającą dodatkowe przecinki (Woliński 2004). Analogiczne rozwiązanie zastosowano w gramatyce POLFIE (Krasnowska-Kieraś i Patejuk 2015, przypis 4). Taki opis jest jednak jedynie przybliżeniem, pozwala bowiem na akceptację zdań zawierających nadmiarowe przecinki w odpowiednich miejscach.

W analizatorze Świgra 2 zastosowano inne rozwiązanie. Wszystkie jednostki gramatyki (z wyjątkiem pojedynczych form fleksyjnych) wyposażono w parametr przecinkowości, będący parą wartości (reprezentowaną za pomocą funktora /) zdających sprawę ze statusu danej jednostki składniowej względem przecinków umiejscowionych odpowiednio na początku i końcu zasięgu jednostki.

Podstawowymi wartościami przecinkowości są następujące symbole: `p` – sygnalizuje, że w danej pozycji jest przecinek; `bp` – sygnalizuje, że przecinka nie ma, ale jest konieczny do pełności frazy (oczekiwanie przecinka); `0` – przecinka nie ma, ale nie ma i oczekiwania przecinka (neutralność).

Jednostka **przec** ma wartość przecinkowości `p/p`. Frazy, które nie zawierają przecinka ani na początku, ani na końcu, typowo mają przecinkowość `0/0`. Dla każdej frazy wymagającej „ujęcia w przecinki" gramatyka przewiduje realizacje wariantowe. Tak więc fraza *, ponieważ chciał,* (z przecinkami na obu końcach) ma wartość przecinkowości `p/p`. Fraza *ponieważ chciał,* – z pominiętym przecinkiem początkowym – wartość `bp/p`. Analogicznie dla frazy z pominiętym przecinkiem końcowym uzyskuje się wartość `p/bp`, a z pominiętymi oboma przecinkami – `bp/bp`.

Schemat działania tego parametru jest następujący: dla każdej pary stojących obok siebie elementów prawej strony reguły sprawdza się, czy przecinkowość końcowa danego składnika jest zgodna z przecinkowością początkową następnego składnika. Wynikowa przecinkowość jednostki po lewej stronie reguły jest parą złożoną z przecinkowości początkowej pierwszego składnika i przecinkowości końcowej ostatniego.

Zasady uzgadniania przecinkowości między sąsiadującymi frazami są następujące: nie mogą spotkać się dwa przecinki, czyli wartości p przecinkowości końcowej pewnej frazy i przecinkowości początkowej frazy następnej. „Luka na przecinek" musi zostać wypełniona, czyli wartość bp musi spotkać się z wartością p. Wartość neutralna `0` może spotkać się z neutralną lub p. Dla uniknięcia niepotrzebnego powielania drzew przyjęto, że przy spotkaniu p i bp luka powinna być na końcu frazy poprzedzającej, a przecinek na początku frazy następnej (we wszystkich regułach produkujących lukę jest też wariant z jawnym przecinkiem w tym miejscu).

Przedstawione zasady można ująć w postaci następującej definicji predykatu wyrażającego zgodność wartości przecinkowości:

```
zgodne_pk(p, 0).
zgodne_pk(bp, p).
zgodne_pk(0, 0).
zgodne_pk(0, p).
```

Niestety zasady te nie obejmują sytuacji, gdy pominięty powinien zostać przecinek na początku wypowiedzenia. Na tę okoliczność wprowadzono wartość specjalną wp (wymuszony przecinek), która przypisywana jest kontekstowi początku wypowiedzenia. Wartość ta nie może spotkać się z jawnym przecinkiem, wymusza więc pominięcie przecinka początkowego następującej po niej frazy:

```
zgodne_pk(wp, bp).
zgodne_pk(wp, 0).
```

Komplikacje w przetwarzaniu wartości przecinkowości są też związane z częstym pomijaniem przez piszących przecinków okalających zdania podrzędne. Aby dopuścić akceptowanie niektórych takich zdań, wprowadzono kolejne wartości specjalne parametru przecinkowości, a poziomem restryktywności gramatyki można sterować, włączając odpowiednie klauzule predykatu zgodne_pk. Szczegóły można znaleźć w kodzie źródłowym.

Można powiedzieć, że bieżąca wersja gramatyki bardzo precyzyjnie opisuje interakcje dotyczące przecinków w tekście polskim. Ceną jest jednak komplikacja reguł gramatyki. Elegancki mechanizm powinien pozwalać na zapisanie ogólnej zasady postaci: gdy dwa składniki następują linearnie po sobie, zawsze sprawdź zgodność ze sobą ich wartości przecinkowości. W obecnej implementacji konieczne było wprowadzenie warunków sprawdzających ową zgodność do prawie wszystkich reguł gramatyki.

4.4.6. Nadrzędność

Parametr nadrzędności (oznaczany w regułach *Sub*) służy do reprezentowania centrów składniowych. Przysługuje on wszystkim jednostkom składniowym z wyjątkiem jednostki **wypowiedzenie**. W typowej regule (zob. przykład na stronie 182) wartość tego parametru jest nieustalona po lewej stronie reguły, natomiast wartości ustalone mają wszystkie składniki prawej strony. Wartość na oznacza centrum składniowe konstrukcji (choć nazwa sugerowałaby nadrzędnik). Wartość ta występuje tylko raz w obrębie reguły[9]. Pozostałe składniki są oznaczane symbolem po.

Jak z tego wynika, parametr ten wyłamuje się z ogólnej zasady, że wartości atrybutów jednostki wynikają z jej składu wewnętrznego, a nie kontekstu użycia. Wartość tego parametru zależy wyłącznie od funkcji, jaką dana fraza pełni w strukturze frazy bezpośrednio obejmującej. Warto zwrócić uwagę, że parametr ten nie jest przedmiotem żadnych uzgodnień ani warunków. Służy jedynie oznaczeniu centrów w strukturze.

Wartość tego parametru nie jest wypisywana wśród atrybutów danego węzła w drzewie składniowym. Jest ona interpretowana jako typ krawędzi wiążącej węzeł z nadrzędnikiem, a więc w zasadzie jako atrybut owej krawędzi. W wizualizacji przyjętej w tej pracy krawędź odpowiadająca wartości na jest przedstawiana z wyróżniającym szarym tłem.

4.4.7. Pozycja

Parametr pozycji przysługuje frazom wymaganym **fw** i frazom nominalnym **fno**. Służy on do reprezentowania położenia wymaganej frazy nominalnej względem centrum finitywnego. Wartość post sygnalizuje frazę wymaganą stojącą po frazie finitywnej (i jej centra nominalne wszystkich poziomów). Wartością pre są oznaczane wszystkie inne frazy wymagane i nominalne.

Potrzeba przetwarzania tego atrybutu jest związana z nietypowym uzgodnieniem rodzaju i liczby skoordynowanych fraz nominalnych. Jak opisano w p. 2.9.3, pewne takie frazy są możliwe tylko w postpozycji względem czasownika, a pewne – w prepozycji.

Parametr pozycji jest wykorzystywany także (w powiązaniu z parametrem przecinkowości) do wymuszenia, aby wymagana mowa niezależna ($Tf\omega$ = or) była linearnie pierwszym lub ostatnim elementem wypowiedzenia.

---

[9] W starszych wersjach gramatyki wprowadzono koncepcję oznaczania wszystkich składników wyrazów składniowych jako nadrzędników, aby zasygnalizować ich atomowość z punktu widzenia składni. Pomysł ten utrudnia jednak konwersję na drzewa zależnościowe i dlatego został zarzucony. Takie oznaczenie może wystąpić w starszych drzewach Składnicy frazowej.

4.5. Sposób realizacji wymagań

Mechanizm realizacji wymagań składniowych jest jednym z bardziej skomplikowanych elementów przedstawianej tu gramatyki, a uwzględnienie koordynacji fraz różnych typów przewidzianej w słowniku Walenty skomplikowało go dodatkowo (Woliński 2015).

W GFJP wymagania składniowe czasowników były reprezentowane w postaci trzech parametrów (co pozwalało opisać do trzech wymagań danej jednostki). Reprezentacja taka jest zupełnie niepraktyczna w wypadku użycia słownika Walenty, którego schematy często są dłuższe. W analizatorze Świgra 2 przyjęto, że parametr rekcja odpowiednich jednostek będzie zawierał listę wymagań faktycznie zrealizowanych przy danym predykacie.

Inaczej niż w GFJP, w której wymagania składniowe były rozpatrywane jedynie w odniesieniu do czasowników (ściślej fraz werbalnych), w analizatorze Świgra 2 walencję wprowadzono także do fraz rzeczownikowych, przymiotnikowych i przysłówkowych. Wysycanie wymagań następuje w regułach opisujących następujące jednostki: zdanie (elementarne), fraza werbalna niefinitywna (bezokolicznikowa lub z imiesłowem przysłówkowym), fraza nominalna (także z centrum gerundialnym), fraza przymiotnikowa (także z centrum imiesłowowym, w szczególności realizującym konstrukcję bierną), fraza przysłówkowa.

Ponieważ w GFJP informacja o konkretnym lemacie predykatu jest dostępna tylko w jednostce terminalnej, w implementacji gramatyki pobranie informacji ze słownika walencyjnego musiało odbywać się na tym poziomie, a następnie komplet informacji walencyjnych był przekazywany przez kolejne poziomy drzewa do miejsca, w którym następuje wysycanie wymagań. Ze względów historycznych mechanizm ten pozostał niezmieniony, chociaż wprowadzenie informacji o centrach leksykalnych, które nastąpiło w analizatorze Świgra 2, można by wykorzystać do pobierania informacji ze słownika dopiero w regule, w której konieczne jest wykorzystanie tej informacji.

Schematy walencyjne notowane w słowniku Walenty są maksymalne w tym sensie, że nie notuje się schematów, które byłyby podzbiorami innych schematów obecnych w słowniku. Jednocześnie większość argumentów notowanych w słowniku jest opcjonalna – niepomijalne są jedynie argumenty zleksykalizowane oraz wymaganie partykuły się. W związku z tym konieczny jest mechanizm, który dopuści częściowe realizowanie schematów walencyjnych w zdaniu.

Rozwiązaniem z pewnego punktu widzenia najprostszym byłoby wyliczenie „z góry" wszystkich dopuszczalnych podzbiorów schematów walencyjnych i przekazywanie do analizy takich skróconych schematów. Wadą tego rozwiązania jest znaczne zwiększenie liczby alternatywnych schematów, które muszą zostać przetestowane przy analizie danego zdania. Mediana długości schematu w Walentym wynosi 3, więc typowo pojedynczy schemat zostałby rozpisany na 8 schematów „nieeliptycznych". Jednak czasowniki zwykle mają w Walentym kilka schematów, co prowadzi do mediany liczby schematów

równej 33 (przy pomijaniu podschematów powtarzających się w różnych schematach). Maksymalna liczba wygenerowanych podschematów wynosiłaby 813, jest ona osiągana dla czasownika DAĆ. Przedstawione oszacowanie nadal nie jest jednak miarodajne, ponieważ częściej używane czasowniki mają schematy bardziej skomplikowane od czasowników rzadkich. Jeżeli uwzględnić częstość tekstową poszczególnych czasowników, to średnia liczba podschematów na czasownik wyniesie 76 (oszacowanie na korpusie Składnica, por. Woliński 2015).

Takie rozwiązanie jest stosowane w gramatyce POLFIE (Patejuk 2015), ponieważ, jak się wydaje, jest to jedyne możliwe rozwiązanie w obrębie systemu XLE. Informacje pochodzące ze schematu walencyjnego muszą bowiem zostać zapisane w haśle leksykonu w postaci w zasadzie gotowych fragmentów reprezentacji odpowiednich argumentów w generowanych strukturach, muszą więc uwzględniać konkretną listę argumentów.

Taka konieczność nie zachodzi w wypadku stosowania formalizmu DCG dzięki możliwości użycia w warunkach dowolnych predykatów zdefiniowanych w Prologu, w szczególności dowolnego przetwarzania zbioru schematów walencyjnych. W analizatorze Świgra 2 hasła słownika Walenty są przechowywane w postaci zbliżonej do oryginalnej i przetwarzane dopiero w momencie wysycania poszczególnych wymagań predykatu.

4.5.1. Wypełnianie pozycji składniowych

Proces ten odbywa się tak samo dla wszystkich typów predykatów uwzględnionych w Walentym (czasowników, rzeczowników, przymiotników i przysłówków). Algorytm operuje na dwóch listach. Pierwsza z nich zbiera już zrealizowane argumenty, a więc służy budowaniu częściowo zrealizowanego schematu składniowego; na zakończenie przetwarzania stanie się ona atrybutem rekcja odpowiedniego wierzchołka w drzewie. Druga lista zawiera części schematów, które mogą jeszcze zostać zrealizowane. Pierwsza lista jest inicjowana jako pusta, druga na początku przetwarzania zawiera wszystkie schematy walencyjne dla danego predykatu. Schemat należy rozumieć jako listę pozycji składniowych, które z kolei są listami typów fraz wymaganych.

Gdy analizator skonstruuje frazę kandydującą do wypełnienia pozycji składniowej, czyli gdy rozpozna jednostkę **fw** z pewnym typem frazy wymaganej *Tfw*, wykonywane są następujące operacje:

1. Z listy schematów wybierane są te, które mają pozycje zawierające argument danego typu *Tfw*.
2. Z każdego z tych schematów usuwana jest pozycja zawierająca typ argumentu *Tfw*.
3. Uzyskana w ten sposób lista staje się nową wartością pozostałych do zrealizowania schematów, a typ *Tfw* jest dodawany do listy argumentów już zrealizowanych.

Warto podkreślić, że pozycje są traktowane dyzjunktywnie: jeżeli zostanie zrealizowany argument zawarty w pewnej pozycji, to ta pozycja jako całość zostaje uznana za zrealizowaną.

Działania te są wykonywane dla każdej frazy wymaganej **fw** rozpoznanej przez analizator. Po rozpoznaniu wszystkich fraz wymaganych lista pozostałych do wykorzystania schematów jest sprawdzana na obecność niezrealizowanych argumentów obowiązkowych. Schematy zawierające takie argumenty są usuwane z listy. Rozpoznana konstrukcja jest akceptowana przez analizator, jeżeli lista schematów pozostała niepusta, czyli jeżeli był choć jeden schemat, w którym wszystkie obowiązkowe argumenty zostały zrealizowane.

4.5.2. Specyfikacje argumentów składniowych

Specyfikacje typów fraz wymaganych są co do zasady zapisane identycznie jak w słowniku Walenty, są to więc termy takie jak np(dat), cp(int) lub prepncp(przeciw, dat, żeby) (por. p. 3.1.2). Termy te stają się wartością typu frazy wymaganej *Tfw* stanowiącego pierwszy parametr fraz wymaganych **fw**.

Jak wspomniano w punkcie 3.1.13, argumenty zleksykalizowane w słowniku Walenty mają ogólną postać lex(*Tfw*, *Cechy_gramatyczne*, *Lemat*, *Modyfikatory*). Ten zestaw informacji jest też przechowywany w listach schematów składniowych. Jednak we frazie wymaganej realizującej takie wymaganie informacje są rozbite na typ *Tfw* i informacje dodatkowe [*Cechy_gramatyczne*, *Lemat*] (ten zbiorczy paramter nazwywany jest *Lex*; postać *Cech_gramatycznych* zależy od typu frazy). Dzięki temu nie są tworzone osobne frazy, jeżeli nie ma pewności, czy dana fraza ma być interpretowana jako realizująca schemat składniowy zwykły czy zleksykalizowany. W tym pierwszym wypadku opis wymagania jest tylko jego typem *Tfw* i on jest konfrontowany z odpowiadającym parametrem frazy wymaganej. W wypadku argumentu zleksykalizowanego oba elementy są konfrontowane z zapisem ze słownika. W bieżącej wersji analizatora nie są przetwarzane *Modyfikatory*.

Schematy z wymaganiem partykuły się (oznaczanym symbolem sie) są rozpatrywane łącznie ze schematami bez się. Jeżeli predykat jest niefinitywną formą czasownika, schematy podlegają przekształceniu zgodnie z opisem w p. 3.1.10. Argument sie jest obowiązkowy przy formach finitywnych czasowników, ale staje się opcjonalny przy odsłownikach. Przy imiesłowach przeszłych argument sie nie może być realizowany. Odsłowniki zachowują się w tej kwestii nieregularnie, przy niektórych się wydaje się obowiązkowe, jednak przy innych może być pominięte lub wręcz jest trudne do pomyślenia. Uwarunkowania te nie są modelowane w słowniku Walenty.

4.5.3. Przykład realizacji wymagań

Dla przykładu przedstawiona zostanie analiza argumentów czasownika chcieć w zdaniu:

(39) *Jan chce, żeby dać mu spokój.*

Dla uproszczenia wypisano jedynie kilka spośród schematów dla tego czasownika i pominięto informacje o kontroli składniowej. Początkowa lista schematów dla finitywnej formy tego czasownika ma postać:

```
[ % schemat 1
  [[sie], [np(dat)], [infp(_)]],
  % schemat 2
  [subj([np(nom)]),
   [np(accgen), cp(żeby), infp(_), advp(misc)]],
  % schemat 3
  [subj([np(nom)]),
   [np(gen), cp(żeby), ncp(gen,żeby)],
   [prepnp(od,gen)]]
]
```

Analizator Świgra 2, przetwarzając przykład (39) od lewej do prawej, natrafia najpierw na frazę nominalną w mianowniku *Jan*. Jest ona dobrym kandydatem na podmiot, jej rodzaj, liczba i osoba zgadzają się z formą czasownikową. Zaakceptowanie tego argumentu spowoduje odrzucenie schematu 1, jako że nie dopuszcza on podmiotu. Następnie z pozostałych schematów zostanie usunięta pozycja podmiotowa, w wyniku czego powstanie następująca lista:

```
[% schemat 2:
  [[np(accgen), cp(żeby), infp(_), advp(misc)]],
 % schemat 3:
  [[np(gen), cp(żeby), ncp(gen,żeby)],
   [prepnp(od,gen)]]
]
```

Drugim kandydatem na argument rozpoznanym przez analizator jest fraza zdaniowa *, żeby dać mu spokój*. Jej centrum jest spójnik ŻEBY, może więc ona realizować argument typu `cp(żeby)`. Oba pozostałe schematy zawierają pozycję dopuszczającą argument tego typu. Żaden z nich nie zostaje więc odrzucony, a po uwzględnieniu wysycenia tej pozycji lista schematów przyjmuje postać:

```
[% schemat 2:
  [],
 % schemat 3:
  [[prepnp(od,gen)]]
]
```

Analizator nie wykrywa więcej kandydatów na argumenty, przechodzi więc do sprawdzenia warunków końcowych. Następuje sprawdzenie obecności

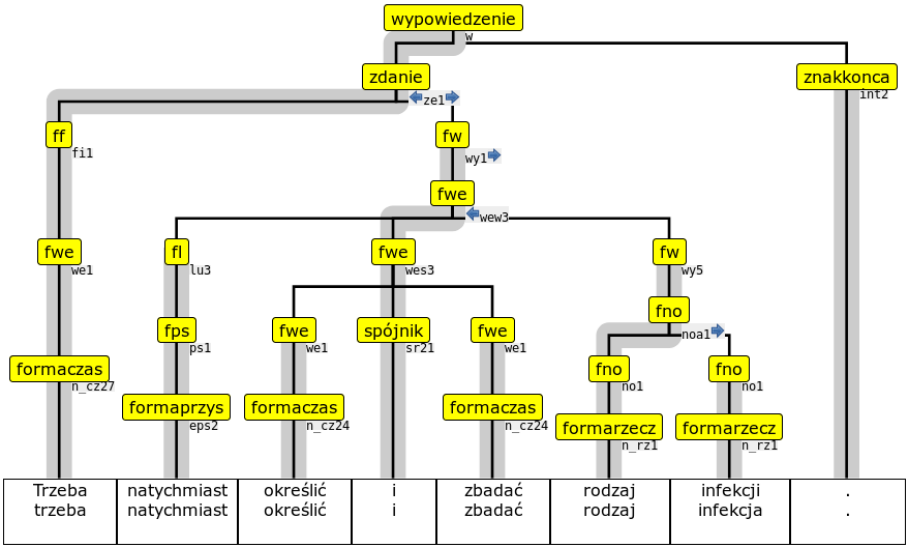
niezrealizowanych argumentów niepomijalnych. W wyjściowych schematach jedynym argumentem obowiązkowym było zwrotne *się*, jednak schemat zawierający je został pominięty. Pozostałe fragmenty schematu 2 i 3 nie zawierają argumentów obowiązkowych. Lista pozostałych schematów pozostaje niepusta, a więc przetwarzanie kończy się sukcesem.

Warto zauważyć, że zbiór rozpoznanych argumentów czasownika pasuje zarówno do schematu 2 (zrealizowanego w całości), jak i 3 (który dopuszczałby jeszcze frazę przyimkowo-nominalną OD z dopełniaczem). Analizator nie rozstrzyga, który ze schematów został użyty, i tworzy tylko jedno drzewo składniowe odpowiadające temu zestawowi argumentów. Oczywiście nie jest to jedyny rozbiór tego zdania, fraza zdaniowa typu żeby może także być luźna, jak w zdaniu *Jan czyta, żeby dać mu spokój.*; dalsze niejednoznaczności składniowe kryją się w zdaniu podrzędnym.

4.5.4. Wymagania a konstrukcje z koordynacją

Typowe konstrukcje skoordynowane, w których połączenie współrzędne następuje wewnątrz frazy danego typu nie wymagają szczególnego traktowania, gdy pełnią funkcję frazy wymaganej, np. skoordynowana fraza nominalna w pozycji wymaganej zachowuje się jak inne frazy nominalne.

Szczególnej uwagi wymaga jednak sytuacja, gdy to frazy będące źródłem wymagań składniowych podlegają koordynacji (por. p. 2.9.5). Rysunek 4.1 przedstawia drzewo dla zdania, w którym dwie formy czasownikowe dzielą się luźnym podrzędnikiem *natychmiast* oraz frazą wymaganą typu np(accgen)



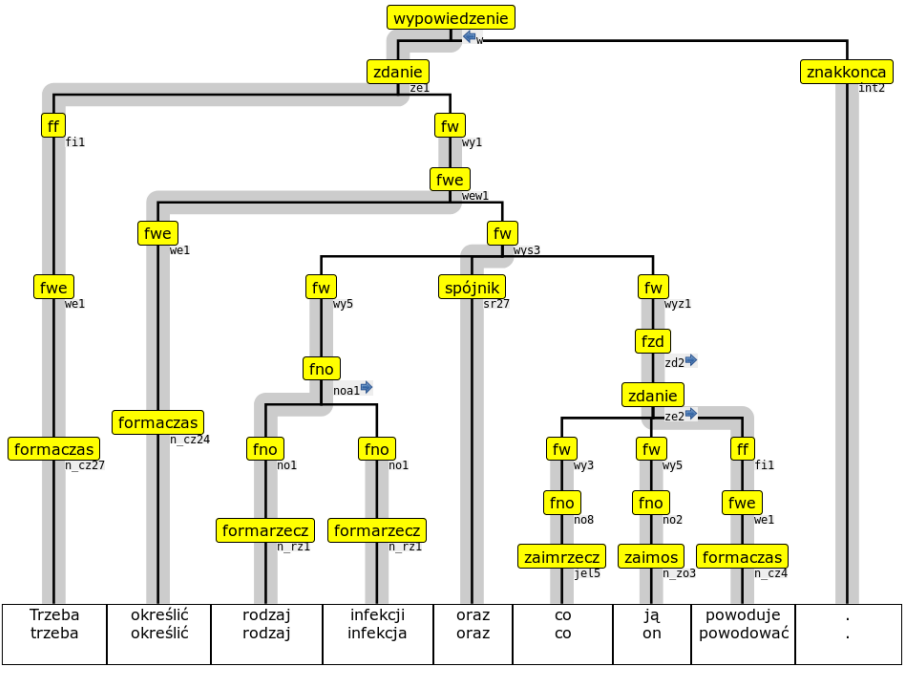
Rysunek 4.1. Uwspólnienie podrzędników przez dwie formy czasownikowe

rodzaj infekcji. W trakcie tworzenia tego drzewa analizator buduje dla frazy *określić i zbadać* specjalną reprezentację schematów składniowych zawierającą listy dopuszczalnych schematów dla obu czasowników. W momencie wysycania wymagań tej frazy skoordynowanej procedura opisana w punkcie 4.5.1 będzie wykonana na reprezentacji schematów każdego z czasowników z osobna, a powodzenie odniesie, jeżeli można ją było wykonać dla wszystkich.

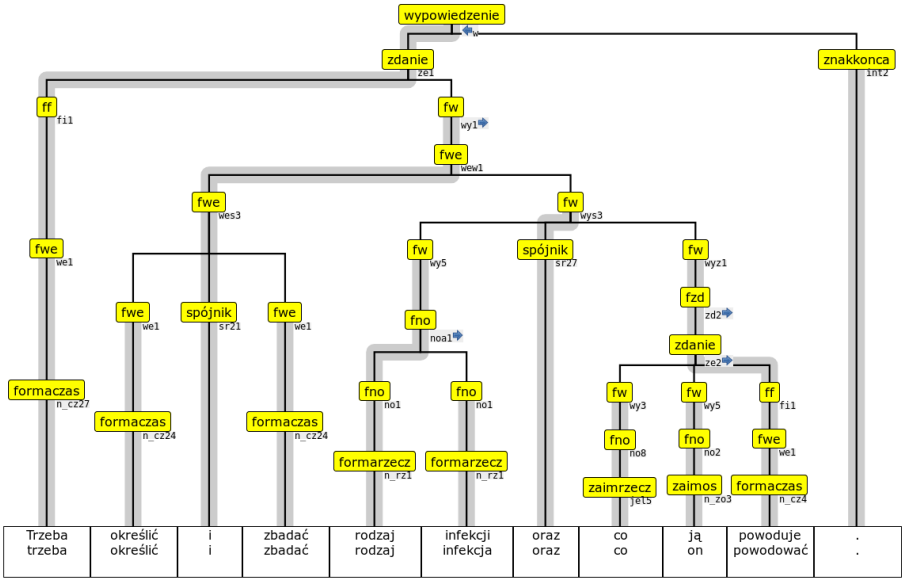
Drugim zjawiskiem wymagającym specjalnego przetwarzania jest uwzględniona w słowniku Walenty możliwość koordynacji w obrębie jednej pozycji składniowej fraz różnych typów (ang. *unlike coordination*, por. p. 2.9.4 i 3.1.3). W tym celu do gramatyki dodano reguły pozwalające realizować schematy konstrukcji współrzędnych omówione w p. 2.9 z frazą wymaganą **fw** jako składnikiem konstrukcji.

Rysunek 4.2 przedstawia przykład zdania, w którym spójnikiem współrzędnym połączono frazę nominalną *rodzaj infekcji* (typu np(accgen)) z frazą zdaniową pytajnozależną cp(int) *co ją powoduje*. Frazy te zostają zinterpretowane jako wymagane **fw**, a następnie podlegają koordynacji. Współrzędna fraza wymagana *rodzaj infekcji oraz co ją powoduje* otrzymuje jako swój typ listę typów składników: [np(accgen), cp(int)].

Analizator przy próbie zaakceptowania takiego argumentu składniowego sprawdza, czy istnieje w jakimś schemacie pozycja zawierająca wszystkie wy-



Rysunek 4.2. Struktura zawierająca wymaganą frazę nominalną skoordynowaną z frazą zdaniową pytajnozależną



Rysunek 4.3. Struktura łącząca zjawiska przedstawione na rysunku 4.1 i 4.2

mienione typy. Tak się faktycznie dzieje dla jednego ze schematów możliwych dla czasownika OKREŚLIĆ:

subj	obj		
np(nom)	np(accgen) cp(int) cp(że) ncp(accgen,int) ncp(accgen,że)	prepnp(w,loc)	xp(mod)

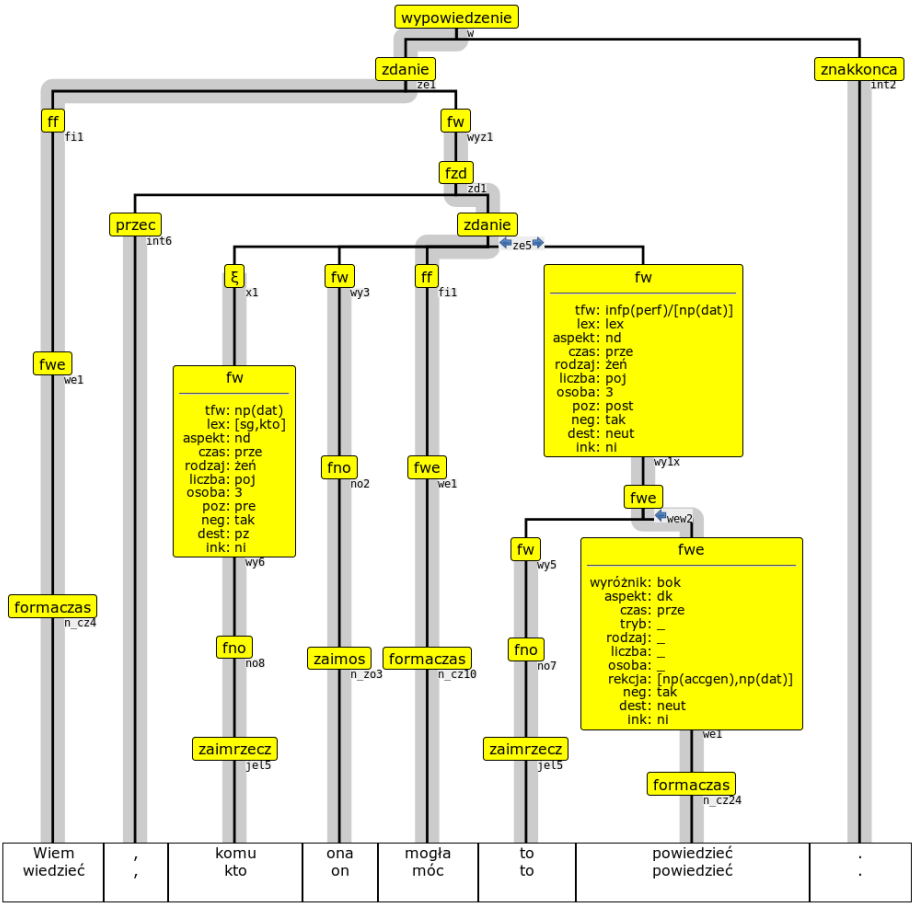
Oba omówione zjawiska mogą wystąpić wspólnie, co zilustrowano na rysunku 4.3.

4.5.5. Wyjęcie wymagania poza frazę

Ostatnim elementem mechanizmu wysycania wymagań w analizatorze Świgra 2 jest implementacja fraz nieciągłych omówionych w punkcie 2.14. Na rysunku 4.4 przedstawiono przykład takiej analizy dla zdania:

(40) *Wiem, komu ona mogła to powiedzieć.*

W zdaniu tym forma *wiem* wymaga następującej po przecinku frazy zdaniowej pytajnozależnej. Forma *mogła* ma zrealizowany podmiot mianownikowy *ona* i wymaganą frazę bezokolicznikową nieciągłą *komu [...] to powiedzieć*. Fraza wymagana *to powiedzieć* ma typ określony jako infp(perf)/[np(dat)], aby zasygnalizować jej niepełność – brak podrzędnika w celowniku. Warunki w regule za-



Rysunek 4.4. Układ zależności w obrębie zdania z frazą bezokolicznikową nieciągłą

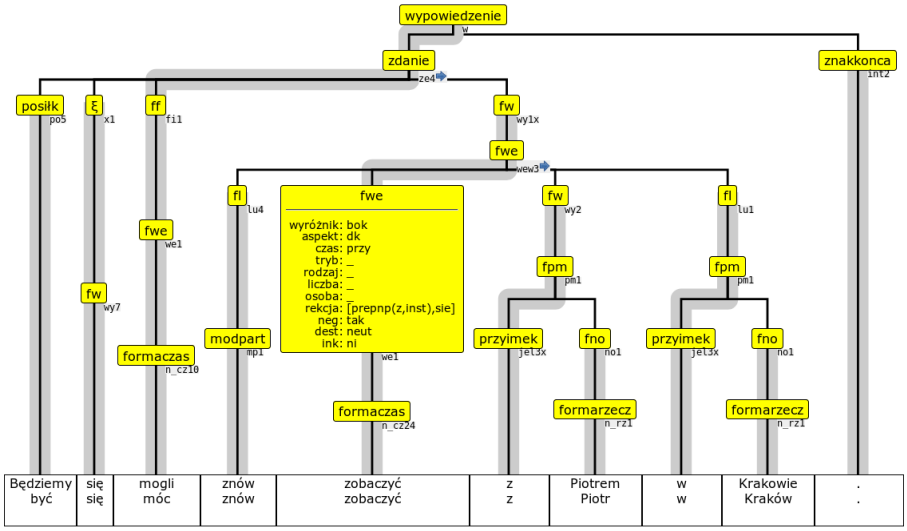
pewniają, że jednostka o takim typie może zostać użyta tylko wtedy, gdy wśród składników jest również jednostka ξ (zob. p. 2.14) z typem równym części po ukośniku. Formie *powiedzieć* przypisano dwa wymagania: frazy np(accgen) lokalnie zrealizowanej przez *to* i np(dat) zrealizowanej poprzez frazę ξ stojącą wyżej w strukturze zdania.

Warto zwrócić uwagę, że fraza pytajnozależna musi mieć inicjalny składnik pytajny, w związku z czym nieciągłość w tym zdaniu jest niejako obowiązkowa, niemożliwe jest zdanie **Wiem, ona mogła komu to powiedzieć.*

„Oderwane" frazy wymagane są dopuszczane przez gramatykę wyłącznie na poziomie zdania (elementarnego). Argument może zostać oderwany od fraz wymaganych typów infp(*Asp*), adjp(pred) i np(*P*).

(41) *Metody te* *można* *podzielić na dwie grupy*. [Skł.]

(42) *Ubogiej oświacie* *chcieli* *zabrać 200 tysięcy*. [Skł.]



Rysunek 4.5. Zdanie z typową nieciągłością: partykuła SIĘ przynależna do czasownika ZOBACZYĆ została wyniesiona przed nadrzędny czasownik MÓC

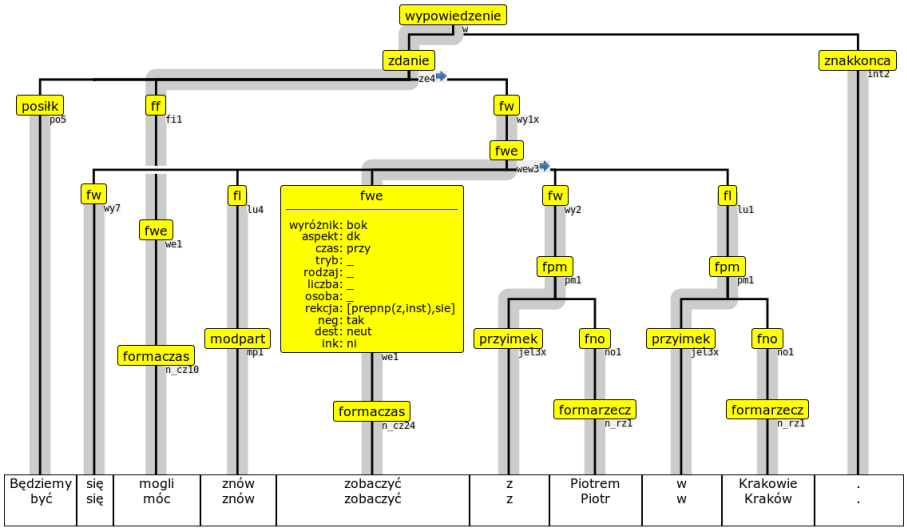
(43) *Jeszcze do niedawna proastmin dostępny był bez recepty.* [Skł.]

(44) *Od zwracanego cła wyrównawczego nie płaci się odsetek.* [Skł.]

Reguły gramatyki pozwalają, aby „oderwaniu" podległa co najwyżej jedna fraza wymagana przez dany predykat (por. Maier i Lichte 2011). Ponadto przyjęto, że jeżeli fraza, w której brakuje składnika wymaganego, stoi po czasowniku głównym w zdaniu, to oderwana fraza wymagana musi stać przed nim. Możliwa też jest konfiguracja odwrotna, zawsze jednak czasownik główny oddziela frazę oderwaną od frazy z luką. Badania korpusowe wskazują na zasadność takiego ograniczenia. Przyjęto, że nie wykonuje się takiej operacji dla fraz luźnych, co oznacza, że frazy luźne bywają przyłączane w drzewie wyżej, niż by wskazywała intuicja.

Aby uniknąć powstania bardzo wielu hipotetycznych fraz z brakującym argumentem, reguły gramatyki tworzą jedną frazę typu na przykład infp(perf)/*Wym*, gdzie *Wym* zawiera całą informację o niewykorzystanych częściach schematów walencyjnych frazy bezokolicznikowej (czyli drugą z list utrzymywanych przez algorytm z punktu 4.5.1). Na poziomie zdania lista ta zostanie skonfrontowana z typem frazy wymaganej rozpoznanej tam jednostki ξ, a w wynikowym drzewie w miejscu *Wym* pojawi się typ konkretnej frazy.

Jak wspomniano poprzednio, jednostka ξ stanowi w istocie jedynie znacznik dla osób znakujących korpus składniowy, że dana fraza wymagana jest elementem nietypowym i nie przynależy do danego poziomu drzewa. Konstrukcję taką zaproponowano w związku z tym, że nieciągłości stanowią pro-



Rysunek 4.6. Drzewo z krzyżującymi się gałęziami, któremu w istocie odpowiada notacja przedstawiona na rysunku 4.5

blem dla formalizmów składnikowych. Przedstawione rozwiązanie generuje więc struktury ciągłe (por. rys. 4.5), które jednak mogą zostać automatycznie skonwertowane na faktycznie nieciągłe (rys. 4.6).

4.6. Prezentacja wyników analizy

Reprezentacja jednostek nieterminalnych w postaci termów prologowych jest mało czytelna, ponieważ tożsamość poszczególnych atrybutów jest kodowana ich pozycją w termie. Staje się to szczególnie uciążliwe, gdy jednostki gramatyki mają wiele atrybutów (np. jednostka **zdanie** ma 11 atrybutów). Dlatego w wynikach analizatora (kodowanych w XML) i w wizualizacji (widocznej na rysunkach w tej książce) jednostki nieterminalne są przekształcane na listę par atrybut–wartość. Nazwa jednostki nieterminalnej jest kodowana jako atrybut o nazwie cat (*category*, kategoria składniowa).

Wykonywanych jest także kilka innych przekształceń atrybutów wierzchołków na potrzeby prezentacji: ukrywana jest złożona reprezentacja schematów walencyjnych odpowiadających danym jednostkom; pokazywane są tylko typy fraz zrealizowanych w danej frazie. Wartość req(*Neg*) (por. p. 4.4.2) jest zastępowana *Neg*. Atrybut przecinkowości jest pomijany. Powoduje to komplikację polegającą na tym, że drzewa różniące się przecinkowością poszczególnych węzłów mogą stać się identyczne po jej pominięciu. Dlatego w przetwarzaniu uwzględniono dodatkową fazę usuwania z lasu powtórzonych (pod)drzew.

Drzewa zapisywane są w XML-u (z użyciem poręcznej biblioteki dostępnej w SWI Prologu) w postaci upakowanego (dzielonego) lasu (por. p. 6.3). Stosowany jest prosty schemat XML stworzony *ad hoc*. Odzwierciedla on bezpośrednio strukturę lasu: podawana jest lista wierzchołków, a dla każdego z nich lista ich alternatywnych realizacji.

Format ten stanowi narzędzie wymiany danych między analizatorem Świgra 2 a systemem Dendrarium (p. 6.5) i automatycznymi narzędziami ujednoznaczniającymi (zob. rozdz. 7), a także służy do udostępniania korpusu Składnica (zob. rozdz. 6).

5

Inne formalne opisy składni języka polskiego

W poprzednich rozdziałach przedstawiono koncepcję opisu formalnego polszczyzny oraz naszkicowano sposób jej technicznej realizacji. W tym rozdziale krótko przedstawiono inne formalne opisy składni polszczyzny – ich zakres i sposób wyrażenia, a więc użyte formalizmy. Szczególną uwagę poświęcono formalizmom HPSG i LFG, które różnią się istotnie od formalizmu stosowanego w tej pracy, co może być interesujące dla czytelnika.

Przedmiotem zainteresowania są opisy o charakterze językoznawczym i ujmujące w założeniu język ogólny. Pominięto więc skądinąd interesujące prace, takie jak gramatyka systemu POLINT (Vetulani 2004; Vetulani *et al.* 2010); gramatyka systemu tłumaczenia maszynowego POLENG (Jassem 2006); gramatyki tworzone z użyciem systemu parsowania powierzchniowego Spejd (Przepiórkowski 2008).

5.1. Opisy składnikowe

5.1.1. Gramatyka Szpakowicza

Najstarszym opisem należącym do nurtu, z którego wywodzi się gramatyka prezentowana w tej pracy, jest praca doktorska Stanisława Szpakowicza (1978), opublikowana następnie w postaci książkowej (Szpakowicz 1983). Przedstawia ona dystrybucyjny powierzchniowy opis składniowy z nastawieniem na pragmatyczny cel jego przydatności do implementacji komputerowej.

Przedmiotem opisu są wyłącznie zdania, a więc konstrukcje z centrum finitywnym oraz konstrukcje współrzędne o składnikach będących tak rozumianymi zdaniami. Autor zastrzega, że nie są rozważane konstrukcje eliptyczne, mowa niezależna ani wtrącenia (w tym wołaczowe). Opis uwzględnia tylko konstrukcje ciągłe o nienacechowanym szyku wyrazów (jak zaznacza autor, jakie warianty szyku zaliczyć do tej grupy, jest decyzją arbitralną). Zakłada się, że analizowane wypowiedzenia muszą być poprawne interpunkcyjnie. Mimo że gramatyka Szpakowicza jest wcześniejsza od GFJP Świdzińskiego, zawiera ona reguły opisujące konstrukcje współrzędne również w obrębie fraz rzeczownikowych, przymiotnikowych i przysłówkowych oraz konstrukcje z liczebnikami. Podobnie jak w GFJP nie są uwzględnione konstrukcje apozycyjne.

Gramatyka jest zapisana w formalizmie gramatyk metamorficznych (Colmerauer 1978). Stosowane jest specjalne oznaczenie sygnalizujące dopuszczalność dowolnej permutacji danego ciągu składników, nie ma jednak mechanizmu pomijania lub powielania składników.

W zasadzie stosowany jest więc ten sam formalizm i styl opisywania jednostek nieterminalnych co w późniejszej GFJP (oczywiście oznacza to, że GFJP w naturalny sposób wzoruje się na gramatyce Szpakowicza). Jednostki nieterminalne tworzą rozbudowane hierarchie (np. zdanie, zdanie złożone, zdanie szeregowe, zdanie pojedyncze, zdanie elementarne, zdanie ograniczone), w których obecne są cykle. Parametry jednostek nieterminalnych są mniej liczne niż w GFJP, w szczególności niektóre jednostki zdaniowe nie mają parametrów. Autor zauważa, że dla wysubtelnienia przedstawianego opisu konieczne byłoby zwiększenie liczby parametrów.

W eksperymentach komputerowych Szpakowicz operował około połową reguł gramatyki przedstawionej w książce, słownik był ograniczony do kilkudziesięciu form fleksyjnych. Na potrzeby pracy doktorskiej wygenerował drzewa składniowe dla kilkudziesięciu zdań, co stanowiło osiągnięcie przy możliwościach dostępnych ówcześnie komputerów. Próbę implementacji pełnej gramatyki Szpakowicza podjął Bień w końcówce lat 90., nie doprowadziła ona jednak do zbudowania programu o zadowalającej efektywności działania (zob. Bień 2009).

5.1.2. Gramatyka Świdzińskiego (GFJP)

Gramatyka Świdzińskiego została przedstawiona w pracy habilitacyjnej (Świdziński 1987), opublikowanej następnie w postaci książki (Świdziński 1992). Ponieważ jest ona zasadniczym punktem odniesienia dla gramatyki Świgra 2, wiele jej cech zostało już zreferowanych we wcześniejszych rozdziałach. Świdziński następująco charakteryzuje zakres GFJP (Świdziński 1992, s. 20):

> Niniejszy opis obejmuje praktycznie wszystkie typy konstrukcji składniowych dzisiejszej polszczyzny. Co więcej, uwzględniam w nim wiele nie dostrzeganych przez gramatyków cech składniowych definiowanych konstrukcji, jak na przykład uzgodnienie aspektu, czasu, trybu, wymaganie składniowe, negacja, typ zdania podrzędnego, typ spójnika itp. Opisuję konstrukcje pomijane w opracowaniach gramatycznych, takie jak zdania złożone ze spójnikiem nieciągłym, zdania złożone pytajne i rozkazujące, zdania pytajne z kilkoma zaimkami pytajnymi itp.

W stosowanym formalizmie GFJP ściśle odpowiada opisowi Szpakowicza i jest rozwinięciem wcześniejszych wspólnych prac Szpakowicza i Świdzińskiego. Rozpatrywane jednostki składniowe są termami złożonymi z nazwy jednostki i parametrów. Wartości parametrów są w GFJP atomowe. Stosowane są reguły przepisywania z warunkami. Reguły można by zapisać w postaci bezkontek-

stowej z wyjątkiem trzech reguł kontekstowych dotyczących pomijania przecinków.

Prace dążące do implementacji GFJP podjął Bień (1997, 1996). Pierwszą próbę stanowił analizator AMOS opracowany w latach 1994–1996, wykorzystujący dostępną w interpreterach Prologu implementację DCG ze zstępującą strategią analizy. Program był realizowany za pomocą implementacji MIM-Prolog w systemie DOS i analizatora fleksyjnego SAM-95 (Szafran 1993, 1996). W gramatyce zablokowano za pomocą dodanych warunków reguły lewostronnie rekurencyjne, okazało się też konieczne opóźnianie obliczania pewnych warunków za pomocą prologowego mechanizmu freeze. Program cechował się bardzo małą efektywnością, udało się jedynie przeanalizować kilkuwyrazowe zdania przykładowe.

W latach 1997–1999 został zaimplementowany analizator AS (Bień *et al.* 2001), w którym strategia analizy została zmieniona na wstępującą z zapamiętywaniem wyników pośrednich. Zreorganizowano też kluczowe parametry GFJP (w szczególności parametr zależności był reprezentowany w postaci listy wartości). Program realizowano w SWI-Prologu, kłopotliwa pozostawała analiza fleksyjna prowadzona w preprocesingu.

Ostateczną formą implementacji GFJP był analizator Świgra 1. Więcej szczegółów na jego temat można znaleźć w pracach Wolińskiego (2004, 2005).

Ogrodniczuk (2005, 2006) zintegrował GFJP z bardziej rozbudowanym opisem fraz nominalnych (Szpakowicz i Świdziński 1990) i dodał konstrukcje liczebnikowe.

5.2. Opis w formalizmie HPSG

Na przełomie tysiącleci podjęto w Instytucie Podstaw Informatyki prace nad wykorzystaniem do opisu języka polskiego formalizmu Head-driven Phrase Structure Grammar (HPSG, Pollard i Sag 1987, 1994). Cykl powstałych wówczas prac doktorskich obejmuje: opis konstrukcji względnych języka polskiego (Mykowiecka 1999), teorię nadawania wartości przypadka (Przepiórkowski 1999), analizę zachowania klityk, w tym partykuły *się* (Kupść 2000) i opracowanie zagadnienia koreferencji zaimków (Marciniak 2001). Zwieńczenie cyklu stanowi praca włączająca poprzednie do ogólnej gramatyki języka polskiego w przyjętym formalizmie (Przepiórkowski *et al.* 2002).

Podstawę formalną HPSG może stanowić logika typowanych struktur atrybutów (RSRL), dla której można udowodnić podobne twierdzenia o poprawności i pełności wywodu jak dla programowania w logice. Sytuacja formalna jest więc podobna jak w wypadku gramatyk DCG, ale formalizm zapewnia system silnego typowania struktur.

HPSG należy do formalizmów gramatycznych opartych na operacji unifikacji (podobnie jak DCG). W odróżnieniu jednak od wcześniej omawianych

formalizmów gramatyka HPSG nie opisuje struktur frazowych za pomocą reguł przepisywania. Opis języka składa się z sygnatury oraz teorii. Sygnatura definiuje hierarchię typów struktur z wielodziedziczeniem oraz przysługujące poszczególnym typom atrybuty (dla każdego atrybutu określony jest typ struktury, która może być jego wartością). Teoria jest zbiorem ograniczeń (zwanych zasadami), które muszą spełniać poprawnie zbudowane struktury (ograniczenia te tłumaczą się na formuły RSRL; ich częścią mogą być wywołania predykatów zdefiniowanych w sposób bardzo bliski programowaniu w logice, choć operujących na typowanych strukturach zamiast na termach).

HPSG jest formalizmem i zarazem teorią językoznawczą, a więc kompleksem sposobów modelowania poszczególnych zjawisk językowych w postaci struktur atrybutów i zasad ich interakcji.

W teorii językoznawczej HPSG przyjmuje się, że struktury reprezentujące składniki danej frazy są wartościami pewnych atrybutów struktury reprezentującej całą frazę. W podstawowej wersji teorii dla każdej struktury typu *phrase* wartością atrybutu HD-DTR jest nadrzędnik konstrukcji, a wartością atrybutu NONHD-DTRS – lista podrzędników. Drzewiasta struktura składników jest więc ukryta wśród atrybutów budowanych struktur.

Powszechnie przyjmowaną zasadą, od której bierze się nazwa HPSG, jest Zasada Elementu Głównego, która mówi mniej więcej, że każda struktura reprezentująca frazę dziedziczy swoją istotną charakterystykę od struktury będącej wartością atrybutu HD-DTR, reprezentującej centrum składniowe konstrukcji (*head*).

Przyjęte w HPSG powiązanie struktur z analizowanym wypowiedzeniem polega na tym, że wszystkie struktury należące do typów odpowiadających formom fleksyjnym i konstrukcjom mają atrybut PHON, którego wartością jest lista segmentów (ogólniej: reprezentacja fonologiczna). Wypowiedzenie należy do języka opisywanego przez daną teorię HPSG, jeżeli istnieje struktura, która spełnia wszystkie ograniczenia należące do teorii i której wartością atrybutu PHON jest owo wypowiedzenie. Relacja, która musi zachodzić między wartościami tego atrybutu dla frazy i jej składników, jest wyrażona odpowiednimi ograniczeniami. W pracy (Przepiórkowski *et al.* 2002) przyjęto, że relacje te zaniedbują porządek linearny składników, ale nie ich ciągłość.

Teoria HPSG jest zleksykalizowana, co oznacza, że własności składniowe poszczególnych form fleksyjnych są zapisywane w słowniku poprzez przypisanie im odpowiednich struktur atrybutów. Można to osiągnąć, definiując podtypy typu *word* odpowiadającego formie fleksyjnej. Korzystając z wielodziedziczenia, można na przykład powiedzieć, że typ *powiedział* jest podtypem typu *trzecioosobowy-przeszlik-męski-pojedynczy* (którego atrybutami są odpowiednie wartości osoby, rodzaju oraz liczby i który jest podtypem typu *przeszlik* reprezentującego finitywne centrum zdania w czasie przeszłym) i jednocześnie typu *leksem-powiedzieć*, którego atrybutem jest m.in. struktura reprezentująca schemat walencyjny dla form tego czasownika. Inna organizacja słownika polegałaby na posłużeniu się jednym typem *word* i odpowiednim układem ograniczeń.

Z powyższych rozważań wynika, że za pomocą formalizmu HPSG można budować bardzo różne teorie, w szczególności niekoniecznie zgodne z HPSG jako teorią językoznawczą. Ta ogólność powoduje jednak problemy z efektywną implementacją. Jedną z przyczyn jest to, że atrybut PHON wiążący struktury z przetwarzanym tekstem nie jest w żaden sposób wyróżniony, trudno więc optymalizować przetwarzanie tak, aby następstwo segmentów w tekście efektywnie sterowało procesem analizy zdania, co ma miejsce w parserach dla poprzednio omawianych gramatyk opartych na regułach przepisywania.

Teoria przedstawiona w pracy (Przepiórkowski *et al.* 2002) została zaimplementowana w systemie ALE (The Attribute Logic Engine). Jak jednak piszą autorzy, nie zaimplementowano wszystkich szczegółowych reguł i zasad, a jedynie wykazano poprawność i spójność podstawowych elementów teorii. Implementacja nie została nigdy poddana weryfikacji korpusowej.

Wyraźną zaletą formalizmu HPSG jest modularność opisu: mechanizm ograniczeń pozwala formułować warunki dotyczące każdej struktury danego typu. Dzięki temu, dokładając ograniczenia, można rozszerzać gramatykę o nowe mechanizmy. To zaś oznacza, że mogą istnieć obok siebie w zasadzie niezależne mechanizmy realizujące różne warstwy opisu – na przykład składnię i semantykę.

Oparcie formalizmu na strukturach atrybutów daje większą czytelność niż posługiwanie się termami logicznymi w DCG. Jednak struktury HPSG dla rozbudowanych konstrukcji językowych szybko stają się bardzo skomplikowane, więc ukrycie w nich struktury frazowej można widzieć jako czynnik zmniejszający czytelność i łatwość interpretacji struktur. Wadą formalizmu jest też trudność jego efektywnej implementacji.

Z inżynierskiego punktu widzenia silne typowanie jest mechanizmem chroniącym przed błędami, co ma znaczenie przy tworzeniu dużych gramatyk. Każdy atrybut każdej struktury musi zostać przewidziany przez sygnaturę gramatyki, nie można więc przypisać omyłkowo atrybutu strukturze, której nie powinien on przysługiwać. Każdy atrybut przewidziany przez sygnaturę musi mieć określoną wartość, aby struktura została uznana za zgodną z gramatyką, nie można więc omyłkowo pominąć nadania wartości.

5.3. Opis w formalizmie LFG

Kolejny opis języka polskiego (Patejuk i Przepiórkowski 2012) został opracowany w formalizmie Lexical Functional Grammar (LFG, Dalrymple 2001). Gramatyka POLFIE powstała poprzez adaptację reguł analizatora Świgra 2 oraz rozwiązań opracowanych na gruncie HPSG (Przepiórkowski *et al.* 2002). Gramatyka została zaimplementowana za pomocą platformy Xerox Linguistic Environment, do której w szczególności podłączono analizator fleksyjny Morfeusz (Krasnowska-Kieraś i Patejuk 2015). Pierwsze wersje gramatyki wykorzy-

stywały także słownik walencyjny analizatora Świgra 2 uzupełniony o schematy wyekstrahowane z korpusu składniowego Składnica. W kolejnych wersjach rolę tę przejął słownik Walenty. Gramatyka została rozbudowana także o wyrafinowany opis konstrukcji skoordynowanych, który stał się podstawą pracy doktorskiej (Patejuk 2015). Równolegle z gramatyką tworzony jest generowany za jej pomocą bank struktur LFG (Patejuk i Przepiórkowski 2014).

LFG to formalizm gramatyki unifikacyjnej oparty na regułach frazowych. Struktury przypisywane w LFG konstrukcjom językowym składają się z co najmniej dwóch warstw. Pierwsza z nich, c-struktura, czyli *constituent structure*, jest drzewem składników bezpośrednich etykietowanych kategoriami składniowymi. Druga – f-struktura, czyli *functional structure* – składa się ze struktur atrybutów (ang. *attribute-value matrices*). Każda ze struktur atrybutów jest przypisana do jakiegoś zbioru wierzchołków c-struktury. Do korzenia drzewa składnikowego jest przypisana f-struktura reprezentująca całe wypowiedzenie. Możliwe jest dodawanie do opisu kolejnych warstw, np. struktury semantycznej.

Opis języka w formalizmie LFG składa się z dwóch części: reguł gramatycznych i leksykonu. Reguły gramatyczne są regułami przepisywania operującymi atomowymi kategoriami składniowymi, są to więc reguły bezkontekstowe. Z poszczególnymi elementami prawej strony reguły są wiązane ograniczenia definiujące f-struktury. Oto przykład hipotetycznej reguły:

(1) $S \longrightarrow \underset{(\uparrow \text{SUBJ})=\downarrow}{N} \quad \underset{\uparrow=\downarrow}{V}$

Symbol ↓ oznacza f-strukturę związaną ze składnikiem, przy którym stoi; symbol ↑ – f-strukturę związaną z jednostką po lewej stronie reguły. Zatem równanie ↑=↓ określa, że struktura opisująca dany składnik ma być jednocześnie strukturą opisującą całą jednostkę. W wypadku przykładowej reguły charakterystyka czasownika staje się charakterystyką zdania. Drugie równanie, (↑ SUBJ)=↓, oznacza, że struktura reprezentująca dany składnik ma się stać wartością atrybutu SUBJ w strukturze nadrzędnej (a więc, że reprezentacja rzeczownika ma się stać wartością atrybutu SUBJ w reprezentacji zdania). W ograniczeniach można stosować ścieżki złożone z wielu atrybutów, na przykład specyfikacja (↑ SUBJ CASE)=NOM oznacza, że struktura ↑ musi mieć atrybut SUBJ, którego wartością jest struktura, której atrybut CASE ma wartość NOM. Możliwe jest także odwołanie się do struktury zawierającej daną, (SUBJ ↓) jest strukturą, w której dana struktura ↓ jest wartością atrybutu SUBJ. Możliwości specyfikowania ścieżek w strukturach atrybutów są dość rozbudowane. Ścieżki mogą być niedospecyfikowane poprzez zadanie wartości alternatywnych lub powtarzających się.

Leksykon składa się z uporządkowanych trójek obejmujących: segment, odpowiadającą mu część mowy, która zostanie użyta w regule frazowej, oraz zestaw ograniczeń specyfikujących odpowiadającą danej formie f-strukturę. Ograniczenia te są formułowane w ten sam sposób co w regułach przepisywania.

LFG, podobnie jak HPSG, jest formalizmem gramatycznym i jednocześnie teorią językoznawczą. Fundamentalnym pojęciem dla teorii językoznawczej LFG są funkcje gramatyczne. Zbiór przyjętych funkcji gramatycznych obejmuje: podmiot, dopełnienie bliższe i dalsze, kilka szczegółowych typów dopełnień, okolicznik. Składnikom konstrukcji składniowych przypisywane są funkcje gramatyczne. Technicznie oznacza to, że reprezentacja danego składnika staje się wartością atrybutu noszącego nazwę danej funkcji gramatycznej. W wypadku elementów, które mogą się powtarzać, np. okoliczników oraz składników konstrukcji współrzędnych, wartością odpowiedniego atrybutu staje się zbiór struktur opisujących poszczególne wystąpienia.

Podobnie jak w HPSG opis jest zleksykalizowany, więc elementy leksykonu zawierają informacje sterujące analizą składniową, np. schematy walencyjne. Jednak możliwe jest także umieszczanie odpowiednich ograniczeń w regułach składniowych. Oznacza to, że formalizm pozwala na wybór, które uwarunkowania lepiej umieścić w danym miejscu. Umieszczenie ograniczeń w słowniku powoduje, że informacja składniowa powtarza się w bardzo wielu elementach leksykonu. Dlatego twórca gramatyki operuje raczej na metapoziomie: posługuje się szablonami specyfikującymi zestawy ograniczeń, które aplikuje w poszczególnych pozycjach leksykonu.

LFG jako formalizm gramatyczny

LFG jest formalizmem opartym na regułach przepisywania, co ułatwia efektywną implementację. Jawne reprezentowanie drzewa składników bezpośrednich zapewnia czytelność struktury przypisywanej wypowiedzeniu. Jednocześnie zakłada się, że c-struktura i f-struktura są od siebie w zasadzie niezależne, co oznacza, że f-struktura powiązana z korzeniem drzewa składnikowego musi reprezentować całe wypowiedzenie łącznie z jego składnikami wszystkich poziomów. W f-strukturze jest więc powtarzana hierarchiczna struktura wypowiedzenia (przy czym może się ona różnić od tej zadanej przez c-strukturę). To oznacza, że f-struktury szybko stają się bardzo skomplikowane, podobnie jak w HPSG.

Struktury atrybutów są w LFG typowane dynamicznie. Jest to podejście diametralnie różne od DCG i HPSG, w których zestaw atrybutów struktury danego typu jest ustalony dla wszystkich instancji tego typu. W LFG zestaw atrybutów struktury zależy m.in. od opisu centrum tej struktury w leksykonie (gdzie podana jest lista atrybutów odpowiadających funkcjom gramatycznym przysługującym temu centrum). Oznacza to rezygnację z kontroli poprawności zapewnianej przez system typów.

Metoda opisywania f-struktur za pomocą równań towarzyszących regułom i elementom leksykonu jest bardzo elegancka. Jednak to, że ograniczenia są wypisywane w regułach, oznacza, że wymagane jest powtarzanie ich we wszystkich regułach opisujących struktury danego typu, względnie we wszystkich elementach leksykonu opisujących formy danej klasy. Praktycznie oznacza to

tworzenie skomplikowanego systemu zależących od siebie szablonów ograniczeń. Szablony stanowią w istocie rodzaj makr rozpisujących się do danych ograniczeń. To skutkuje koniecznością programowania systemu nieutypowionych makr, co jest błędogenne. Brak formalnie wyróżnionych typów struktur powoduje, że nie można formułować ograniczeń zależnych od typów, które dla odmiany stanowią bardzo nośny mechanizm w HPSG.

Praca Patejuk (2015) pokazuje, że formalizm LFG jest wystarczająco bogaty, by zaimplementować w nim gramatykę polszczyzny. Jednak pewne elementy opisu wymagały dość kreatywnego wykorzystania formalizmu. Dotyczy to opisu koordynacji fraz różnych typów za pomocą mechanizmu LFG nazywanego *off-path constraints*, który został użyty nie dlatego, że był to naturalny sposób odwołania się do pewnych elementów struktur, ale dlatego że tak sformułowane ograniczenia są interpretowane inaczej niż zapisane jako „zwykłe" ograniczenia.

Pewną niedogodnością LFG dla opisu sterowanego zawartością słownika walencyjnego (wspomnianą w p. 4.5) jest to, że jedynym sposobem wprowadzenia schematów składniowych do analizy jest umieszczenie ich rozpisanych wariantów w elementach leksykonu. Pod tym względem dużo bardziej elastyczny jest formalizm DCG, który pozwala odwołać się w warunkach umieszczonych w regułach do dowolnych zewnętrznych źródeł danych.

LFG jako teoria językoznawcza

Teoria LFG powstała na gruncie języka angielskiego i choć zastosowano ją do opisu wielu innych języków, to przyjęta w niej granica kompetencyjna między c-strukturą i f-strukturą wydaje się być obciążona cechami języka pozycyjnego o ubogiej fleksji.

W języku silnie fleksyjnym, gdzie widać duże zróżnicowanie i specjalizację form fleksyjnych, a w związku z tym znaczną zależność funkcji pełnionych przez formę w zdaniu od wartości jej kategorii gramatycznych, same kategorie składniowe stanowią bardzo zgrubny opis własności danej jednostki. Tak więc samo oznaczenie frazy rzeczownikowej NP nie determinuje, czy fraza ta może pełnić funkcję podmiotu. To zależy w szczególności od wartości przypadka. Tymczasem w języku angielskim ta sama forma rzeczownika może równie dobrze pełnić funkcję podmiotu, jak i dopełnienia bliższego. Tak więc o ile drzewa składnikowe dekorowane wyłącznie nazwami jednostek nieterminalnych w języku angielskim stanowią rozsądne przybliżenie struktury składniowej, to w języku takim jak polski przybliżenie to jest daleko niedoskonałe. Proponowany w LFG sposób rozdzielenia struktury składnikowej od funkcyjnej powoduje, że dla języka polskiego większość istotnej informacji składniowej znajduje się w f-strukturach.

Sugeruje to, że zaproponowane w LFG rozdzielenie c-struktury od f-struktury jest nieoptymalne dla języka fleksyjnego. Struktura modelująca zjawiska powierzchniowoskładniowe dla takiego języka powinna zawierać nie tylko

nazwy jednostek składniowych, ale i przysługujące im cechy formalnogramatyczne (tak dzieje się w etykietowanych drzewach składnikowych analizatora Świgra 2). Osobną warstwą mogłaby być dopiero struktura semantyczna, operująca np. rolami ze słownika Walenty.

Wątpliwości budzi też pojęcie funkcji gramatycznych. Powstaje bowiem pytanie, z jakiego poziomu jest to pojęcie i czemu służy jego obecność w strukturach składniowych. Pewną sugestię może dać refleksja nad następującą ramą walencyjną dla czasownika GIVE z projektu FrameNet (przytoczoną za Hajnicz *et al.* 2016a)[1]:

(2)	rola	*Donor*	*Theme*	*Recipient*
	typ frazy	NP	NP	NP
	funkcja gramatyczna	Ext	Comp	Obj

W przykładzie tym występują trzy wymagane frazy nominalne. Wprowadzenie funkcji gramatycznych ułatwia przypisanie im ról semantycznych poprzez składniowe odróżnienie tych fraz od siebie. W wypadku języka fleksyjnego z wyrazistym pojęciem przypadka sytuacja jest jednak inna. Oto analogiczna rama ze słownika Walenty dla czasownika DAĆ:

(3)	rola	*Initiator*	*Theme*	*Recipient*
	typ frazy	np(nom)	np(accgen)	np(dat)

W tym wypadku typy fraz użyte w schemacie składniowym są wystarczająco szczegółowe, by skojarzyć je z rolami semantycznymi. Ewentualne wprowadzenie funkcji gramatycznych nie wniesie więc nowej informacji.

Autorzy gramatyki POLFIE również postulują odejście od fundamentalnego dla LFG pojęcia funkcji gramatycznych (Patejuk i Przepiórkowski 2016). W ich pracy można znaleźć postulat pozostawienia jedynie funkcji SUBJ i OBJ odpowiadających pozycjom składniowym wyróżnionym w słowniku Walenty. Pozostałe podrzędniki proponują reprezentować w f-strukturach w postaci niezróżnicowanej listy.

5.4. Opisy zależnościowe

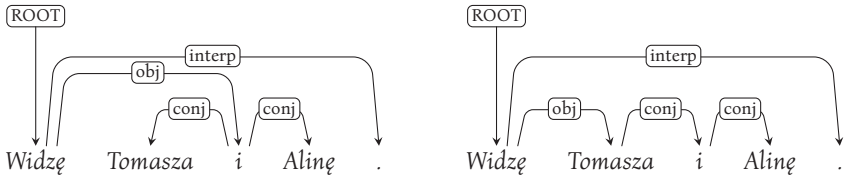
Nurt opisu zależnościowego jest wyraźnie obecny w polskim językoznawstwie. Reprezentuje go w szczególności gramatyka Klemensiewicza (1969), wykorzystywana w nauczaniu szkolnym. Wymienić warto także prace Świdzińskiego (1989) i Derwojedowej (2011).

[1] Przytoczone etykiety funkcji gramatycznych są inne niż w LFG, nie zmienia to jednak istoty argumentu.

Metodę automatycznej analizy zależnościowej opartej na regułach przedstawił Obrębski (2002). Opracował on efektywny algorytm analizy składniowej oraz przedstawił reguły opisujące nietrywialny podzbiór języka polskiego (obejmujący m.in. konstrukcje podrzędne, koordynację, apozycję, frazy względne z uwzględnieniem uzgodnień odległych). Analizator był testowany na zbiorze zdań przeznaczonym do testowania analizatorów języka polskiego (Marciniak *et al.* 2000).

Inne podejście, oparte na metodach maszynowego uczenia, reprezentuje praca doktorska Wróblewskiej (2014). Przedstawiono w niej wynik wytrenowania stochastycznych parserów zależnościowych na danych złożonych ze Składnicy zależnościowej rozszerzonej o ręcznie utworzone drzewa dla zdań zawierających różne klasy trudnych zjawisk składniowych (zob. też Wróblewska 2014; Wróblewska i Woliński 2012; Wróblewska 2012).

Ogromnymi zaletami analizy zależnościowej są prostota stosowanej reprezentacji oraz efektywność algorytmów analizy. Problemem w analizie zależnościowej są struktury współrzędne (skoordynowane). Na rysunku 5.1 przedstawiono dwie propozycje struktury zależności dla frazy współrzędnej ze spójnikiem i. Pierwsza byłaby wynikiem konwersji odpowiedniego drzewa



Rysunek 5.1. Alternatywne struktury zależnościowe dla zdania *Widzę Tomasza i Alinę.*

składnikowego Składnicy: uznano w nim, że centrum składniowym konstrukcji *Tomasza i Alinę* jest spójnik, więc to do niego prowadzi zależność od formy czasownikowej. Oba składniki koordynowane są połączone krawędzią zależnościową ze spójnikiem.

Niektórzy badacze (np. Mel'čuk 1988) uważają jednak, że krawędź obj powinna łączyć czasownik ze składnikiem, który może być reprezentantem konstrukcji, a więc z formą *Tomasza* albo *Alinę*. Ten wariant ilustruje druga struktura. Jej wadą jest to, że bezpośrednie powiązanie z czasownikiem ma jedynie forma *Tomasza*, podczas gdy drugi kandydat na reprezentanta – forma *Alinę* – jest ukryty głębiej w strukturze.

Żadna z tych struktur nie zdaje jednocześnie sprawy z dwóch faktów: że zależność konotacyjna zachodzi w istocie między formą *widzę* i każdą z form *Tomasza* i *Alinę* oraz że fragment *Tomasza i Alinę* stanowi pewną całostkę składniową. Wprowadzenie takiej całostki składniowej jest łatwe w drzewach składnikowych. Analogiczny problem dotyczy leksykalnych jednostek wielocłonowych jak *na pewno* albo *co do* – chciałoby się je w analizie zależnościowej traktować jako obiekty atomowe.

Dodatkowe komplikacje wiążą się z przypisywaniem cech gramatycznych. W drzewie zależnościowym wierzchołki odpowiadają formom fleksyjnym. Cechy wierzchołków są więc jednocześnie cechami formy stanowiącej centrum frazy i cechami frazy tworzonej przez wierzchołek ze swoimi potomkami. Gdy te dwie charakterystyki się różnią, pojawia się problem reprezentacji (por. Obrębski 2002, p. 2.1.2.3). Przykładem może być zdanie z nieredukowalną frazą nominalną (por. p. 2.9.3):

(4) *Tomasz i Alina przyszli.*

Niezależnie od tego, którą z proponowanych reprezentacji przyjmie się dla frazy *Tomasz i Alina*, pojawi się problem, polegający na niezgodności wartości liczby. Mnoga forma czasownika wymaga bowiem podmiotu w liczbie mnogiej. Żaden z dostępnych elementów drzewa zależnościowego nie ma jednak takiej wartości liczby.

5.5. Pożądane cechy formalizmu składniowego

Doświadczenie budowy gramatyki analizatora Świgra 2 daje pewne wskazówki co do pożądanych cech formalizmu składniowego, za pomocą którego wygodnie byłoby opisywać język polski[2].

Prace zreferowane w poprzednich punktach wskazują, że praktycznie implementowalne są gramatyki operujące składnikami (lub formami, jak w wypadku opisów zależnościowych), a nie zestawiające ograniczenia na bardziej abstrakcyjnych strukturach, jak w HPSG.

Wygodnym narzędziem opisu własności obiektów językoznawczych są zagnieżdżone struktury atrybutów. Są one wyraźnie bardziej czytelne od termów prologowych używanych w implementacji analizatora Świgra 2. Użyteczne jest też zbieranie pewnych zbiorów atrybutów w podstruktury, zwłaszcza w wypadku parametrów, którymi gramatyka operuje łącznie (w odniesieniu do GFJP por. Bień 2009, rozdz. 5). Bardzo elegancki sposób manipulowania takimi strukturami daje formalizm LFG.

Czytelności reprezentacji sprzyja jawna struktura drzewiasta. Jest ona lepsza od struktury ukrytej w zagnieżdżeniach struktur atrybutów. Jeśli chodzi o wybór między strukturą składnikową i zależnościową, przewagę wydaje się mieć struktura składnikowa ze względu na naturalną reprezentację większej klasy struktur (w szczególności współrzędnych). Drzewa zaproponowane w rozdziale 2 wydają się rozsądnym kompromisem między szczegółowością i czytelnością.

[2] Przemyślenia zawarte w tym punkcie należy traktować jako subiektywne odczucia autora.

Adekwatną reprezentacją wypowiedzenia może być drzewo składnikowe, którego węzły są powiązane z pewnymi strukturami atrybutów. W gramatyce Świgra 2 przyjęto, że reprezentację stanowi drzewo wraz z owymi strukturami, tak więc struktura powiązana z wierzchołkiem może reprezentować jego lokalne właściwości, a nie własności całego poddrzewa. Jest to mechanizm abstrakcji: w strukturze związanej z danym wierzchołkiem powinny występować wyłącznie cechy mające wpływ na interakcję tego wierzchołka z jego otoczeniem składniowym. Inaczej dzieje się w formalizmach HPSG i LFG – struktura atrybutów reprezentuje zawsze całe poddrzewo ze wszystkimi jego szczegółami. Pierwsze podejście zabezpiecza przed tworzeniem bardzo głęboko zagnieżdżonych struktur atrybutów, ale wymaga szczegółowych decyzji, które cechy muszą być widoczne w reprezentacji danej frazy. Drugie podejście zapewnia łatwe odwołanie się do dowolnej cechy dowolnie zagnieżdżonej frazy, ale kosztem jest komplikacja tworzonych struktur, która wydaje się przekładać na problemy z efektywnością analizy.

Tworzenie drzew odpowiadających intuicjom językowym wymaga pozbycia się ograniczenia do węzłów o ustalonej arności. Mechanizm zaproponowany w punkcie 4.2 spełnia ten postulat, ale jego użycie skutkuje dużą komplikacją sposobu formułowania warunków.

Formułowanie ograniczeń nakładanych na budowane struktury byłoby łatwiejsze przy obecności silnego systemu typów. Można by sobie wyobrazić mechanizm analogiczny do warunków iterowanych w prezentowanej tu gramatyce, wyrażony poprzez przypisanie ograniczeń do typów struktur (jak w HPSG). Ciekawa byłaby próba wyrażenia mechanizmu pochłaniania przecinków (por. p. 4.4.5) za pomocą tego rodzaju globalnie wyrażonych ograniczeń.

Wyraźna jest potrzeba tworzenia wielu warstw opisu: do warstwy składniowej powinna być dodana warstwa semantyczna i być może inne. Optymalny formalizm powinien umożliwiać opisywanie owych warstw w jak największym stopniu niezależnie.

Doświadczenie rozbudowy gramatyki Świgry 2 wskazuje też, że gramatyka zyskałaby na przejrzystości, gdyby udało się sformułować schematy dla konstrukcji współrzędnych jako rodzaj szablonów. Szablony takie musiałyby pozwalać na zaaplikowanie do różnych typów fraz z możliwością uzupełnienia pewnych szczegółów w zależności od typu frazy.

6

Korpus składniowy Składnica

Przy tworzeniu gramatyki dla języka naturalnego istotne jest zweryfikowanie, w jakim stopniu gramatyka ta adekwatnie opisuje zjawiska językowe. Można to zrobić, analizując z użyciem danej gramatyki wybrany zbiór tekstów i oceniając struktury wygenerowane przez gramatykę. W wyniku takiej procedury powstaje korpus składniowy (ang. *treebank*). Korpus składniowy może też być wdzięcznym obiektem badań językoznawczych, w szczególności badań językoznawstwa kwantytatywnego. Innym zastosowaniem korpusu składniowego jest wykorzystanie go jako danych treningowych w metodach uczenia maszynowego. Pozwala to w szczególności uzyskać wersję analizatora składniowego wskazującą jedno konkretne drzewo jako najbardziej prawdopodobne dla danego zdania (por. rozdz. 7).

Tematem bieżącego rozdziału jest Składnica – zbudowany za pomocą analizatora Świgra 2 pierwszy korpus struktur składniowych języka polskiego o zauważalnej wielkości (Woliński *et al.* 2011; Świdziński i Woliński 2010; Świdziński *et al.* 2013).

Na świecie korpusy składniowe mają trwałą pozycję w pracach badawczych związanych z automatyczną analizą składniową. „Banki drzew" stworzono dla wielu języków, w szczególności dla angielskiego (np. The Penn Treebank – zanalizowany tekst o długości ponad 1 000 000 słów, struktura frazowa; Marcus *et al.* 1993), niemieckiego (NEGRA – tekst gazetowy długości 20 602 zdań, 355 096 słów, struktura frazowa; Brants *et al.* 2003), a także czeskiego (Prague Dependency Treebank – 1 500 000 słów w wersji 2, struktura zależnościowa, korpus ten jest rozwijany od ok. 20 lat; Böhmová *et al.* 2003; Hajič *et al.* 2001).

Wymienione korpusy drzew składniowych były tworzone ręcznie lub półautomatycznie. Całkowicie automatyczna budowa banku drzew jest niemożliwa, o wartości korpusu stanowi bowiem w znacznym stopniu element weryfikacji znakowania składniowego przez ekspertów.

Wcześniejsza próba budowy polskiego korpusu składniowego (Marciniak *et al.* 2000) różniła się istotnie od przedstawianego tu projektu. Przyjęto mianowicie konstrukcję drzew w ramach teorii HPSG oraz założono całkowicie ręczną budowę drzew i w związku z tym bardzo ograniczoną wielkość korpusu. Wynikiem wspomnianego projektu jest zbiór 340 zdań testowych ilustrujących różne zjawiska składniowe polszczyzny.

Realnym prototypem korpusu składniowego dla polszczyzny była tzw. baza wypowiedników polskich (Świdziński 1996), obejmująca 4500 wypowiedzeń wylosowanych z korpusu. Była to jednak analiza tylko poziomów zdaniowych wypowiedzeń (nie powstawały więc pełne drzewa), wykonana przez ludzi, a nie automatycznie. Zasady opisu były zgodne z GFJP, jednak opisywano również struktury, których GFJP nie obejmuje, w szczególności wypowiedzenia nieciągłe i niezdaniowe. Nie analizowano też problemów, które nasunęłyby składniki niezdaniowe. Z założenia każdemu wypowiedzeniu przypisywano jedną składnikową interpretację składniową. Dla poszczególnych składników zdaniowych analizowano odpowiadające im schematy zdaniowe (Szpakowicz i Świdziński 1981), a także odnotowywano rozmaite charakterystyki ilościowe.

Istotną zachętą do budowy korpusu Składnica w oparciu o GFJP było to, że Maciej Ogrodniczuk (Ogrodniczuk 2006) pokazał, posługując się implementacją GFJP (Woliński 2004), że uzupełnienie reguł gramatyki pozwala zwiększyć udział zdań akceptowanych do 84% korpusu wypowiedników Świdzińskiego.

Korpus Składnica jest budowany poprzez automatyczny rozbiór wypowiedzeń za pomocą analizatora Świgra 2. Następnie wynikowe drzewa są weryfikowane przez językoznawców. Świgra 2 często generuje dla zdania liczne zbiory dopuszczalnych alternatywnych interpretacji składniowych zgodnych z gramatyką. Podstawowym zadaniem ekspertów jest odnalezienie w takim zbiorze właściwego drzewa dla danego zdania. Językoznawcy oceniają również poprawność wypowiedzeń. Gdy dla poprawnego wypowiedzenia nie ma właściwego drzewa, potrzebna jest zmiana w gramatyce formalnej. Proces jest więc iteracyjny: rozbudowywana jest jednocześnie i gramatyka, i korpus składniowy (por. Branco 2009).

Świgra 2 generuje drzewa składnikowe, więc zasadniczą postacią korpusu jest *Składnica frazowa*. Dzięki oznaczeniom centrów składniowych (por. p. 2.6) drzewa te można skonwertować do postaci zależnościowej, tworząc wtórną reprezentację korpusu – *Składnicę zależnościową*. System etykiet w tej formie Składnicy zaprojektowała Alina Wróblewska i ona jest też autorką parserów zależnościowych wytrenowanych na Składnicy (Wróblewska i Woliński 2012). Składnica zależnościowa była wykorzystana w zadaniu SPMRL2013 (Statistical Parsing of Morphologically Rich Languages, Seddah *et al.* 2013). Wtórnie drzewa zostały skonwertowane również do systemu Universal Dependencies (Nivre *et al.* 2016) i udostępnione na stronie projektu[1].

Bardzo ważną cechą korpusu Składnica jest to, że wybrane 20 000 zdań jest poddawane analizie z użyciem kolejno ulepszanych wersji gramatyki[2]. Uzyskany w ten sposób korpus daje realne dane o skuteczności zbudowanej gramatyki i parsera, umożliwiając tym samym zbliżenie się do odpowiedzi na pytanie, w jakim stopniu język polski poddaje się takiej formie opisu.

[1] http://universaldependencies.org/treebanks/pl_sz/index.html

[2] Jak się zdaje, Składnica jest do tej pory jedynym korpusem składniowym dla polskiego, w którym utrzymano tę zasadę.

W tej chwili publicznie dostępne korpusy składniowe dla polszczyzny obejmują oprócz Składnicy: bank drzew zależnościowych zawierający Składnicę zależnościową, ale rozbudowany o kolejne drzewa tworzone ręcznie (Wróblewska 2014) oraz bank struktur LFG opracowywany z użyciem gramatyki POLFIE (Patejuk i Przepiórkowski 2014).

6.1. Podstawa tekstowa

Źródłem tekstów dla Składnicy jest Narodowy Korpus Języka Polskiego (NKJP, Przepiórkowski *et al.* 2012). Ściślej rzecz biorąc, wykorzystany został podkorpus NKJP o objętości około 1 miliona słów (oznaczany NKJP1M), który ma dwie korzystne cechy: jest on – zdaniem autorów – zrównoważony, jeśli chodzi o reprezentację stylów funkcjonalnych, oraz został ręcznie oznakowany fleksyjnie. Na potrzeby Składnicy z NKJP1M wylosowano próbki łącznie zawierające 20 000 zdań. W losowaniu zostały pominięte, wyodrębnione na podstawie metadanych NKJP, teksty internetowe i mówione, chodziło bowiem o zbadanie adekwatności gramatyki w odniesieniu do redagowanych tekstów pisanych[3].

Uwzględnienie w próbkach kilkuzdaniowego kontekstu pomagało w ujednoznacznianiu analiz składniowych poszczególnych zdań. Dzięki niemu można również wykorzystać korpus do prac, w których przetwarzanie tekstu nie ogranicza się do osobnych zdań, na przykład do wykrywania nawiązań.

Znakowanie fleksyjne NKJP1M zostało przeprowadzone w ten sposób, że wyniki automatycznego analizatora fleksyjnego Morfeusz SGJP ręcznie ujednoznaczniono oraz dodano interpretacje fleksyjne segmentów nieznanych analizatorowi (zwłaszcza nazw własnych)[4]. W wyniku tych prac każdy segment korpusu ma dokładnie jedną, zweryfikowaną przez człowieka interpretację fleksyjną. Wykorzystanie w Składnicy tak przygotowanego tekstu pozwoliło ekspertom skupić się na kwestiach składniowych.

Początkowo zakładano, że w ramach prac nad znakowaniem składniowym będzie się przyjmować za dane i niezmienne znakowanie fleksyjne NKJP1M. W trakcie projektu okazało się jednak, że interpretacje w ręcznie znakowanym NKJP1M nie są wolne od błędów, a i pewne decyzje systematyczne trzeba zmienić w wyniku refleksji składniowej. Tak więc warstwa opisu fleksyjnego Składnicy zaczęła się różnić od opisu NKJP1M.

---

[3] Rozszerzenie opisu na teksty odbiegające od normy językowej (co często zdarza się tekstom internetowym) można widzieć jako osobne zadanie.

[4] Praca ta była wykonana ok. 10 lat temu. Od tego czasu zmienił się słownik fleksyjny, jak również wyewoluowały poglądy na zasady znakowania. W tej chwili toczą się prace mające na celu uzgodnienie NKJP1M z najnowszą wersją Morfeusza.

6.2. Zasady znakowania korpusu

Przyjęto technikę pracy nad korpusem składniowym polegającą na generowaniu lasów składniowych za pomocą analizatora Świgra 2, które były następnie przedstawiane ekspertom-językoznawcom do akceptacji. Każde drzewo w korpusie musi być wynikiem pracy automatycznego analizatora, co zapewnia pełną zgodność drzew z gramatyką formalną.

Oczywiście nie wszystkie zdania w korpusie są akceptowane przez analizator. Co więcej, nawet jeśli analizator generuje drzewa dla danego zdania, zdarza się, że wśród nich nie ma takiego, które eksperci by zaakceptowali. Takie sytuacje wymagają zmian w gramatyce. W ten sposób analiza korpusu wpływa na rozwój gramatyki, a zmiany w gramatyce odzwierciedlają się w kolejnych wersjach korpusu składniowego. Za takim iteracyjnym podejściem do konstrukcji korpusów składniowych opowiadali się np. Branco (2009) i Rosén *et al.* (2006).

Od ekspertów oczekuje się wyboru jednego konkretnego drzewa dla każdego zdania. Mimo że gramatyka opisuje składnię powierzchniową (a więc niesemantyczną), w wyborze drzew uwzględniana jest semantyka i wszelkie inne uwarunkowania tekstu, o ile da się je wywnioskować z kontekstu zdań otaczających.

W początkowej fazie pracy wszystkie zdania z jednej próbki korpusowej były przydzielane do obróbki tej samej osobie, aby ułatwić interpretowanie ich w kontekście. W późniejszych iteracjach do rozpatrzenia trafiają pojedyncze poprawione zdania, więc sposób przydziału nie ma znaczenia.

Praca nad wyborem i weryfikacją drzew składniowych odbywa się w systemie Dendrarium przedstawionym w punkcie 6.5. Do zadań jego użytkowników, zwanych w żargonie projektu dendrologami, należy wybór drzewa i zbadanie go, a w szczególności: 1) sprawdzenie, czy podział na frazy na wszystkich poziomach struktury jest poprawny, 2) zbadanie poprawności wymagań składniowych wszystkich predykatów, 3) ustalenie, czy właściwe elementy zostały oznaczone jako centra fraz, 4) weryfikacja poprawności wszystkich znaczników fleksyjnych w wybranym drzewie. W wypadku braku poprawnego drzewa dendrolog wybiera tzw. odpowiedź specjalną i opatruje ją komentarzem, pomagającym później w poprawieniu gramatyki.

W początkowej fazie tworzenia korpusu składniowego stosowano następujące odpowiedzi specjalne:

– „wypowiedzenie niepoprawne" – zawierające błąd gramatyczny lub poważny błąd interpunkcyjny (np. brak przecinka zamykającego zdanie podrzędne);
– „nie zdanie" – wypowiedzenie nie dające się rozłożyć na składniki zbudowane wokół form finitywnych czasowników;
– „zbyt trudne" – wypowiedzenie zawierające element z góry uznany za zbyt trudny do uwzględnienia (frazy nieciągłe, przytoczenia, pewne znaki interpunkcyjne: cudzysłowy, nawiasy, myślniki);

– „błąd w opisie fleksyjnym" – wypowiedzenie zawierające błąd w przejętym z NKJP1M opisie fleksyjnym;
– „brak drzewa" – wypowiedzenie zawierające jakiś problem wymagający naprawienia po stronie gramatyki.

Wszystkie wymienione odpowiedzi specjalne z wyjątkiem ostatniej zatrzymywały dalsze przetwarzanie wypowiedzenia. Ostatnia odpowiedź specjalna powodowała powrót wypowiedzenia do puli wypowiedzeń oczekujących na niezbędne poprawki w gramatyce.

W dalszych fazach projektu podjęto próbę opisania również konstrukcji niezdaniowych (zob. p. 2.12) i innych „zbyt trudnych" (w szczególności częstych konstrukcji nieciągłych, p. 2.14). Zaczęto także poprawiać opisy fleksyjne. Nie podjęto jedynie próby opisywania wypowiedzeń wyraźnie niepoprawnych (choć dopuszczono pewne częste wykolejenia, np. dotyczące interpunkcji). Tak więc w późniejszych fazach ograniczono odpowiedzi specjalne do trzech: „wypowiedzenie niepoprawne", „błąd w opisie fleksyjnym" i „brak drzewa".

6.3. Niejednoznaczności analizy składnikowej

Gramatyki języków programowania są z zasady jednoznaczne. Wymusza to albo dobór formalizmu, albo ustalenie z góry, która interpretacja ma być preferowana. W wypadku sformalizowanej gramatyki języka naturalnego typowa jest sytuacja przeciwna: niejednoznaczności wkradają się do opisu na wszystkich poziomach: dany segment może mieć wiele interpretacji fleksyjnych; ten sam wierzchołek w drzewie może być wywiedziony z użyciem różnych reguł gramatyki; słownik walencyjny zwykle zawiera wiele schematów dla danej jednostki. Rozstrzygnięcie niektórych z niejednoznaczności wymaga odwołania do semantyki, co jest niemożliwe na poziomie przetwarzania opisywanym w tej pracy.

Typem wierzchołka w drzewach składniowych, który jest najbogatszy w różne interpretacje, a przez to daje najwięcej niejednoznaczności, jest **zdanie**. W jego wypadku długość listy alternatywnych realizacji może przekroczyć 100.

Typowym źródłem wieloznaczności jest możliwość zinterpretowania ciągu segmentów jako jednej frazy lub rozbicia go na wiele fraz. Dzieje się tak na przykład w obecności fraz przyimkowych, które mogą stać się składnikiem zarówno fraz składnikowych (mogą być podrzędnikiem rzeczownika, przymiotnika, przysłówka), jak i bezpośrednio składnikiem zdania. Biorąc pod uwagę, że analizator nie ma podstaw do rozstrzygnięcia, która z tych możliwości jest właściwa w danym wypadku, musi generować wszystkie możliwe struktury. Na przykład następujące wypowiedzenia składają się z form o bardzo podobnych

charakterystykach, więc analizator dla obu generuje te same struktury składniowe:

(1) *Maria do jutra* *jest zajęta.*
(2) *Droga do domu* *jest zajęta.*

Jednak w pierwszym zdaniu fraza przyimkowa *do jutra* w oczywisty sposób jest podrzędnikiem czasownika, a w drugim fraza *do domu* raczej jest podrzędnikiem rzeczownika *droga*.

Wiele niejednoznaczności jest także wprowadzanych przez rozróżnienie między frazami wymaganymi **fw** i luźnymi **fl**, ponieważ jest wiele typów fraz składnikowych, które mogą realizować obie możliwości. Jak zresztą wspomniano, samo rozróżnienie jest nieostre i czasami decyzja, czy w schemacie danego czasownika dane wymaganie należy uwzględnianić, jest w istocie arbitralna.

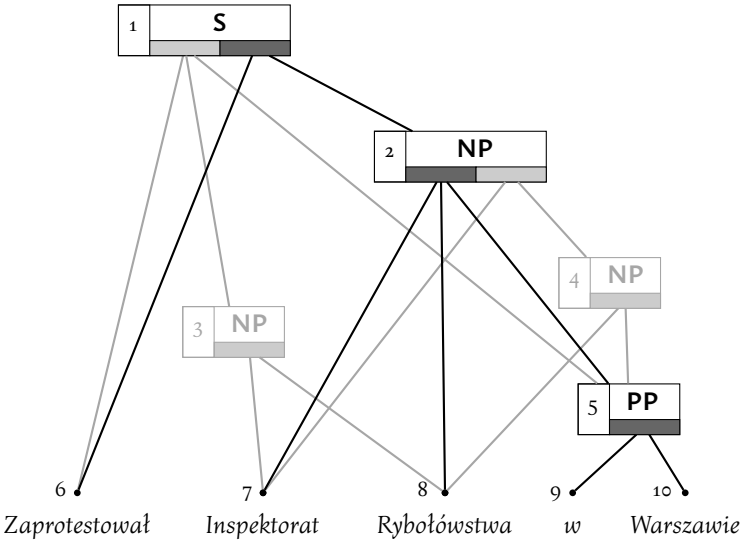
Najprostszym przykładem takiej niejednoznaczności są frazy nominalne biernikowe:

(3) *Czytał książkę.*
(4) *Czytał godzinę.*

Pierwsza fraza nominalna jest wymagana jako obiekt czytania, druga, wyrażająca czas trwania czynności, jest typową frazą luźną. Różnica między ich interpretacjami wynika z uwarunkowań semantyczno-leksykalnych: co może być miarą czasu. Klasa rzeczowników, które mogą wystąpić w tym miejscu jest ograniczona, ale trudno wymienić wszystkie. Oprócz bowiem oczywistych miar czasu jak *godzinę*, *minutę*, *miesiąc*, *chwilę* możliwe są określenia specyficzne dla pewnych kontekstów, np. *kolejkę* (w czasie gry w chińczyka).

Ciekawym źródłem niejednoznaczności składniowych są także ciągi form rzeczownikowych w dopełniaczu, które mogą układać się w różne struktury drzewiaste (Bartosiak i Woliński 2015). Najdłuższy taki ciąg w NKJP ma 11 elementów: *głosu wiceprezesa Rady Ministrów ministra skarbu państwa pana profesora Mirosława Pietrewicza*.

W gramatyce DCG niejednoznaczność oznacza, że pewien nieterminal o ustalonym zasięgu może być rozpisany na wiele zestawów składników bezpośrednich. Przyczyną takiej sytuacji jest istnienie wielu reguł o takiej samej lewej stronie (uwzględniając wartości argumentów jednostki nieterminalnej), a różnych prawych stronach. Jeżeli realizacja danego nieterminala jest niejednoznaczna, to w każdym drzewie, które go zawiera, można w miejsce jego realizacji podstawić dowolną z możliwych realizacji alternatywnych i otrzymać drzewo akceptowane przez gramatykę. Jeżeli dwa węzły niejednoznaczne nie mają wspólnego poddrzewa, to wprowadzane przez nie warianty wymnażają się – można je dowolnie zestawiać ze sobą. Powoduje to, że liczba niejednoznaczności szybko rośnie wraz z długością wypowiedzenia, a wzrost może mieć charakter wykładniczy (Woliński 2004). Dlatego drzewa składniowe generowane przez analizator Świgra 2 są przechowywane w postaci upakowanej (ang. *shared parse forests*, Billot i Lang 1989). Rysunek 6.1 przedstawia las



Rysunek 6.1. Przykład upakowanego lasu

dla zdania, w którym fraza przyimkowa *w Warszawie* może być składnikiem w trzech miejscach struktury:

(5) *Zaprotestował Inspektorat Rybołówstwa w Warszawie.*

Ciemniejsze linie i wierzchołki pokazują drzewo wybrane przez dendrologów, elementy szare składają się na pozostałą część lasu. Wierzchołek reprezentujący całe zdanie, oznaczony liczbą 1, jest niejednoznaczny: jego składnikami bezpośrednimi mogą być albo wierzchołki 6, 3 i 5 (w tej interpretacji Warszawa jest miejscem, gdzie zaprotestował Inspektorat Rybołówstwa), albo wierzchołki 6 i 2. Dendrolodzy wybrali tę drugą możliwość. Niejednoznaczny jest także wierzchołek 2: zależnie od jego rozkładu na składniki Inspektorat Rybołówstwa mieści się w Warszawie (7, 8 i 5) albo jest to inspektorat zajmujący się rybołówstwem w Warszawie (7 i 4). Wierzchołek 2 reprezentuje więc dwie możliwe struktury wewnętrzne fragmentu *Inspektorat Rybołówstwa w Warszawie*.

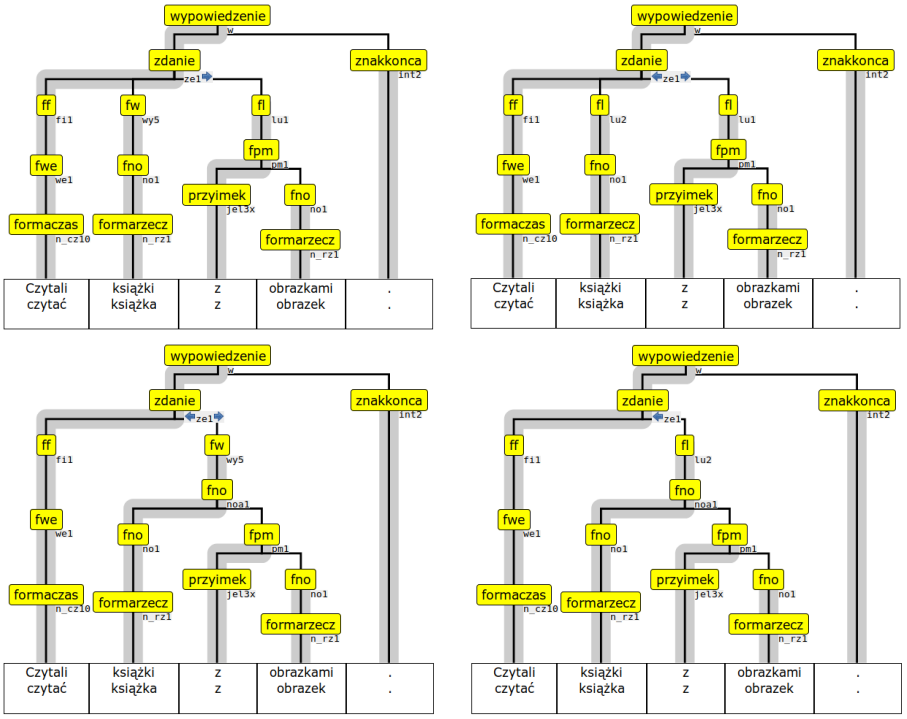
Upakowany las składniowy zawdzięcza zwartość reprezentacji temu, że każde poddrzewo występujące w wielu drzewach jest reprezentowane tylko jeden raz. W przykładzie każdy z wierzchołków 5, 7 i 8 jest dzielony przez 3 możliwych rodziców, a wierzchołek 6 ma dwóch możliwych rodziców. Liczba wierzchołków upakowanego lasu jest zawsze wielomianowa ze względu na długość wypowiedzenia (Billot i Lang 1989).

Ważną zaletą upakowanego lasu jest to, że pewne operacje można na nim wykonać bez generowania z niego wszystkich drzew. Przykładem takiej prostej operacji jest wyznaczenie liczby drzew zawartych w lesie. Również przedstawione w rozdziale 7 algorytmy ujednoznaczniające analizy składniowe operują bezpośrednio na upakowanym lesie.

Wizualizacja drzew składniowych ma duże znaczenie dla komfortu pracy ekspertów ujednoznaczniających wyniki analizy, a także dla użytkowników analizatora i korpusu składniowego.

W pierwszej implementacji GFJP (Woliński 2004) generowane były pliki PDF z „pionowym" zapisem drzew, przypominającym wyświetlanie katalogów w programach zarządzających plikami w komputerze. Istotnym krokiem było zaproponowanie interaktywnej wersji plików PDF, w których dowiązania hipertekstowe pozwoliły przechodzić między drzewami różniącymi się danym poddrzewem (Woliński 2006a). Wreszcie idea ta została zrealizowana w języku JavaScript w postaci biblioteki pozwalającej wyświetlić las składniowy w przeglądarce WWW i nawigować po poszczególnych niejednoznacznościach analizy (Zaborowski 2010).

Rysunek 6.2 przedstawia wizualizacje wszystkich drzew składniowych wygenerowanych dla zdania *Czytali książki z obrazkami*.. Drzewa te różnią się podczepieniem frazy *z obrazkami* do nadrzędnika *książki* albo do formy czasownikowej oraz potraktowaniem frazy z centrum *książki* jako wymaganej albo



Rysunek 6.2. Wizualizacje z elementami nawigacyjnymi wszystkich drzew składniowych dla zdania *Czytali książki z obrazkami*.

luźnej. Jedyny wierzchołek niejednoznaczny w tym lesie – wierzchołek dla jednostki **zdanie** – jest zaopatrzony w strzałki (przy symbolu reguły ze1, która posłużyła do wygenerowania tego wierzchołka). Kliknięcie strzałki poniżej danego wierzchołka niejednoznacznego w interfejsie analizatora powoduje przełączenie realizacji tego wierzchołka na kolejny wariant. W wypadku drzew na rysunku kliknięcie strzałki w prawo powoduje przejście do kolejnego przedstawionego drzewa.

Istotną cechą tej wizualizacji jest to, że kliknięcie strzałki nawigacyjnej powoduje zmianę jedynie w obrębie poddrzewa zaczepionego w skojarzonym wierzchołku niejednoznacznym. Pozostała część drzewa (włącznie z wierzchołkami niejednoznacznymi nienależącymi do tego poddrzewa) pozostaje bez zmian.

Taka forma nawigacji po strukturze dość dobrze dzieli przestrzeń poszukiwań. Wyjątkiem jest sytuacja, w której lista możliwych interpretacji dla danego wierzchołka jest bardzo długa (zdarza się to zwłaszcza dla wierzchołków zdaniowych). Usprawnienie nawigacji w tym wypadku mogłoby być przedmiotem przyszłych prac, do pomyślenia jest też zmiana sposobu wyszukiwania na podawanie cech poszukiwanego drzewa (zob. dalej uwagi o systemie INESS).

Przedstawiona forma wizualizacji jest stosowana do prezentacji drzew składniowych w systemie Dendrarium, sieciowej wersji analizatora Świgra 2 i w wyszukiwarce drzew składniowych (zob. p. 6.7).

6.5. Dendrarium

Technika tworzenia korpusu Składnica polega na wybieraniu drzew składnikowych z lasów składniowych generowanych przez analizator regułowy. W momencie rozpoczynania prac nad Składnicą (w roku 2008) nie istniał system, który pozwalałby na taki tryb pracy na składnikowych lasach składniowych. Podjęto więc decyzję o budowie systemu Dendrarium. Jedyny dostępny wówczas korpus składniowy dla typologicznie podobnego języka – czeskiego – Prague Dependency Treebank (Böhmová *et al.* 2003) – był tworzony w formalizmie zależnościowym, nie mógł więc być bezpośrednim źródłem rozwiązań. Obecnie można by chyba przystosować do lasów Składnicy system INESS (Rosén *et al.* 2007); kilka słów o różnicach między nim a Dendrarium znajdzie się więc na końcu niniejszego punktu. Porównanie z innymi systemami dostępnymi w momencie tworzenia Dendrarium można znaleźć w artykule (Woliński 2010).

System Dendrarium (Woliński 2010) wypełnia podstawowe założenie projektowe: pozwala grupie ekspertów na symultaniczną pracę na drzewach składniowych. Zostało to osiągnięte poprzez realizację Dendrarium w postaci apli-

kacji webowej operującej bazą danych przechowywaną na serwerze[5]. Zadaniem użytkownika jest wybór i weryfikacja struktury składniowej dla zdania. Biorąc pod uwagę, że liczba drzew generowanych dla niektórych zdań jest bardzo duża, wybór ten nie może odbywać się poprzez przegląd kolejnych proponowanych drzew.

W Dendrarium ujednoznacznienie drzewa polega na wybraniu wariantów interpretacyjnych dla poszczególnych wierzchołków niejednoznacznych w upakowanym lesie składniowym. Dendrolog operuje na zwartej reprezentacji całego lasu składniowego poprzez wybieranie odpowiednich fragmentów drzew. System jest odpowiedzialny za to, aby prezentowane fragmenty dawały się skomponować w pełne drzewo.

Każde zdanie przekazywane jest dwóm dendrologom, którzy opracowują je niezależnie. Jeżeli przedstawione opisy są zgodne, wynik jest akceptowany. W razie wykrycia rozbieżności zdanie jest ponownie pokazywane obu dendrologom z prośbą o weryfikację i ponowne zatwierdzenie. Faza ta ma pozwolić dendrologom na wyłapanie prostych omyłek (w tej fazie nie widzą oni konkurencyjnej odpowiedzi, ażeby nie sugerować się wzajemnie). Jeżeli rozbieżność nadal istnieje, zdanie jest kierowane do „superdendrologa", który porównuje odpowiedzi i wybiera lepszą. Superdendrolog może także interweniować w dowolnych miejscach struktury zdania. Taki tryb postępowania wydaje się być przyjęty w znakowaniu fleksyjnym korpusów, w szczególności tak znakowano NKJP1M (Przepiórkowski *et al.* 2012), a także polskie korpusy historyczne (Kieraś *et al.* 2017; Kieraś i Woliński 2018).

Analizując dane wypowiedzenie, dendrolodzy oceniają, czy kształt drzewa jest właściwy, a więc czy nastąpiły właściwe hierarchiczne podziały na składniki; czy struktura argumentów jest zgodna ze schematami w słowniku Walenty; czy właściwy jest opis fleksyjny w liściach drzewa. Jeżeli w lesie składniowym brak drzewa spełniającego te warunki, dendrolodzy przypisują wypowiedzeniu jedną z odpowiedzi specjalnych wymienionych w punkcie 6.2. W wypadku odpowiedzi specjalnych różnych od „wypowiedzenie niepoprawne" po wprowadzeniu odpowiednich poprawek w gramatyce lub opisie fleksyjnym wypowiedzenie jest poddawane ponownej analizie automatycznej i ponownej ocenie przez ekspertów.

W związku z taką organizacją pracy bardzo istotnym elementem systemu Dendrarium jest mechanizm pozwalający uniknąć ponownego ujednoznaczniania drzew generowanych przez kolejne wersje gramatyki. Mechanizm ten wyszukuje w nowym lesie wygenerowanym przez analizator drzewo poprzednio oznaczone jako poprawne i zastępuje stary las nowym lasem z oznaczonym tym samym drzewem. Jeżeli operacja ta się powiedzie, uznaje się, że odpo-

[5] Kod programu Dendrarium został udostępniony na licencji GPL pod adresem git://git.nlp.ipipan.waw.pl/constituency/Dendrarium/. Zrąb systemu zaimplementowali studenci: Karolina Sołtys, Piotr Achinger, Tomasz Badowski i Dominika Pawlik. Wizualizację struktur składniowych zrealizował Andrzej Zaborowski. Późniejsze zmiany wprowadzali Jan Szejko i Tomasz Bartosiak.

wiedź dla danego zdania pozostaje w mocy: jeżeli jakieś drzewo zostało uznane za poprawne, będzie oznaczone jako wybrane drzewo w nowym lesie, nawet jeżeli zawiera on również możliwości interpretacyjne, których wcześniej nie przedstawiano dendrologom. Dlatego też dendrolodzy powinni akceptować jedynie drzewa, które są w całości poprawne (a nie najlepsze z dostępnych). Może się zdarzyć również, że nowy las nie zawiera poprzednio wybranego drzewa. Taka sytuacja zwykle jest sygnałem, że zmiany wprowadzone do kolejnej wersji gramatyki zawierają jakiś błąd, eliminujący wcześniej akceptowane poprawne interptetacje. Dlatego sprawdzenie, czy występują takie wypadki, jest wykonywane próbnie bez faktycznej aktualizacji drzew – ewentualne wykryte problemy są analizowane przez autora gramatyki.

Ponieważ podczas nieomal dziesięciu lat rozwoju korpusu koncepcja struktury drzew ewoluowała (w szczególności zmieniono system oznaczeń fraz wymaganych na zgodny ze słownikiem Walenty), nierealistyczne okazało się założenie, że wystarczy wybierać drzewo literalnie identyczne z poprzednio wybranym. Dlatego też powstało kilka wersji procedury aktualizacyjnej uwzględniających systematyczne zmiany w drzewach wprowadzone w pewnych wersjach gramatyki (np. zmianę oznaczeń przypadków na łacińskie). Opracowano algorytm, który wykonuje systematyczne przekształcenie starego drzewa według założonych reguł i dopiero tej „unowocześnionej" postaci poszukuje w nowym lesie składniowym. Dzięki temu udało się doprowadzić Składnicę do zgodności z bieżącą wersją gramatyki.

Przykład ujednoznacznienia zdania w systemie Dendrarium

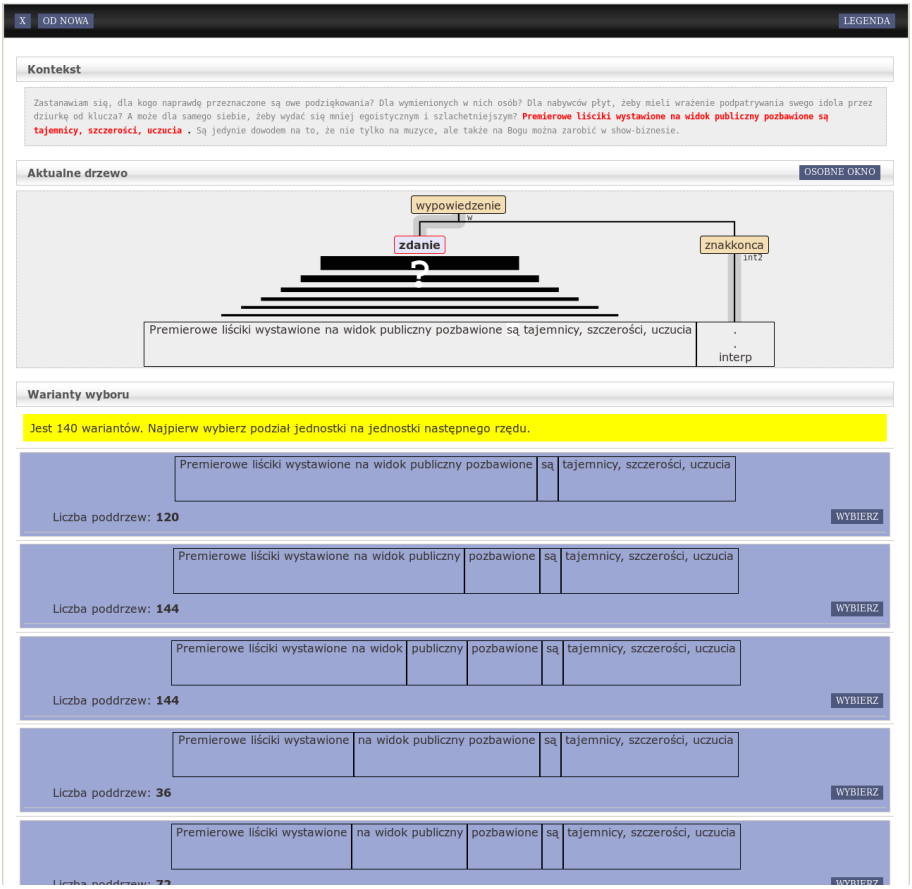
Ujednoznaczenienie struktury składniowej odbywa się poprzez wybieranie interpretacji dla węzłów niejednoznacznych, czyli takich, dla których istnieje więcej niż jeden sposób rozpisania ich na składniki bezpośrednie. Proces postępuje od węzłów niejednoznacznych najbliższych wierzchołka reprezentującego całe wypowiedzenie, ponieważ wybór interpretacji dla nich wpływa na to, jakie wierzchołki niejednoznaczne są dostępne w ich poddrzewach. Taka kolejność ujednoznaczniania gwarantuje, że rozpatrywane będą jedynie te wierzchołki, które wystąpią w docelowo wybranym drzewie (o ile ono istnieje).

W ten sposób drzewo jest budowane *top-down*, przy czym fragmenty jednoznaczne są od razu wypełniane przez system. Dendrolog jedynie dokonuje wyboru spośród możliwości dopuszczonych przez analizator automatyczny.

Proces ujednoznaczniania struktury składniowej w systemie Dendrarium zostanie przedstawiony na przykładzie zdania:

(6) *Premierowe liściki wystawione na widok publiczny pozbawione są tajemnicy, szczerości, uczucia.* [Skł.]

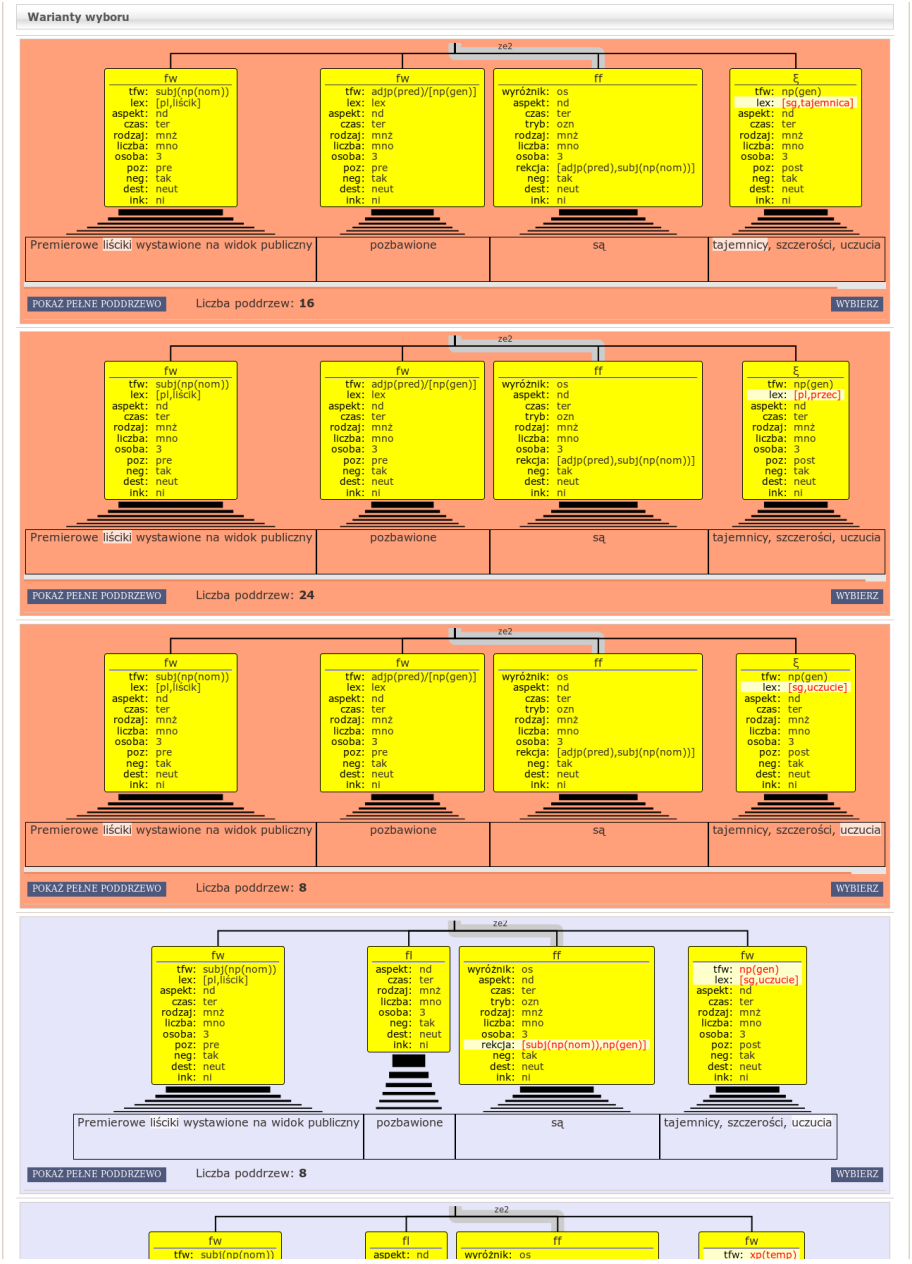
Rysunek 6.3 przedstawia pierwszy ekran widziany w przeglądarce WWW przez dendrologa, któremu przydzielono do opracowania to zdanie. Na górze



Rysunek 6.3. Pierwszy wierzchołek niejednoznaczny w zdaniu (6) – początek listy podziałów zdania na składniki

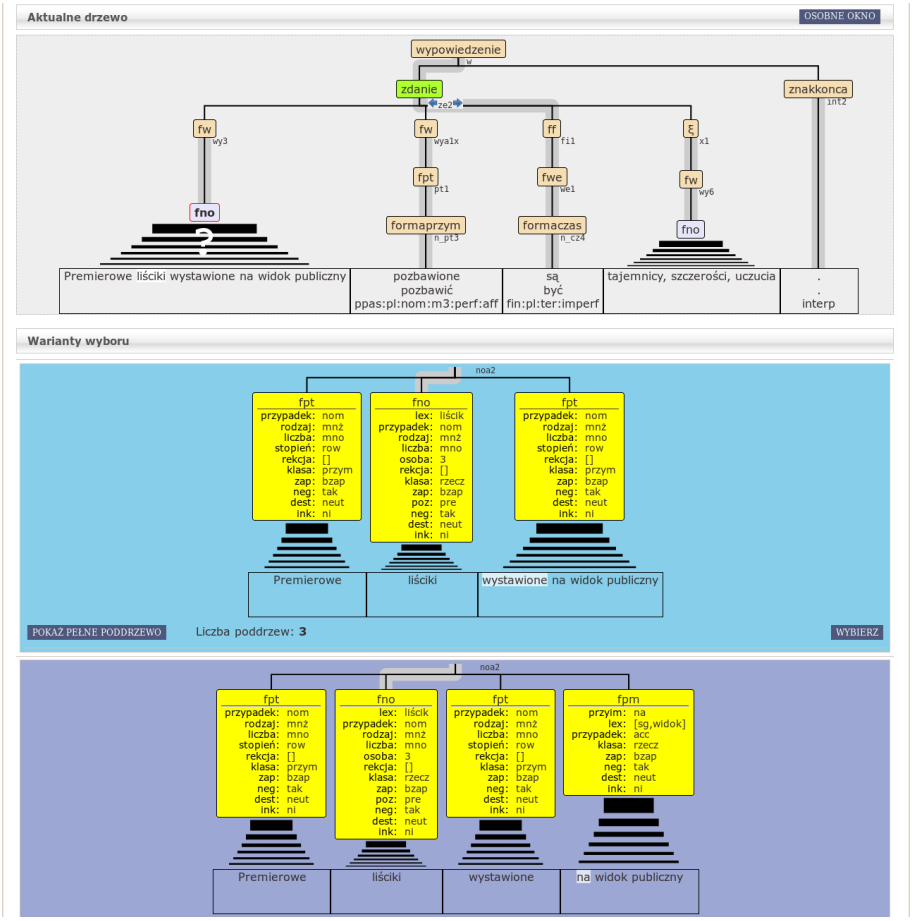
strony pokazywana jest w postaci tekstowej cała próbka korpusowa zawierająca opracowywane wypowiedzenie. Pozwala to odwołać się do kontekstu sąsiednich wypowiedzeń przy wyborze struktury składniowej.

Poniżej wizualizowany jest ujednoznaczniony dotychczas fragment drzewa składniowego. Dla rozważanego wypowiedzenia w pierwszym kroku jednoznaczny fragment składa się z korzenia **wypowiedzenie**, którego składnikami bezpośrednimi są jednostki **zdanie** i **znakkonca**. Węzeł odpowiadający jednostce **zdanie** jest najbliższym korzenia węzłem niejednoznacznym. Dlatego jest on podświetlony przez system, a poniżej przedstawione są możliwe sposoby rozpisania tego wierzchołka na składniki. Wybór spośród bardzo wielu możliwości jest niewygodny. Dlatego, jeżeli sposobów realizacji wierzchołka jest więcej niż dziesięć, system wyświetla najpierw tylko listę możliwych podziałów danego fragmentu na składniki bezpośrednie – bez typów fraz i atrybutów. Tak dzieje się w wypadku omawianego zdania (por. rys. 6.3 i 6.4).



Rysunek 6.4. Pierwszy wierzchołek niejednoznaczny w zdaniu (6) – warianty realizacji zdania wyświetlane po wyborze podziału na składniki

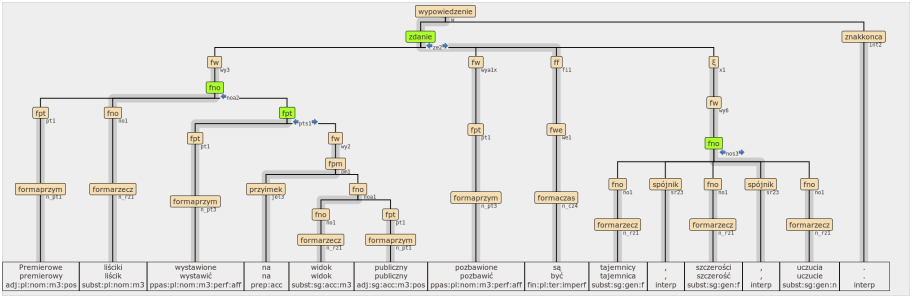
Analizowane zdanie jest konstrukcją w stronie biernej. Występuje w nim nieciągłość: fragment *tajemnicy, szczerości, uczucia* został oderwany od swojego nadrzędnika *pozbawione*. Zgodnie z zasadami opisu takich konstrukcji przed-



Rysunek 6.5. Drugi wierzchołek niejednoznaczny w zdaniu (6)

stawionymi w punkcie 2.14, należy więc wybrać drugi z podziałów proponowanych na rysunku 6.3. Wybór ten prowadzi do listy wariantów realizacji tego wierzchołka, której początek przedstawia rysunek 6.4. Pierwsze trzy warianty składają się z frazy wymaganej podmiotowej *Premierowe liściki wystawione na widok publiczny*, frazy wymaganej przymiotnikowej adjp(pred) *pozbawione*, w której brakuje składnika nominalnego w dopełniaczu (por. p. 4.5.5), frazy finitywnej *są* i odłączonej frazy wymaganej nominalnej w dopełniaczu *tajemnicy, szczerości, uczucia*. W pozostałych czterech realizacjach składnik *pozbawione* jest interpretowany jako fraza luźna, należy więc je odrzucić. Dla ułatwienia pracy dendrologa, wspólny kolor tła pierwszych trzech wariantów sygnalizuje, że składają się one z tych samych jednostek nieterminalnych, różniących się jedynie atrybutami. Dalej kolor się zmienia, bo i składniki są inne (fraza luźna).

Trzy realizacje z frazą wymaganą różnią się opisem składnika oderwanego. Różniące się atrybuty są w interfejsie zaznaczone kolorem czerwonym.



Rysunek 6.6. Wybrane drzewo dla zdania (6)

W tym wypadku jest to atrybut lex, niosący charakterystykę centrum leksykalnego frazy. Centrum jest także zaznaczane podświetleniem w liściach. Na tej podstawie można ustalić, że poszukiwaną interpretację skrywa wariant drugi, w którym centrum konstrukcji jest przecinek pełniący funkcję spójnika współrzędnego. Właściwa interpretacja tego zdania zawiera frazę skoordynowaną szeregową *tajemnicy, szczerości, uczucia*.

Wybór drugiego wariantu prowadzi do sytuacji przedstawionej na rysunku 6.5. Ustalony został skład zdania, niejednoznaczne pozostają jeszcze dwie frazy składnikowe. W tym wypadku dwa wierzchołki niejednoznaczne nie wpływają na siebie, dendrolog może więc wybrać, który ujednoznaczniać jako pierwszy.

Pierwsza niejednoznaczność dotyczy struktury podmiotu. Pokazane na ilustracji dwa warianty struktury odpowiadają uznaniu frazy przyimkowej *na widok publiczny* albo za podrzędnik imiesłowu *wystawione*, albo też rzeczownika *liściki*. Wybór wariantu, w którym fragment *wystawione na widok publiczny* w całości jest składnikiem, ujawnia jeszcze niejednoznaczność wewnątrz niego (nie pokazaną na ilustracjach) – fraza przyimkowa może stanowić frazę wymaganą lub luźną. Ponieważ słownik Walenty zawiera schemat dla frazeologizmu *wystawić na widok publiczny*, właściwym wyborem jest fraza wymagana typu prepnp(na, acc) z odpowiednim opisem leksykalizacji.

W drugim z niejednoznacznych wierzchołków widocznym na ilustracji 6.5 możliwości interpretacyjne dotyczą struktury frazy skoordynowanej *tajemnicy, szczerości, uczucia*. Właściwym wyborem jest struktura szeregowa w układzie II (por. p. 2.9.2).

Rozstrzygnięcie obu niejednoznaczności prowadzi do pełnego drzewa składniowego (pokazanego na ilustracji 6.6). Po osiągnięciu tego etapu operator systemu powinien upewnić się, że kształt drzewa i wartości atrybutów są właściwe.

Uzyskanie jednoznacznego drzewa dla omawianego przykładu wymagało od operatora podjęcia decyzji co do czterech wierzchołków niejednoznacznych (wierzchołki te są oznaczone wyróżniającym kolorem jak na ilustracji 6.6, co ułatwia ewentualny powrót do podjętych decyzji). Las składniowy mógł

zawierać więcej wierzchołków niejednoznacznych, które kryły się w odrzuconych fragmentach struktury. Kolejność rozstrzygania niejednoznaczności narzucona przez system ma na celu to, by ocenie były poddawane tylko wierzchołki występujące w docelowym drzewie.

Decyzje w kwestii podziału na składniki bazują na pojęciach dobrze zrozumiałych i intuicyjnych dla dendrologów. Jednak konieczne jest rozumienie pewnych konwencji opisu, na przykład w omawianym zdaniu trzeba rozumieć konwencję opisu konstrukcji nieciągłych.

Trudniejsze bywają wybory, w których różnice struktur objawiają się w różnicach wartości atrybutów. W kroku przedstawionym na rysunku 6.4 różnice strukturyzacji ostatniego składnika przejawiają się tylko w różnych opisach centrum frazy. Wybór jest zrozumiały pod warunkiem, że operator ma świadomość, że centrum składniowym frazy skoordynowanej jest element spójnikowy (w tej frazie – przecinek).

W razie trudności można za pomocą przycisków „Pokaż pełne poddrzewo" widocznych na ilustracjach wyświetlić dokładniejszy widok lokalnego fragmentu struktury. Pozwala to rozwiać wątpliwości, gdy odpowiedniość między wartościami atrybutów a wynikową strukturą jest zbyt zawiła.

Proces ujednoznaczniania jest szybki i skuteczny dzięki ograniczonej liczbie wierzchołków, w których trzeba podjąć decyzję.

Dendrarium a INESS

System Dendrarium i system INESS (Rosén *et al.* 2007) reprezentują dwa różne podejścia do problemu wyodrębniania konkretnego drzewa z lasu składniowego. W Dendrarium poszczególne decyzje polegają na wyborze spośród możliwych realizacji danego wierzchołka niejednoznacznego. W systemie INESS krok ujednoznacznienia polega na wyborze cechy, która przysługuje poprawnej strukturze, lub też jej zaprzeczenia. Przykładami takich cech mogą być: „trzeci segment powinien być interpretowany jako forma rzeczownika", „forma czasownikowa *przeczytał* ma w tym zdaniu dwa argumenty", „forma *biały* jest podrzędnikiem formy *domek*".

Obie metody przy ujednoznacznianiu skupiają się na wierzchołkach niejednoznacznych, więc obie wymagają podobnego nakładu pracy w celu ujednoznacznienia struktury. W Dendarium liczba decyzji jest równa liczbie wierzchołków niejednoznacznych w wybranym drzewie (która może być znacząco mniejsza od liczby wierzchołków niejednoznacznych w całym lesie). W INESS niekoniecznie każdy wybór cechy prowadzi do ujednoznacznienia choćby jednego wierzchołka, więc liczba wyborów może być większa. Być może jednak koncentracja na cechach jest bardziej intuicyjnym sposobem pracy. W Dendrarium zdarza się bowiem, że różnica w prezentowanych przez system strukturach jest tylko echem zjawiska składniowego, które wystąpiło w którymś poddrzewie. Na przykład jeśli tylko w jednej z dwóch interpretacji **zdanie** ma ustalony atrybut rodzaj, znaczy to, że w tej interpretacji zdanie to ma

podmiot o ustalonym rodzaju w odróżnieniu od drugiej. Wartość nieustalona może oznaczać brak podmiotu lub podmiot o nieustalonej wartości rodzaju, np. w postaci zaimka MY. Wybór cechy „to zdanie ma podmiot" byłby prostszy od wykonania tego rodzaju analizy.

Metoda oparta na cechach nie narzuca kolejności dokonywania wyborów, co można postrzegać jako zaletę lub wadę, ponieważ może się przy niej zdarzyć, że późniejszy wybór wyeliminuje całe poddrzewo, którego dotyczyły cechy wybrane wcześniej.

Inaczej też w obu systemach odbywa się zapamiętywanie wybranego drzewa i aktualizacja lasów do zmienionej wersji gramatyki. W Dendrarium zapamiętywane jest konkretne drzewo, a więc pełny opis składających się na nie wierzchołków i ich zasięgi. W systemie INESS zapamiętywane są wartości cech wybrane przy ujednoznacznianiu lasu.

Można w tym widzieć zasadniczą wadę systemu opartego na wyborze cech. Rozwój gramatyki odbywa się bowiem przede wszystkim poprzez opisywanie konstrukcji wcześniej nieuwzględnionych. To zaś oznacza, że w kolejnych wersjach przybywa interpretacji dla już zanalizowanych wypowiedzeń (zwiększa się liczba niejednoznaczności). W takiej sytuacji mechanizm aktualizujący w Dendrarium będzie w stanie znaleźć wybrane drzewo w nowym (obszerniejszym) lesie i zatwierdzić aktualizację automatyczną. To, że pewne wierzchołki były jednoznaczne w starym drzewie, a są niejednoznaczne w nowym, nie wpływa na ten proces.

Jednak w INESS zbiór cech, który w poprzedniej wersji gramatyki determinował wybór konkretnego drzewa, wobec zwiększonej liczby możliwości interpretacyjnych może przestać wystarczać. Może się okazać, że wiele nowych drzew pasuje do poprzednio wybranych cech. Nie sposób wtedy przywrócić jednoznaczności drzew bez ponownego przejrzenia zdań dotkniętych zmianą, których potencjalnie może być bardzo dużo.

6.6. Ewaluacja korpusu Składnica

Prace nad korpusem Składnica toczyły się w dwóch etapach wyraźnie różniących się zasadami opracowywania drzew. Etap pierwszy obejmuje prace wykonane w kierowanym przez autora projekcie *Budowa banku drzew składniowych dla języka polskiego z wykorzystaniem automatycznej analizy składniowej* (projekt N N104 224735 finansowany przez MNiSW w latach 2009–2011). Korpus powstały w tej fazie ochrzczono mianem Składnica 0.5.

Etap drugi obejmuje rozbudowę korpusu w ramach projektu POIG NEKST (*Adaptacyjny system wspomagający rozwiązywanie problemów w oparciu o analizę treści dostępnych źródeł*) w latach 2009–2014 i jego dalszy rozwój w ramach projektu infrastrukturalnego CLARIN-PL w latach 2015–2018.

Punktem wyjścia prac nad korpusem Składnica była implementacja GFJP (Woliński 2004), która akceptowała około 30% wypowiedzeń w korpusie Składnicy (Świdziński i Woliński 2009). Gramatyka dla nowej wersji analizatora, Świgry 2, została jednak w zasadzie napisana od nowa na wzór poprzedniej z wykorzystaniem nowych elementów formalizmu (punkt 4.2). Zmieniono kształt drzew zgodnie z opisem w rozdziale 2. Wprowadzono koordynację składników niezdaniowych. Dopracowano słownik walencyjny tak, aby opisywał ok. ¾ wystąpień czasowników w korpusie. Wprowadzono wiele drobnych ulepszeń w regułach gramatyki, w tym specyficznych realizacji leksykalnych form składniowych. Przyjęto jednak, że w tym projekcie opisane zostaną jedynie konstrukcje zdaniowe (z centrum finitywnym lub z centrum spójnikowym i wszystkimi składnikami będącymi zdaniami w tym sensie). Założono, że pozostałe wypowiedzenia zostaną odrzucone, ale sklasyfikowane.

W wyniku tych prac 8227 zdań (spośród 20 000, a więc 41,1%) otrzymało pełny opis w postaci drzewa, który został zaakceptowany przez dendrologów. Wynik klasyfikacji wypowiedzeń w korpusie przedstawia tabela 6.1. Wśród zdań odrzuconych 6% stanowiły zdania uznane przez dendrologów za niegramatyczne, 5% zaś zawierało błąd w opisie fleksyjnym. W większości były to błędy interpretacyjne, np. przypisanie biernika zamiast mianownika. Segmentom z literówką przypisywano w NKJP opis fleksyjny odpowiadający zapisowi poprawnemu, nie stanowiły więc one problemu przy analizie. Gros odrzuceń stanowiły jednak wypowiedzenia uznane za niemieszczące się w zaplanowanym dla tego etapu opisie (wypowiedzenia niezdaniowe i „zbyt trudne"). Stanowiły one 89% odrzuconych, czyli 29% całości korpusu.

Tabela 6.1. Klasyfikacja wypowiedzeń na pierwszym etapie prac nad Składnicą

	całości	odrzuconych	„zbyt trudnych"
podlegające opisowi	67%		
odrzucone	33%		
niepoprawne	2%	6%	
problem w opisie fleksyjnym	2%	5%	
nie zdanie	11%	34%	
„zbyt trudne"	18%	55%	
mowa niezależna	7%	21%	39%
cudzysłowy	5%	15%	27%
myślnik	3%	10%	19%
nawiasy	3%	8%	14%
dwukropek	1%	2%	3%
nieciągłość	1%	3%	5%
inne	0%	1%	2%

Dendrolodzy podawali powody uznania wypowiedzenia za „zbyt trudne" w postaci swobodnego opisu. Dlatego przedstawione w tabeli rozbicie tej klasy zostało wykonane na losowo wybranej próbce 100 zdań z tej grupy. Klasy nie są wykluczające się, więc liczby sumują się do więcej niż 100%. Widać, że największa liczba zdań została odrzucona z powodu obecności mowy niezależnej. Wymienione w tabeli znaki interpunkcyjne uznano za wygodne indykatory zjawisk składniowych. Część wystąpień cudzysłowów i myślników towarzyszy przytoczeniom. Myślniki bywają jednak także używane w funkcji zbliżonej do przecinka, np. objęte nimi bywają fragmenty wtrącone. Dwukropki mogą pełnić funkcję składniową zbliżoną do spójnika CZYLI. Przy przyjętych zasadach opisu nieciągłość była powodem odrzucenia tylko 1% wypowiedzeń (5 zdań „zbyt trudnych"). Dane te ukierunkowały dalszy rozwój gramatyki.

Stosunek liczby opracowanych zdań (8227) do liczby wypowiedzeń podlegających opisowi w tym etapie projektu (13412) pokazuje, że analizator automatyczny wygenerował poprawne drzewo dla 61% opisywanych wypowiedzeń.

Jak wspomniano wcześniej, lasy składniowe dla każdego wypowiedzenia są weryfikowane niezależnie przez dwóch dendrologów, a konflikty rozstrzygane przez superdendrologa. Dendrolodzy uzyskali zgodność opinii dla 88% wypowiedzeń. W wypadku wypowiedzeń z kolizją superdendrolog w 71% wypadków uznał za poprawną odpowiedź jednego z dendrologów, w pozostałych wypadkach udzielił odpowiedzi własnej (często będącej kombinacją fragmentów odpowiedzi dendrologów).

W tabeli 6.2 zestawiono statystyki kolizji według ostatecznie zatwierdzonego typu odpowiedzi. Jak widać w trzeciej kolumnie, w wypadku zdań z poprawnym drzewem częściej zdarzało się, że przynajmniej jeden z dendrologów dał odpowiedź zatwierdzoną potem przez superdendrologa.

Tabela 6.2. Kolizje w Składnicy według typów odpowiedzi

typ odpowiedzi	kolizja	jedna z odpowiedzi zatwierdzona
pełna	40%	79%
brak drzewa	37%	65%
odrzucone	23%	66%

Podjęto także próbę oszacowania, ile błędów mogło pozostać w danych po konfrontacji odpowiedzi. W tym celu wylosowano 100 wypowiedzeń z pełnymi drzewami. W próbie tej znaleziono aż 18 wypowiedzeń z błędami w drzewach, w niektórych wypadkach było to kilka błędów. W 6 wypadkach dendrolodzy wybrali złe miejsce zaczepienia podrzędnika, przy czym nie były to miejsca wątpliwe, ale ewidentnie złe. Dla 8 zdań źródłem problemu było uznanie frazy wymaganej za luźną lub na odwrót (z tego 4 wypadki dla centrów niefinitywnych). Raz błędnie wybrano typ frazy wymaganej. Dwukrotnie dendrolodzy

mieli problem z dopełniaczem negacji dla wyrażenia *nie ma [czegoś]*. W 4 wypadkach wybrano wyraźnie błędną strukturę drzewa. W jednym drzewie błędnie oznaczono podmiot predykatywu TO.

Wyniki te oznaczają niestety, że weryfikacja drzew składniowych jest dla wykonawców trudna i nawet zgodność poglądów dwóch osób lub sprawdzenie wyniku przez trzecią nie zapewnia pełnej bezbłędności. Wymienione typy błędów są takie, jakich można by z góry oczekiwać: podrzędniki często dają się podczepiać w różnych miejscach struktury (w szczególności frazy przyimkowe); rozróżnienie fraz wymaganych i luźnych często jest niejasne lub arbitralne. Interesującym sposobem poprawienia części błędów wydaje się statystyczna analiza prawidłowości w korpusie (zob. Krasnowska *et al.* 2012).

Etap drugi

Analiza wykonana w etapie pierwszym wskazała główne elementy, o które trzeba było rozbudować gramatykę w etapie drugim budowy korpusu. Przede wszystkim opracowano więc opis struktur zdaniopodobnych bez centrum finitywnego (por. p. 2.12).

Ponadto wprowadzono do gramatyki reguły opisujące mowę niezależną. Zjawisko to zostało uwzględnione w słowniku Walenty w postaci typu wymagania or (*oratio recta*) przysługującego niektórym czasownikom (por. p. 3.1.2). W regułach uwzględniono obecność przytaczanego zdania na początku lub na końcu wypowiedzenia z różnymi kombinacjami wydzielających je znaków interpunkcyjnych:

(7) *– Sekretarz pokazał swój lwi pazur – komentował potem Zdzich.* [Skł.]
(8) *Mówił: „Za rok ja pozbieram wszystkie dyplomy".* [Skł.]

Układy te obejmują przytłaczającą większość występujących w tekście przytoczeń. Ciekawym elementem rozbudowy opisu było uwzględnienie struktur z nieciągłością polegającą na oderwaniu frazy wymaganej (por. p. 2.14 i 4.5.5).

Niezmiernie istotną zmianą, zarówno dla gramatyki, jak i dla korpusu, było wdrożenie słownika walencyjnego Walenty. Wymagało to dostosowania gramatyki do zmienionego formatu zapisu typów fraz i do bardziej wyrafinowanej postaci schematów zdaniowych (por. p. 4.5). Konieczna była także aktualizacja istniejących drzew w Składnicy (Woliński 2015; Woliński *et al.* 2018).

Zastosowanie słownika Walenty zwiększyło liczbę wypowiedzeń akceptowanych przez analizator Świgra 2 z 13 194 (66% korpusu Składnicy) do 14 103 (70,5%). Wersje gramatyki, które posłużyły do uzyskania podanych wyników, różniły się jedynie elementami koniecznymi do użycia innej postaci słownika. Oczywiście w obu wypadkach część wygenerowanych lasów zostałaby odrzucona w całości przez dendrologów, jednak stosunek tych liczb pokazuje zysk z wdrożenia nowego słownika.

W procesie aktualizacji drzew Składnicy udało się automatycznie przenieść na nowe lasy składniowe 95,5% spośród poprzednio zatwierdzonych

drzew. Pozostałe 4,5% wypowiedzeń zostało skierowane do ponownego rozstrzygnięcia przez dendrologów. Dodatkowej pracy wymagało wprowadzenie w Składnicy fraz typów xp(...), które są bardziej szczegółowe niż wcześniej stosowane oznaczenie advp. W związku z tym trzeba było rozstrzygnąć około 300 niedopasowań oraz 200 nowych niejednoznaczności. W efekcie prac wprowadzono poprawki zarówno w drzewach Składnicy, jak i w schematach Walentego, a także w definicjach niektórych typów xp, dopuszczając nowe możliwości ich realizacji.

W chwili pisania tych słów (początek czerwca 2018) Składnica frazowa obejmuje 11 920 drzew zweryfikowanych przez dendrologów przy takiej samej zasadzie pracy jak w pierwszej fazie tworzenia korpusu. Można więc powiedzieć, że analizator Świgra 2 wygenerował poprawne drzewa dla 59,6% wypowiedzeń w całym korpusie Składnicy. Jeżeli odrzucić wypowiedzenia niepoprawne i błędy w opisie fleksyjnym, stwierdzi się, że uzyskano poprawne drzewa dla 62% wypowiedzeń zakwalifikowanych do analizy.

Na tym etapie trudno już wskazać wyraźne duże grupy wypowiedzeń z tym samym typem nieuwzględnionego w gramatyce problemu składniowego. Dalsze zwiększenie odsetka wypowiedzeń akceptowanych przez gramatykę wymaga więc rozstrzygnięcia bardzo wielu problemów o dużym stopniu szczegółowości.

6.7. Wyszukiwarka drzew składniowych

Ważnym narzędziem umożliwiającym wykorzystanie korpusu jest wyszukiwarka pozwalająca odwołać się do elementów znakowania w danym korpusie. Znakowany fleksyjnie korpus NKJP można przeszukiwać za pomocą przeszukiwarki Poliqarp (Przepiórkowski 2006), której język zapytań pozwala wyszukiwać formy fleksyjne o określonych wykładnikach, lematach i cechach gramatycznych oraz ciągi takich form. Analogicznie korpus składniowy powinien być wyposażony w wyszukiwarkę pozwalającą wyszukać struktury składniowe o zadanych cechach.

Przykładem wyszukiwarki składnikowych drzew składniowych jest Tiger Search (König i Lezius 2003; Lezius 2002; König *et al.* 2003). Jej język zapytań pozwala wyszukiwać drzewa zawierające wierzchołki o zadanych wartościach atrybutów i formułować warunki wskazujące sposób powiązania wierzchołków ze sobą z wykorzystaniem relacji następstwa liniowego oraz bycia potomkiem lub współskładnikiem. Dane Składnicy zostały zapisane w formacie wyszukiwarki Tiger Search[6].

[6] Składnica 0.5 zapisana w tym formacie jest dostępna na stronie http://zil.ipipan.waw.pl/Składnica wraz z obrazem płyty CD pozwalającej używać wyszukiwarki Tiger Search w systemie Windows. Niestety prace nad wyszukiwarką zostały później przerwane przez jej twórców.

Opracowano także wyszukiwarkę przystosowaną specjalnie do korpusu Składnica, która daje ciekawą możliwość znaną użytkownikom wyszukiwarki Poliqarp w odniesieniu do opisu fleksyjnego. Mianowicie w zapytaniach możliwe jest odwołanie się zarówno do drzewa składniowego wybranego przez dendrologów, jak i do wszystkich innych drzew wygenerowanych przez analizator Świgra 2 (wyszukiwarka działa więc na całym upakowanym lesie składniowym). Funkcja ta jest interesującym narzędziem nie tylko dla zewnętrznego użytkownika korpusu, ale może przede wszystkim dla osób pracujących nad jego rozwojem. Pozwala bowiem na przykład wyszukać konstrukcje, dla których istnieją zadane konkurencyjne interpretacje generowane przez analizator, i zbadać, na ile konsekwentnie dendrolodzy wybierali między nimi. Umożliwia to wykrywanie błędów pewnej klasy w ręcznym znakowaniu korpusu składniowego.

Wyszukiwarka (Woliński i Zaborowski 2012), zrealizowana w formie aplikacji internetowej, jest dostępna pod adresem http://treebank.nlp.ipipan.waw.pl/. Język zapytań jest wzorowany na zapytaniach Tiger Search. Istotnym rozszerzeniem języka względem oryginału jest uwzględnienie w języku kwantyfikatora egzystencjalnego (por. Marek *et al.* 2008) w postaci „nie istnieje wierzchołek taki, że..." (zob. dalej).

Realizacja wyszukiwarki była eksperymentem, mającym pokazać, na ile technologia relacyjnych baz danych tworzona do wyszukiwania informacji w dużych zbiorach może być wykorzystana do przetwarzania danych składniowych. Formułowane przez użytkownika zapytania są tłumaczone na SQL i przekazywane silnikowi bazy danych PostgreSQL. Taka koncepcja wyszukiwarki pozwoliła stworzyć ją (włącznie z modułem wizualizującym drzewa) w ramach pracy magisterskiej (Zaborowski 2010). Wynik można uznać za pozytywny: dla zapytań o 3–5 węzłów na wersji Składnicy obejmującej ok. 8200 zdań wyszukiwania w TIGER Search trwają 3–8 sekund, a w wyszukiwarce opartej na PostgreSQL – 5–15 sekund. Wyszukiwarka generuje więc wyniki ze sprawnością porównywalną z Tiger Search, w której indeksowanie było projektowane od zera specjalnie pod dany problem wyszukiwania. Należy przy tym zaznaczyć, że wyszukiwarka drzew Składnicy pracuje na danych o rząd wielkości większych niż Tiger Search, aby dać dostęp do pełnych lasów składniowych. Prawdopodobnie jest więc ona w istocie sprawniejsza od mechanizmu indeksującego Tiger Search.

Język zapytań

W dalszej części bieżącego podrozdziału przedstawiono język zapytań wyszukiwarki drzew w Składnicy frazowej. Obiektem, stanowiącym przedmiot wyszukiwania, są węzły drzew składniowych spełniające pewne warunki. Tak więc każde zapytanie musi się odnosić do co najmniej jednego węzła w drzewie składnikowym, a więc do co najmniej jednej jednostki nieterminalnej lub terminalnej gramatyki. Najprostsze zapytanie składa się z pary nawiasów kwa-

dratowych: `[]`. Dopasowuje się ono do dowolnego węzła w każdym drzewie, więc zbiorem wyników dla takiego zapytania jest cały korpus (zadanie tego zapytania pozwala więc sprawdzić liczbę wypowiedzeń w korpusie). Zbiór wyników można ograniczać poprzez dodawanie warunków określających atrybuty węzłów oraz zadawanie relacji między węzłami, co wyjaśniono w następnych punktach.

Atrybuty węzłów

Każdy wierzchołek drzewa jest charakteryzowany pewnym zbiorem atrybutów. Wartości atrybutów, wyróżniające poszukiwane wierzchołki, można zadawać w postaci równości wewnątrz nawiasów specyfikujących wierzchołek. Atrybut o nazwie `cat` daje dostęp do nazwy jednostki nieterminalnej. Tak więc następujące zapytanie znajduje wierzchołki reprezentujące wystąpienia jednostki **zdanie** w korpusie:

(9) `[ cat = "zdanie" ]`

Jeżeli zadana wartość atrybutu jest ciągiem liter i cyfr (nie zawierającym w szczególności odstępu), to znaki cudzysłowu można pominąć:

(10) `[ cat = zdanie ]`

W obrębie specyfikacji wierzchołka można umieścić wiele warunków określających wartości atrybutów, połączonych operatorami koniunkcji `&`, alternatywy `|` i negacji `!`. Odpowiednie połączenie warunków można wyrazić za pomocą nawiasów. Na przykład zapytanie (11) pozwala znaleźć frazy frazy nominalne i przymiotnikowe w bierniku:

(11) `[(cat=fno | cat=fpt) & przypadek=acc]`

To zapytanie można również wyrazić, łącząc spójnikiem logicznym wartości po prawej stronie znaku równości:

(12) `[cat = (fno | fpt) & przypadek=acc]`

Atrybuty przysługujące wierzchołkom terminalnym (formom fleksyjnym w liściach drzewa) obejmują:
- `orth` – segment (wykładnik tekstowy);
- `base` – lemat leksemu reprezentowanego przez dany segment;
- `tag` – znacznik fleksyjny formy reprezentowanej przez dany segment (według systemu przedstawionego w rozdziale 1);
- `pos` – część znacznika fleksyjnego przed pierwszym dwukropkiem, czyli klasa gramatyczna (np. `subst`, `adj`, `fin`, `praet`).

Atrybuty wierzchołków nieterminalnych (jednostek składniowych) to:
- `cat` – kategoria składniowa, czyli nazwa jednostki nieterminalnej;
- atrybuty fleksyjne dziedziczone od centrum jednostki, w zależności od jej typu: `przypadek`, `rodzaj`, `liczba`, `osoba`, `aspekt`, `czas`, `tryb`, `stopień`;

– atrybuty czysto składniowe (por. p. 4.4) takie jak: `rekcja` – lista typów zrealizowanych argumentów, `tfw` – typ frazy wymaganej, `wyróżnik` fleksyjny, `dest` – predestynacja, `ink` – inkorporacja, `neg` – negacja. Ich obecność jest zależna od kategorii składniowej[7].

Następujące atrybuty przysługują wszystkim wierzchołkom (terminalnym i nieterminalnym):

– `nid` – arbitralny ale unikalny identyfikator wierzchołka;
– `from` – identyfikator pozycji w tekście, od której zaczyna się wystąpienie danej jednostki;
– `to` – identyfikator pozycji w tekście, na której kończy się wystąpienie danej jednostki;
– `depth` – odległość danego wierzchołka od korzenia drzewa (korzeń jest na głębokości 0);
– `terminal` – czy wierzchołek jest terminalem;
– `sel` – czy dany wierzchołek należy do drzewa wybranego przez dendrologów (zob. dalej punkt o strukturach niejednoznacznych).

Odwołując się do wymienionych identyfikatorów, można na przykład formułować warunki typu „jednostki zaczynają/kończą się w tym samym miejscu". Ostatnie dwa atrybuty są binarne – ich wartości to `true` i `false`.

Wyrażenia regularne

Wartość po prawej stronie znaku równości w specyfikacji atrybutu może także być ujętym w ukośniki wyrażeniem regularnym. Tak sformułowany warunek jest spełniony, jeżeli wyrażenie regularne da się dopasować do całej wartości atrybutu. Na przykład następujące zapytanie wyszukuje segmenty kończące się literami *ąc*:

(13) `[ orth = /.*ąc/ ]`

Wyrażeń regularnych można używać w odniesieniu do dowolnych atrybutów. Następujące zapytanie dotyczy form zaliczonych do jednej z klas zaimków osobowych (`ppron12` i `ppron3`), które można uchwycić łącznie, stawiając warunek, żeby wartość klasy gramatycznej zaczynała się ciągiem `ppron`:

(14) `[ pos = /ppron.*/ ]`

W kolejnym przykładzie przedstawiono sposób wyszukania segmentów oznaczonych na poziomie fleksyjnym jako formy rzeczowników w bierniku (znacznik zaczyna się `subst:` i zawiera dalej fragment `:acc:`):

(15) `[ tag = /subst:.*:acc:.*/ ]`

[7] Przy formułowaniu zapytania najłatwiej jest chyba sprawdzić dopuszczalne wartości poszczególnych atrybutów w drzewach Składnicy (np. wyszukanych pytaniem bardziej ogólnym).

Relacje między węzłami

Również specyfikacje wierzchołków można łączyć operatorami i nawiasami w bardziej skomplikowane wyrażenia. Do poniższego zapytania pasują drzewa zawierające zarówno frazę nominalną **fno**, jak i przymiotnikową **fpt** (bez wymagania konkretnej konfiguracji tych fraz względem siebie):

(16) `[cat=fno] & [cat=fpt]`

Ciekawsze zależności można oddać za pomocą operatorów wyrażających relacje strukturalne między węzłami. Należy do nich przede wszystkim relacja bycia dzieckiem w drzewie `>` i jej domknięcie przechodnie `>*` (relacja bycia potomkiem. Mogą one na przykład posłużyć do wyszukania fraz nominalnych **fno** zawierających jako składnik bezpośredni frazę przymiotnikową **fpt**:

(17) `[cat=fno] > [cat=fpt]`

lub fraz nominalnych zawierających gdzieś w swojej strukturze zdanie:

(18) `[cat=fno] >* [cat=zdanie]`

Wszystkie dostępne operatory określające relacje na wierzchołkach zestawiono w tabeli 6.3. Te z nich, które zapisane są infiksowo, można łączyć w łańcuchy. Na przykład następujące wyrażenie wyszukuje **zdania**, które zawierają frazę wymaganą **fw**, która jest realizowana przez frazę przymiotnikową **fpt**:

(19) `[cat=zdanie] > [cat=fw] > [cat=fpt]`

Operator `arity(A, N, M)` pozwala nałożyć warunek na liczbę składników bezpośrednich jednostki (czyli arność węzła). Warunek ten jest spełniony, jeżeli węzeł A ma nie mniej niż N i nie więcej niż M składników bezpośrednich. Następujące zapytanie znajduje wystąpienia jednostki **zdanie** o co najmniej 6 składnikach bezpośrednich:

(20) `arity([cat=zdanie], 6, 1000)`

Zmienne

Zmienne pozwalają nadać nazwę pewnemu elementowi zapytania, aby wskazać, że musi on być identyczny z innym elementem. Można ich użyć na dwa sposoby: w odniesieniu do wartości atrybutów oraz do specyfikacji wierzchołków. Nazwy zmiennych wprowadzane są znakiem `#`.

Pierwszy sposób użycia ilustruje następujący przykład:

(21) `[orth=#w: /.*zny/ & base=#w]`

Na atrybut `orth` nałożono wymaganie, aby pasował do wzorca `/.*zny/` (a więc, aby segment kończył się na *-zny*). Jednocześnie wartości atrybutu `orth` nadano

Tabela 6.3. Operatory języka wyszukiwania wyrażające relacje strukturalne między węzłami w drzewie

`A > B`	A jest rodzicem B w pewnym drzewie, czyli B jest składnikiem bezpośrednim A
`A >* B`	A jest przodkiem B, czyli B jest składnikiem A, czyli B należy do pewnego poddrzewa A (B != A)
`A >R B`	w pewnym drzewie A jest rodzicem B, przy czym wierzchołek wyprowadzono z użyciem reguły o nazwie (symbolu) R
`A >N B`	B należy do pewnego poddrzewa A, ścieżka A–B ma długość N
`A >M,N B`	B należy do pewnego poddrzewa A, ścieżka A–B ma długość między M i N
`A >@l B`	skrajny lewy potomek: B należy do pewnego poddrzewa A i żadna jednostka w tym poddrzewie nie poprzedza B
`A >@r B`	skrajny prawy potomek: B należy do pewnego poddrzewa A i B nie poprzedza żadnej jednostki w tym poddrzewie
`A >@c B`	centralny potomek: B należy do pewnego poddrzewa A i istnieje ścieżka z A do B prowadząca wyłącznie przez centra fraz
`A . B`	A bezpośrednio poprzedza B, czyli fraza B zaczyna się tam, gdzie kończy się A
`A .* B`	A poprzedza B, czyli B zaczyna się dalej niż A się kończy
`A .N B`	B zaczyna się N segmentów od końca A
`A .M,N B`	B zaczyna się między M a N segmentów od końca A
`A $ B`	rodzeństwo: A i B mają w pewnym drzewie wspólnego rodzica
`A $.* B`	rodzeństwo i poprzedzanie: A i B mają w pewnym drzewie wspólnego rodzica i A występuje przed B
`root(A)`	A jest korzeniem lasu (A nie ma rodzica)
`same_tree(A, B, …)`	istnieje drzewo zawierające wymienione wierzchołki (operator ten jest istotny przy pracy ze strukturami niejednoznacznymi, zob. dalej)

nazwę `#w`. W dalszym składniku koniunkcji zawarte jest wymaganie, aby atrybut `base` miał tę samą wartość. W sumie oznacza to, że wykładnik tekstowy zakończony na *-zny* ma być równy lematowi leksemu.

Użycia zmiennych nie muszą występować w obrębie specyfikacji tego samego wierzchołka. W następnym zapytaniu poszukiwane są dwa wystąpienia tego samego segmentu obok siebie (relacja `.`):

(22) `[orth=#s] . [orth=#s]`

Drugi sposób użycia zmiennych pozwala wyspecyfikować udział konkretnego wierzchołka w większej liczbie relacji. Na przykład w poniższym zapytaniu poszukiwane są wierzchołki kategorii **zdanie**, które mają jako składniki bezpośrednie jednocześnie: frazę wymaganą oznaczoną jako podmiot subj, frazę wymaganą nominalną typu np(accgen) oraz frazę wymaganą dowolnego podtypu xp(_). W tym celu wierzchołek reprezentujący zdanie zostaje oznaczony zmienną `#z`, aby mógł być użyty jeszcze dwukrotnie w specyfikacjach kolejnych relacji:

(23) `#z: [cat=zdanie] > [tfw=/subj.*/] &`
`#z > [tfw="np(accgen)"] & #z > [tfw=/xp.*/]`

W kontekście zapytań z udziałem wielu węzłów każda para nawiasów kwadratowych implicite zadaje inny węzeł. Tak więc poniższe zapytanie znajduje węzły, które mają co najmniej dwa składniki bezpośrednie:

(24) `#w: [] > [] & #w > []`

Praca ze strukturami niejednoznacznymi

Przedstawione dotychczas zapytania operowały wyłącznie na strukturach składniowych Składnicy wybranych przez dendrologów. Jednak dla wyszukiwarki dostępne są wszystkie struktury wygenerowane przez analizator Świgra 2. Do należących do nich wierzchołków można się odwołać, specyfikując wierzchołek w podwójnych nawiasach kwadratowych `[[]]`. Tak więc następujące zapytanie wyszukuje wszystkie drzewa, dla których analizator Świgra 2 dopuścił istnienie frazy nominalnej w bierniku, niezależenie od tego, czy dendrolodzy tę właśnie strukturę wybrali:

(25) `[[cat=fno & przypadek=acc]]`

Zapis w pojedynczych nawiasach jest w istocie skrótem notacyjnym równoważnym dodaniu warunku, że wierzchołek ma ustawiony atrybut `sel`, sygnalizujący, że należy on do drzewa wybranego przez językoznawców. Tak więc zapytanie

(26) `[cat=fno & przypadek=acc]`

jest równoważne

(27) `[[cat=fno & przypadek=acc & sel]]`

Następujące zapytanie pozwala wyszukać frazy wymagane, które zostały przez językoznawców uznane za przyimkowe (prepnp), dla których analizator Świgra 2 proponował także interpretację jako frazy typu xp(_):

(28) `[tfw=/prepnp.*/ & from=#f & to=#t]`
`& [[tfw=/xp.*/ & from=#f & to=#t]]`

Użycie zmiennych oraz atrybutów `from` i `to` zapewnia, że dwie znalezione jednostki są rozpięte (w różnych drzewach) nad tym samym fragmentem tekstu.

W zapytaniu (28) chodziło o znalezienie fragmentów struktury należących do dwóch różnych wariantów interpretacyjnych struktury, a więc dwóch różnych drzew. Czasami zachodzi jednak potrzeba wyszukania wierzchołków, które należą do tego samego drzewa, chociaż drzewo to nie zostało wybrane przez dendrologów. Warunek przynależności do tego samego drzewa można zadać za pomocą predykatu `same_tree(…)`, którego argumentami powinny być specyfikacje co najmniej dwóch wierzchołków.

W oryginalnym języku TIGER Search wszystkie specyfikacje wierzchołków są poprzedzone implicite kwantyfikatorem szczegółowym. Wyrażają warunek „istnieje taki wierzchołek, że...". Nie da się zaś wyrazić warunku „nie istnieje taki wierzchołek, że..." ani „wszystkie wierzchołki mają cechę...". W podręczniku wyszukiwarki TIGER można znaleźć stwierdzenie, że dodanie takiej możliwości miałoby niekorzystny wpływ na efektywność wykonywania zapytań (König *et al.* 2003). Język TIGER został jednak w późniejszej wersji rozszerzony o kwantyfikator ogólny (Marek *et al.* 2008). W wyszukiwarce Składnicy uznano za bardziej naturalne formułowanie warunków w postaci „nie istnieje taki wierzchołek, że...".

Operator negacji użyty w obrębie specyfikacji wierzchołka wybiera drzewa posiadające wierzchołek niepasujący do negowanego kryterium. Operator negacji `!` użyty na zewnątrz specyfikacji wierzchołka (czyli na zewnątrz nawiasów kwadratowych) oznacza wymaganie, aby w lesie nie było wierzchołków pasujących do warunków znajdujących się w zasięgu tego operatora.

Na przykład następujące zapytanie znajduje frazy nominalne zawierające jako bezpośredni składnik frazę przymiotnikową i niezawierające nigdzie w swojej strukturze zdania:

(29)
```
#n: [cat=fno] > [cat=fpt] & !( #n >* [cat=zdanie] )
```

W zasięgu operatora negacji przywoływany jest wierzchołek `#n`, który jednak zdefiniowany jest wcześniej w warunku pozytywnym, jest on więc kwantyfikowany egzystencjalnie. Zapytanie to czyta się następująco: istnieje wierzchołek `#n` o własnościach określonych w pierwszej części zapytania i nie istnieje wierzchołek `[cat=zdanie]`, który byłby składnikiem `#n`.

Poprawne zapytanie nie może się składać wyłącznie z wierzchołków wyspecyfikowanych w zasięgu operatora „nie istnieje".

Ten mechanizm można wykorzystywać również w odniesieniu do struktur niejednoznacznych i na przykład poszukiwać zdań, w których dendrolodzy uznali, że nie ma wyrażonego jawnie podmiotu, mimo że wśród dostępnych alternatyw istnieje interpretacja z podmiotem:

(30)
```
#z: [cat=zdanie] > [[tfw=/subj.*/]]
    & (! #z > [tfw=/subj.*/] )
```

6.8. Przykłady wykorzystania wyszukiwania w Składnicy

Dostępność korpusu składniowego Składnica i wyszukiwarki drzew otwiera nowe możliwości badań. Anegdotyczną ilustracją użyteczności Składnicy

może być przykład (31) na stronie 40, zawierający frazy finitywne z centrami we wszystkich trzech osobach, który znaleziono za pomocą wyszukiwarki drzew, zadając zapytanie:

(31)
```
[cat=ff & osoba=1] & [cat=ff & osoba=2]
    & [cat=ff & osoba=3]
```

Zadając pytanie (32) z różnymi wartościami ograniczeń można ustalić, jak wiele składników jest możliwych w zdaniu elementarnym:

(32)
```
#n: [cat=zdanie] > [cat=ff] & arity(#n,7,1000)
```

Wyniki dla zapytania zadającego arność co najmniej 7 obejmują 16 zdań zaakceptowanych przez dendrologów. Nie ma wyników dla zapytania o arność co najmniej 8. Informacja taka może posłużyć do optymalizacji analizatora poprzez ograniczenie liczby podrzędników, które program próbuje podłączyć do danego nadrzędnika.

Analogiczne pytanie dla fraz nominalnych pozwala znaleźć frazę nominalną o największej w korpusie liczbie składników:

(33)
```
#n: [cat=fno] > [cat=fno] & arity(#n,7,1000)
```

Niestety liczbę tę zawyżają przecinki ortograficzne, które stają się składnikami bezpośrednimi frazy. Faktycznych składników jest w owej frazie 5:

(34) *Są to jedynie najważniejsze, opisane dość pobieżnie, zranienia w dziedzinie seksualnej, które wymagają uzdrowienia.* [Skł.]

Można zaproponować wiele podobnych pytań sprawdzających, jakie realizacje poszczególnych typów fraz faktycznie wystąpiły w korpusie. Pytań takich można użyć do badania spójności korpusu oraz optymalizacji analizatora składniowego.

Pozycja podrzędnika przymiotnikowego

Za pomocą wyszukiwarki drzew można także na przykład przeprowadzić badanie pozycji podrzędników przymiotnikowych we frazach nominalnych. Zapytanie (35) pozwala wyszukać frazy nominalne `#n`, które mają zarówno składnik nominalny `#rzecz` (będący nadrzędnikiem), jak i składnik przymiotnikowy `#przym`, przy czym składnik przymiotnikowy poprzedza w porządku segmentów składnik nominalny:

(35)
```
#n: [cat=fno] > #przym: [cat=fpt] &
#n > #rzecz: [cat=fno] & #przym .* #rzecz
```

Zamiana kolejności elementów połączonych operatorem `.*` pozwala znaleźć analogiczne frazy, w których składnik przymiotnikowy następuje po nadrzędniku nominalnym.

przymiotnik w prepozycji
liczba wyników: **3817**

...wzrastających w **szybkim tempie** zbiorów.
...każdego **swoim zimnym tchnieniem**.
...wrogów, ani **wiecznych przyjaciół**.
Często oglądanym w tych dniach obrazkiem byli...
Są już nawet meble do **tych wnętrz**.
...miałam **takie ambicje**...
...wjeżdża **długa, biała limuzyna**.
...nie miały **wysokich stołków**...
...chwalona w **tym kształcie**, nie będzie
...odnotowała **białostocka policja**.
...w **innych gminach** jest podobnie.
...

przymiotnik w postpozycji
liczba wyników: **1818**

Budowę **murów aureliańskich**...
...byli **ludzie wylewający wiadrami wodę z piwnic, komórek i garaży**.
...na **wystawie przemysłowej**, która...
...**ultimatum postawionego przez porywaczy**.
...lecz **fakty składające się na polską rzeczywistość** działają nam na nerwy.
...skutkiem **decyzji nacjonalizacyjnych**.
...akademii **sztuk pięknych**...
...na **bocznicach kolejowych**...
...wielkość **akumulacji zimowej**.
Dobrze znosi **warunki miejskie**.
...

Rysunek 6.7. Pozycje podrzędników przymiotnikowych w Składnicy

Na rysunku 6.7 przedstawiono fragment wyników dla takich zapytań. Jak można zobaczyć, wystąpienia przymiotnika przed nadrzędnikiem są zauważalnie częstsze, można więc powiedzieć, że szyk, w którym przymiotnik poprzedza rzeczownik, jest w jakimś sensie preferowany. Uwaga ta nie dotyczy jednak fraz przymiotnikowych z centrum będącym imiesłowem przymiotnikowym. Dla takich fraz naturalniejsza jest pozycja za centrum nominalnym.

Widać także różnicę w funkcji podrzędnika z centrum będącym faktycznym przymiotnikiem. Podrzędnik taki umieszczony przed nadrzędnikiem jest modyfikatorem, określa cechę pewnego obiektu (*zimne tchnienie*, *wysokie stołki*). Umieszczony za nadrzędnikiem nominalnym tworzy całostkę terminologiczną lub dookreśla, o jaki rodzaj obiektu nazwanego przez nadrzędnik chodzi (*sztuki piękne*, *bocznice kolejowe*, *warunki miejskie*). Oczywiście nie są to rozróżnienia ścisłe, ale różnice między znalezionymi frazami są dość wyraziste.

6.9. Przykłady kwantytatywnych zastosowań Składnicy

Korpus składniowy może stanowić także przedmiot badań kwantytatywnych. Z punktu widzenia inżynierii lingwistycznej korpus taki można przede wszystkim zastosować do budowy modeli statystycznych, na przykład służących do ujednoznaczniania drzew składniowych (por. rozdz. 7), ale może także być przedmiotem badań językoznawczych.

Wypada w tym miejscu zastrzec, że Składnica ma skrzywienie statystyczne wynikające z tego, że część poprawnych wypowiedzeń została odrzuco-

na przez analizator automatyczny, pod pewnymi względami nie jest więc reprezentatywna statystycznie. Jednak w odniesieniu do pewnych konstrukcji można zakładać, że rozkład częstości jest niezaburzony. Ryzykowne byłoby wyciąganie wniosków na temat częstości typów fraz nieciągłych, jako że zjawisko to zostało dotychczas opisane w sposób niepełny. Można jednak mieć zaufanie do własności statystycznych typów zdań i fraz w obrębie zdań elementarnych.

Jako dwa przykłady zjawisk, których reprezentacja frekwencyjna powinna być adekwatna, wybrano konstrukcje skoordynowane i szyk zdań elementarnych. Badania przeprowadzono na wersji Składnicy frazowej z początku czerwca 2018 roku, liczącej 11 920 zatwierdzonych drzew składniowych.

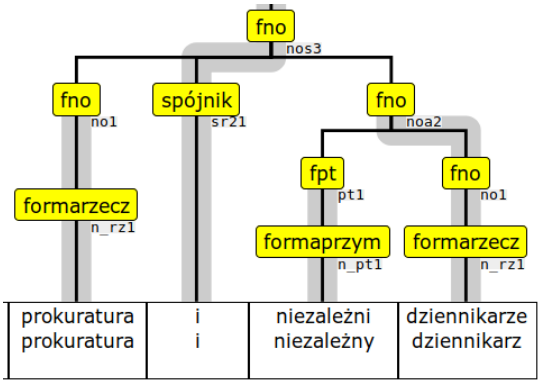
Frekwencje konstrukcji skoordynowanych

W tym punkcie przedstawiono analizę, której celem jest sprawdzenie, na ile wprowadzenie do gramatyki Świgra 2 współrzędnych (skoordynowanych) realizacji poszczególnych typów składników było istotne dla tworzonego korpusu składniowego.

W tabeli 6.4 zestawiono liczby wystąpień różnych typów składników w zatwierdzonych drzewach Składnicy. W drugiej kolumnie podano liczbę wszystkich wystąpień składników danego typu, przy czym zliczane są tylko składniki o różnych centrach. W trzeciej kolumnie podano liczby nietrywialnych konstrukcji podrzędnych, a więc zawierających nadrzędnik i co najmniej jeden podrzędnik. W kolumnie czwartej – liczby konstrukcji odpowiedniego typu z centrami w postaci spójnika współrzędnego. Według tych zasad konstrukcja *prokuratura i niezależni dziennikarze*, której odpowiada poddrzewo zawierające cztery wystąpienia jednostki **fno** (zob. rys. 6.8), wnosi dwa wystąpienia do kolumny „wszystkich" (bo zawiera dwa centra nominalne *prokuratura* i *dziennikarze*), a ponadto jedno wystąpienie frazy z podrzędnikiem *niezależni dziennikarze* i jedno wystąpienie frazy z koordynacją (z centrum *i*).

Tabela 6.4. Częstości składników różnych typów w Składnicy frazowej

	wszystkich	z podrzędnikami	z koordynacją
zdanie	16019	15628	1608
fraza werbalna **fwe**	18480	2339	273
fraza nominalna **fno**	43085	23234	1326
fraza przymiotnikowa **fpt**	14554	2164	354
fraza przysłówkowa **fps**	4245	422	32
fraza przyimkowa **fpm**	13148	364	136
fraza zdaniowa **fzd**	2739	42	30
fraza wymagana **fw**	31481	–	2



Rysunek 6.8. Struktura frazy nominalnej *prokuratura i niezależni dziennikarze*

Zgromadzone dane pokazują, że zdania finitywne z zasady są rozbudowanymi konstrukcjami podrzędnymi – tylko 391 zdań, czyli 2,4%, redukuje się do wystąpienia frazy finitywnej. Podrzędniki towarzyszą też około połowie centrów fraz nominalnych. W innych typach fraz przeważają (w różnym stopniu) konstrukcje nierozbudowane, a więc realizowane przez jedną formę fleksyjną (lub właściwą minimalną realizację w wypadku fraz przyimkowych, zdaniowych i wymaganych).

Konstrukcje współrzędne są generalnie o rząd wielkości rzadsze od konstrukcji podrzędnych. Wyjątkiem są tu frazy przyimkowe i zdaniowe, w których każda realizacja rozbudowana jest na tyle nietypowa, że rozróżnienie to się zaciera.

Koordynacja konstrukcji nominalnych jest w liczbach absolutnych mniej więcej tak samo częsta jak koordynacja na poziomie zdania, chociaż fraz nominalnych jest w strukturach wyraźnie więcej niż zdań, więc mniejszy ułamek fraz nominalnych zawiera koordynację.

Sumaryczna liczba konstrukcji niezdaniowych z koordynacją przekracza liczbę skoordynowanych zdań. Sumaryczna liczba konstrukcji współrzędnych w konstrukcjach z centrum czasownikowym (**zdanie** i **fwe**) jest mniej więcej równa sumarycznej liczbie konstrukcji współrzędnych we wszystkich pozostałych typach fraz.

Liczby te potwierdzają, że opisanie konstrukcji współrzędnych niezdaniowych jest równie ważne jak konstrukcji zdaniowych. Tak więc uwzględnienie różnych typów konstrukcji skoordynowanych w gramatyce Świgra 2 stanowi uzupełnienie ważnej luki w GFJP.

Komentarza wymaga zapewne jeszcze ostatni wiersz tabeli opisujący frazy wymagane. Ich realizacje skoordynowane odpowiadają koordynacji fraz różnych typów w obrębie jednej pozycji składniowej (por. p. 2.9.4). Wymienione w tabeli dwie konstrukcje to następujące zdania:

(36) *Mówią , że mam własne życie i żebym go pilnował .*

(37) *Nazywają ich tu „Żydami góralami", czasami wyśmiewają ich surową religijność, zamknięcie na świat i to, że wybrali handel zamiast bardziej „żydowskiej" tradycji intelektualnej.*

Struktury takie zostały wprowadzone do gramatyki późno, dopiero na etapie zmiany słownika walencyjnego, więc bieżący stan Składnicy niekoniecznie odzwierciedla w pełni ich występowanie. Można jednak odczytać go jako sugestię, że jest to zjawisko bardzo rzadkie. Nie zmienia to oczywiście tego, że takie konstrukcje są możliwe, a ich uwzględnienie w gramatyce było potrzebne, aby w pełni wykorzystać wyrafinowaną postać schematów składniowych słownika Walenty.

Szyk zdania elementarnego

Korzystając ze Składnicy, można powtórzyć na większym materiale badania szyku zdania elementarnego, jakie na bazie korpusu wypowiedników (Świdziński 1996) opisała Derwojedowa (2000). Badania Świdzińskiego polegały na ręcznej analizie według GFJP wypowiedzeń pochodzących z korpusu *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz *et al.* 1990; zob. także Bień i Woliński 2003; Ogrodniczuk 2003b,a). Uzyskana baza danych zdaje sprawę ze struktury składniowej jednostek poziomu zdaniowego zbioru 4484 wypowiedzeń polskich. Zanalizowano strukturę wypowiedzeń z centrum spójnikowym oraz zdań elementarnych i oznajmień elementarnych (konstrukcji bez centrum finitywnego). Łącznie zdania i oznajmienia są w tej pracy nazywane wypowiednikami. Przeprowadzona analiza nie schodzi na poziom fraz składnikowych.

Tabela 6.5 przedstawia najczęstsze warianty szyku zdań elementarnych (stanowiące co najmniej 1% wystąpień w Składnicy). Zestawiono wyniki przedstawione przez Derwojedową (kolumna SFPW) z analogicznymi badaniami przeprowadzonymi na danych Składnicy frazowej. Z korpusu wyizolowano wszystkie zdania elementarne z centrum finitywnym (w wypowiedzeniach złożonych jest zawarte kilka zdań elementarnych). Szyk każdego z nich zapisano symbolami o następującym znaczeniu: V – fraza finitywna; S – fraza wymagana oznaczona jako podmiot; O – inna fraza wymagana; frazy luźne są pomijane. W zapisie szyku zdań ze Składnicy osobno oznaczone zostały także: fraza wymagana realizowana przez *się* (oznaczona literą s, tylko to oznaczenie wystąpiło wśród wariantów przedstawionych w tabeli), jednostka **posiłk** będąca fragmentem nieciągłej frazy finitywnej i jednostka ξ stanowiąca fragment nieciągłej frazy wymaganej.

W korpusie wypowiedników partykuła *się* była traktowana jako część frazy finitywnej w wypadku *się* inherentnego (*bali się*) oraz jako osobna fraza wymagana w wypadku innych użyć (np. *myła się = myła siebie*). W Składnicy wszystkie wystąpienia *się* były interpretowane jako osobna fraza wymagana. Dlatego w zestawieniu w tabeli 6.5 *się* jest w kolumnie Składnicy oznaczane osobno.

Tabela 6.5. Najczęstsze warianty szyku zdań elementarnych z centrum finitywnym

	SFPW		Składnica	
łącznie zdań elem.	4839		16019	
szyk VO	1193	24,65%	3921	24,48%
SVO	993	20,52%	3611	22,54%
OVS	319	6,59%	1166	7,28%
V	439	9,07%	826	5,16%
VS	113	2,34%	704	4,39%
OV	234	4,84%	620	3,87%
SV	250	5,17%	611	3,81%
VOO	188	3,89%	528	3,30%
SVsO			389	2,43%
VsO			378	2,36%
SVOO	130	2,69%	370	2,31%
OVO	112	2,31%	287	1,79%
VSO	134	2,77%	277	1,73%
VOS	104	2,15%	229	1,43%
SOV			213	1,33%
OVsS			175	1,09%
VsS			173	1,08%

Doliczenie go do odpowiednich klas analogicznie do badania Derwojedowej mogłoby nieco zmienić kolejność wierszy w tabeli.

Różnice mogą też dotyczyć uznawania poszczególnych składników za nieciągłe w obu badaniach, jednak zjawiska związane z nieciągłością okazały się na tyle rzadkie, że nie wpłynęły na wyniki zestawione w tabeli.

Jak widać, badanie powtórzone na ponad trzykrotnie większym korpusie, zawierającym inaczej dobrane teksty (NKJP1M), daje bardzo zbliżoną listę najczęstszych wariantów szyku. W tabeli uporządkowano wyniki według malejącej częstości w Składnicy. Dwa pierwsze co do częstości warianty są takie same w obu badaniach: VO i SVO. Potwierdzają więc one w szczególności obserwację Derwojedowej, że najczęstszym szykiem jest VO, która zaprzecza wcześniejszemu poglądowi Klemensiewicza, że preferowanym szykiem jest SVO (por. Derwojedowa 2000, s. 55). Następne dwie pary wierszy w obu badaniach wystąpiły w odwrotnej kolejności. Są one o tyle ciekawe, że różnice między tymi wierszami w obu badaniach są wyraźne. Wyższa pozycja szyku V w korpusie wypowiedników mogłaby być wytłumaczona tym, że zawiera on część wystąpień oznaczonych w Składnicy jako sV i Vs. W sumie jednak wyniki obu badań są bardzo spójne, w szczególności osiem wariantów szyku wymienionych jako najczęstsze powtarza się w obu badaniach.

7

Statystyczne ujednoznacznianie analiz składniowych

Oczywistą wadą regułowego analizatora składniowego jest generowanie dla danego wypowiedzenia wszystkich możliwych interpretacji składniowych. Pożądane byłoby wskazanie jednej konkretnej struktury dla wypowiedzenia. Analizator regułowy modeluje wyłącznie kwestie fleksyjne i formalnoskładniowe. Tymczasem w komunikacji językowej człowiek ujednoznacznia otrzymywany komunikat nie tylko ze względu na te cechy, ale także biorąc pod uwagę jego znaczenie, do czego wykorzystuje swoją wiedzę o świecie. W zastosowaniach praktycznych automatycznej analizy składniowej potrzebne jest analogiczne ujednoznacznienie lasu składniowego generowanego przez analizator.

Jednym z celów stworzenia korpusu składniowego Składnica, omówionego w poprzednim rozdziale, było uzyskanie danych umożliwiających tworzenie tego rodzaju modeli statystycznych. W tym rozdziale zostaną omówione dwa sposoby wykorzystania tych danych do stworzenia statystycznych modeli ujednoznaczniających lasy składniowe, a więc wybierających spośród drzew wygenerowanych przez analizator Świgra 2 jedno drzewo o najwyższym prawdopodobieństwie. W obu wypadkach zadanie jest postawione w ten sposób, że algorytm ma wykonać tę samą pracę, co anotatorzy Składnicy. Celem jest więc wybranie dla niejednoznacznych wierzchołków lasu składniowego jednej z dostępnych realizacji danego wierzchołka. Pierwszy z omawianych modeli opiera się na koncepcji probabilistycznych gramatyk bezkontekstowych (PCFG, Probabilistic Context Free Grammars), drugi – na technice modelowania maksimum entropii (ME, Maximum Entropy).

Warto zauważyć, że modele te „uczą się" na podstawie cech dostępnych w danych, a więc cech fleksyjnych i składniowych. Decyzje podjęte przez anotatorów danych treningowych są skutkiem wykorzystania ich całej wiedzy o tekście. Proces uczenia może więc wykorzystać ewentualne korelacje cech formalnych z własnościami wykraczającymi poza te cechy.

7.1. Sposób oceny jakości modeli

W pracach dotyczących statystycznej analizy składniowej przyjmuje się zwykle ocenę jakości modeli poprzez badanie poprawności opisu poszczegól-

nych węzłów w drzewie (Abney *et al.* 1991). Badana jest więc poprawność wyróżnienia poszczególnych składników i poprawność przypisania im cech gramatycznych, przy czym wszystkie wierzchołki traktowane są z tą samą wagą.

Stosowanymi miarami jakości są dokładność (ang. *precision*, *P*) i pełność (ang. *recall*, *R*):

$$P = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków wybranych przez algorytm}}$$

$$R = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków w danych treningowych}}$$

Dokładność równa 1 oznaczałaby, że algorytm wybrał tylko te wierzchołki, które powinien wybrać. Pełność równa 1 oznaczałaby, że algorytm wybrał wszystkie wierzchołki, które powinien.

W przedstawionych eksperymentach wartości dokładności i pełności różnią się od siebie o nie więcej niż 3 punkty procentowe. Wynika to z tego, że drzewa wybierane przez algorytm mają liczbę wierzchołków zbliżoną do drzew „idealnych". Gdy algorytm wybiera niewłaściwy wierzchołek (co pogarsza dokładność), zwykle jednocześnie pomija wybór jakiegoś właściwego wierzchołka (co pogarsza pełność). Dlatego ocenę wyników można uprościć poprzez podawanie tylko łącznej charakterystyki – tzw. miary *F*, czyli średniej harmonicznej dokładności i pełności:

$$F = \frac{2PR}{P + R}$$

Ocena wyboru poszczególnych wierzchołków była dokonywana na dwa sposoby. Ocena bardziej zgrubna, oznaczana dalej F_N, sprawdza jedynie, czy zgadza się zasięg danego składnika i przypisana nazwa jednostki nieterminalnej. Ocena dokładniejsza, F_A, wymaga zgodności wszystkich atrybutów przypisanych danemu składnikowi. Oceniane są wyłącznie węzły wewnętrzne drzew, ponieważ w rozważanym korpusie analiza fleksyjna została przeprowadzona w sposób jednoznaczny, więc nie ma niejednoznaczności dotyczących terminali.

Dzięki oznaczeniom centrów składniowych drzewa Składnicy można skonwertować na drzewa zależnościowe. Dzięki temu w eksperymentach podawana jest jeszcze jedna miara: ULAS (ang. *unlabelled attachment score*), czyli procent właściwie wybranych krawędzi w drzewie zależnościowym. W wypadku drzew zależnościowych liczba wszystkich krawędzi jest znana z góry, tak więc ta jedna liczba jest wspólną miarą dokładności i pełności. Miara ta jest liczona na wynikach konwersji opartej wyłącznie na oznaczeniach centrów w Składnicy frazowej, nie zaś na Składnicy zależnościowej, której konwersja zawierała elementy heurystyczne, mogące zakłócić wyniki oceny. Nie wpływa to na możliwość wykorzystania tak obliczonej wartości ULAS do porównania wyników proponowanych tu metod z wynikami statystycznych parserów zależnościowych (Wróblewska i Woliński 2012; Wróblewska 2014).

Eksperymenty przedstawione w punkcie 7.2 zostały przeprowadzone równolegle na dwóch wersjach Składnicy. Wersja pierwsza, oznaczona Składnica 0.5, liczy 8227 pełnych drzew i odpowiada końcowi pierwszej fazy jej tworzenia (zob. p. 6.6). Druga, oznaczona Składnica 18.06, to bieżąca wersja rozwojowa z początku czerwca 2018, obejmująca 11 920 drzew zaakceptowanych przez dendrologów. Druga wersja jest większa, ale jednocześnie zawiera bardziej skomplikowane wypowiedzenia, z konstrukcjami nieuwzględnionymi w pierwszym etapie tworzenia, w szczególności konstrukcjami bez centrum finitywnego. Eksperymenty opisane w punkcie 7.3 przeprowadzono wyłącznie na nowszej wersji Składnicy.

Należy zaznaczyć, że sytuacja, w której realizowano eksperyment, jest nieco wyidealizowana z powodu wykorzystania ujednoznacznionych ręcznie interpretacji fleksyjnych. Odpowiada to dokładnie zadaniu wykonywanemu w systemie Dendrarium, przedstawione wyniki pokazują więc, jaką jakość można uzyskać w automatycznej analizie składniowej pozostałej części korpusu NKJP1M. Przejście do analizy tekstu bez ręcznego ujednoznacznienia fleksyjnego wymagałoby użycia tagera, którego pomyłki nakładałyby się na błędy ujednoznacznienia drzew, lub ujednoznaczniania drzew wygenerowanych na nieujednoznacznionych danych fleksyjnych. Tego rodzaju interakcje i strategie radzenia sobie z nimi będą przedmiotem dalszych badań. Interesująca jest w szczególności kwestia, czy model ujednoznaczniający drzewa jest w stanie lepiej wybierać interpretacje fleksyjne niż tager.

We wszystkich eksperymentach stosowana była dziesięciokrotna walidacja krzyżowa, a więc przedstawiane wyniki są średnią z 10 eksperymentów, w których ⁹⁄₁₀ zdań służyło jako dane treningowe, a pozostała ¹⁄₁₀ korpusu składniowego – jako dane testowe.

Punkt odniesienia: małpa w Dendrarium

Ocenę wyników narzędzia statystycznego ułatwia podanie punktu odniesienia, który pokazywałby, na ile trudne jest dane zadanie, ile można osiągnąć bardzo prostymi metodami, a o ile lepsze wyniki pozwala osiągnąć proponowana metoda.

Losowe tworzenie tekstu przedstawia się czasem w postaci obrazu małpy uderzającej losowo w klawisze maszyny do pisania. Adaptacja tego modelu do ujednoznaczniania drzew składniowych polegałaby na posadzeniu owej małpy do interfejsu Dendrarium: klikałaby ona losowo w jedną z przedstawionych możliwości, aż do wybrania pełnego drzewa. Zgodnie z zasadami Dendrarium ujednoznacznienie przebiega od korzenia drzewa, więc kolejne decyzje wpływają na to, jakie wierzchołki niejednoznaczne znajdą się w dalszych poddrzewach.

Można zadać pytanie, jaki jest najgorszy możliwy wynik takiego ujednoznaczniania. Wymodelowano to poprzez modyfikację modelu małpy w ten sposób, że przy wyborze realizacji dla wierzchołka niejednoznacznego naj-

pierw pomija się odpowiedź poprawną (na podstawie ręcznych oznaczeń Składnicy), a następnie małpa losuje z pozostałych możliwości. W modelu tym, nazwanym modelem złośliwej małpy w Dendrarium, nadal jest element losowości, ponieważ decyzje dotyczące wierzchołków bliższych korzenia wpływają na paletę niejednoznaczności bliżej liści.

Wyniki zastosowania obu wariantów modelu małpy w Dendrarium do drzew ze Składnicy przedstawia tabela 7.1.

Tabela 7.1. Wyniki ujednoznaczniania drzew przez małpę w Dendrarium

	Składnica 0.5			Składnica 18.06		
	F_N	F_A	ULAS	F_N	F_A	ULAS
złośliwa małpa	0,826	0,530	0,593	0,836	0,663	0,759
małpa	0,845	0,599	0,655	0,854	0,711	0,784

Dwie przedstawione wersje korpusu różnią się wielkością, co nie ma wpływu na wyniki małpy, ale różnią się też znakowaniem, w szczególności zastosowaniem słownika Walenty w nowszej wersji. Składnica 18.06 okazuje się być łatwiejsza do ujednoznacznienia, szczególnie w zakresie ustalania wartości atrybutów wierzchołków.

Dokonując losowych wyborów zamiast celowo złych, można poprawić wynik o około 6 punktów procentowych dla F_A i tylko o niecałe 2 punkty procentowe dla F_N. Wyniki F_A są wyraźnie niższe niż odpowiadające F_N, co daje wgląd w różnicę trudności przypisania wartości wszystkim atrybutom w stosunku do ustalenia tylko nazw jednostek nieterminalnych. Różnica ta (dla obu wariantów modelu małpy) jest 1,7 raza większa dla Składnicy 0.5 niż dla wersji 18.06.

7.2. Modelowanie w stylu probabilistycznych gramatyk bezkontekstowych

Zasadnicza koncepcja probabilistycznych gramatyk bezkontekstowych polega na przypisaniu prawdopodobieństw użycia poszczególnym regułom używanej gramatyki bezkontekstowej[1]. Prawdopodobieństwa te należy rozumieć jako prawdopodobieństwa warunkowe użycia poszczególnych reguł pod warunkiem wyboru zadanej lewej strony reguły: $p(A \rightarrow \alpha_i|A)$, gdzie α_i dla $1 < i < k_A$ są możliwymi prawymi stronami reguł o lewej stronie A (Booth i Thompson 1973).

Analizator PCFG zaczyna rozkład wypowiedzenia od symbolu startowego gramatyki z prawdopodobieństwem 1 i wybiera spośród reguł rozpisujących

[1] Eksperymenty referowane w tym punkcie były tematem wspólnych prac autora i Dominiki Rogozińskiej (Woliński i Rogozińska 2013, 2016).

ten symbol. Skoro z symbolem startowym związane było prawdopodobieństwo 1, to prawdopodobieństwa warunkowe poszczególnych wyborów w pierwszym kroku są po prostu bezwarunkowymi prawdopodobieństwami wywodu złożonego z tego kroku. W następnych etapach wywodu, aby uzyskać jego prawdopodobieństwo, trzeba pomnożyć prawdopodobieństwo warunkowe danej reguły przez dotychczasowe prawdopodobieństwo wywodu. W ten sposób prawdopodobieństwo kompletnego drzewa składniowego t jest iloczynem prawdopodobieństw warunkowych reguł użytych do jego zbudowania:

$$p(t) = \prod_{A \to \alpha \in \mathrm{der}(t)} p(A \to \alpha | A)$$

gdzie $\mathrm{der}(t)$ jest wywodem drzewa t rozumianym jako wielozbiór użytych reguł gramatyki. Znalezienie najbardziej prawdopodobnego drzewa wymaga wybrania wywodu maksymalizującego iloczyn.

Prawdopodobieństwa reguł są w prosty sposób szacowane na podstawie korpusu drzew treningowych:

$$p(A \to \alpha | A) = \frac{\text{liczba wystąpień reguły } A \to \alpha \text{ w korpusie}}{\text{liczba wystąpień reguł o lewej stronie } A \text{ w korpusie}}$$

Zastosowanie tej koncepcji do ujednoznaczniania drzew generowanych przez analizator Świgra 2 wymaga adaptacji. Reguły rozważanej gramatyki mogą generować wierzchołki o różnej arności, a więc są jakby metaregułami gramatyki bezkontekstowej. Na potrzeby oceny prawdopodobieństwa drzewa różne realizacje reguły muszą być rozważane osobno. Obliczane jest więc prawdopodobieństwo warunkowe rozpisania danego symbolu nieterminalnego na dany ciąg składników.

Warto przy tym zauważyć, że w przyjętym zastosowaniu nie jest konieczne zapisanie gramatyki Świgry 2 jako faktycznej PCFG. Model ma być użyty jedynie do ujednoznaczniania istniejących lasów składniowych, więc używana „gramatyka" PCFG może być znacznie uproszczona, na przykład jednostki nieterminalne mogą być ograniczone do samych nazw kategorii składniowych. Nie musi ona uwzględniać wszystkich uwarunkowań odzwierciedlonych w regułach analizatora (gdyby faktycznie użyć jej do analizy, akceptowałaby ona za dużo struktur). Budowany model ma odzwierciedlać jedynie reguły wybierania najlepszego spośród istniejących drzew.

Eksperyment 1: jednostki nieterminalne

W pierwszym eksperymencie, aby zbadać zachowanie podstawowego schematu PCFG, modelowane reguły ograniczono do samych nazw jednostek nieterminalnych (kategorii składniowych), ignorując wszystkie atrybuty wierzchołków.

Na podstawie Składnicy zostały oszacowane prawdopodobieństwa tego, że dana jednostka nieterminalna rozpisuje się na dany ciąg składników, a więc na

przykład oszacowano prawdopodobieństwa tego, że **zdanie** jest rozpisywane na ciąg ⟨**fw**, **ff**⟩, albo ⟨**fw**, **ff**, **fw**⟩ itd.

Tabela 7.2. Wyniki modelu PCFG uwzględniającego tylko nazwy jednostek

	Składnica 0.5			Składnica 18.06		
	F_N	F_A	ULAS	F_N	F_A	ULAS
małpa	0,845	0,599	0,655	0,854	0,711	0,784
PCFG	0,917	0,721	0,890	0,933	0,856	0,902

Wyniki eksperymentu zaprezentowano w tabeli 7.2. Na danych Składnicy 18.06 to proste podejście poprawia około 54% błędów popełnianych przez małpę w Dendrarium, jeśli liczyć same przyporządkowania jednostek nieterminalnych, i 50% błędów, jeśli uwzględniać przewidywanie wartości atrybutów. Dla wersji 0.5 liczby te wynoszą odpowiednio 46% i 30%.

Wyniki tego eksperymentu pokazują, że w istocie zadanie ujednoznaczniania różni się istotnie od zadania analizy. Trudno sobie bowiem wyobrazić tak wysokie wyniki dla analizy bez uwzględnienia konieczności uzgodnień wartości liczby, rodzaju, osoby i przypadka w odpowiednich typach składników. Zachodzenie odpowiednich uzgodnień zapewnia jednak analizator Świgra 2, dlatego ujednoznacznienie odwołujące się do samych kategorii składniowych może dać stosunkowo dobre wyniki.

Eksperyment 2: uwzględnienie atrybutów

Model testowany w pierwszym eksperymencie opierał się jedynie na nazwach jednostek nieterminalnych, które dają zbyt mało informacji, żeby rozróżnić struktury zdań. Na przykład „reguła" ⟨**fw**, **ff**, **fw**⟩ odpowiada zarówno zdaniu złożonemu z podmiotu, orzeczenia i dopełnienia, jak i z dopełnienia, orzeczenia i podmiotu, a także dopełnienia, podmiotu i drugiego dopełnienia, nie różnicując ponadto typów owych dopełnień.

Naturalnym rozwinięciem jest uwzględnienie w modelu PCFG wybranych cech składników. W następnych eksperymentach uwzględniono atrybut typu frazy wymaganej *Tfw*, tak więc zamiast ciągu ⟨**fw**, **ff**, **fw**⟩ model „widzi" bardziej szczegółowe:

⟨**fw(**subj(np(nom))**)**, **ff**, **fw(**np(accgen)**)**⟩
⟨**fw(**subj(np(nom))**)**, **ff**, **fw(**np(dat)**)**⟩
⟨**fw(**subj(np(nom))**)**, **ff**, **fw(**cp(int)**)**⟩
⟨**fw(**np(accgen)**)**, **ff**, **fw(**subj(np(nom))**)**⟩
itd.

Wartość atrybutu *Tfw* daje sensowną informację o frazach wymaganych. Aby zróżnicować inne jednostki gramatyki, w wariantach eksperymentu dodano także wartości podstawowych cech fleksyjnych: rodzaju, liczby i osoby (w tabeli 7.3 pakiet ten oznaczono *RLO*). Cechy te są dziedziczone przez frazy od

ich centrów składniowych, więc dodanie tych atrybutów dostarcza informacji podobnej jak „leksykalizacja" rozważana w kontekście PCFG, czyli dodanie do etykiet fraz słów stanowiących centra tych fraz (Collins 1997).

Wartość przypadka wydaje się najbardziej istotna dla wymaganych fraz nominalnych, dla których jest sygnalizowana jako część atrybutu *Tfw*. Dlatego nie dołączono jej do pakietu *RLO* przy trenowaniu modelu uwzględniającego atrybut *Tfw*. Przeprowadzono jednakże osobny eksperyment, oznaczony *RLOP*, w którym stosowane są wszystkie cztery cechy fleksyjne, w tym przypadek.

Tabela 7.3. Wyniki modelu PCFG uwzględniającego wybrane atrybuty jednostek

	Składnica 0.5			Składnica 18.06		
	F_N	F_A	ULAS	F_N	F_A	ULAS
PCFG	0,917	0,721	0,890	0,933	0,856	0,902
PCFG+*Tfw*	0,924	0,764	**0,897**	0,931	0,853	**0,903**
PCFG+*RLO*	0,925	0,775	0,890	0,934	0,860	0,899
PCFG+*RLOP*	**0,927**	0,790	0,889	**0,935**	**0,863**	0,896
PCFG+*Tfw*+*RLO*	0,926	**0,807**	0,892	0,932	0,857	0,900

Jak można wyczytać z tabeli 7.3, uwzględnienie typów fraz wymaganych znacząco poprawiło wyniki dla Składnicy 0.5. Poprawy nie obserwuje się jednak w wynikach dla wersji 18.06 – model podstawowy osiągnął w tym wariancie wynik lepszy od wszystkich wzbogaconych modeli trenowanych na Składnicy 0.5. Dalsze uszczegółowienie modelu poprawia nieco wyniki, jednak zestaw *Tfw* + *RLO* osiąga wyniki gorsze niż modele prostsze.

Nasuwającym się wyjaśnieniem tych zjawisk jest to, że w wyniku dodawania cech dane stają się coraz bardziej rzadkie. Aby dane prawdopodobieństwo reguły PCFG „pasowało" do wierzchołka w ocenianych danych, zgadzać się muszą wszystkie uwzględniane w modelu cechy wszystkich składników bezpośrednich. Sytuacja, gdy dany zestaw cech nie wystąpił w danych treningowych, powoduje problem w ocenie prawdopodobieństwa, gdyż użycie prawdopodobieństwa 0 dawałoby 0 również jako wynikowe prawdopodobieństwo całego drzewa. Dlatego zastosowano prostą technikę „wygładzania" modelu: dla nieznanych kombinacji składników algorytm używa małej wartości ustalonej, która jest mniejsza od najmniejszego zaobserwowanego prawdopodobieństwa reguły. Heurystyka ta powoduje, że oceny generowane przez model przestają być prawdopodobieństwami warunkowymi (bo zsumują się do więcej niż 1), jednak wydaje się, że w wypadku modelu PCFG fakt ten nie powinien mieć niepożądanych skutków. Jeżeli wśród możliwych realizacji danego wierzchołka są takie, które wystąpiły w danych uczących, dostaną one wyższą ocenę niż te, które nie wystąpiły. Z kolei wymnażanie prawdopodobieństw dla poszczególnych węzłów będzie powodować preferowanie drzew zawierających jak najmniej kombinacji niewidzianych w danych uczących.

Tabela 7.4. Liczba wierzchołków niewystępujących w danych treningowych w zależności od wariantu modelu

	typów	wystąpień
PCFG	3 434	171 130
PCFG+*Tfw*	15 472	248 946
PCFG+*Tfw*+*RLO*	61 281	416 605

W tabeli 7.4 zestawiono liczbę wypadków, w których algorytm oceniający natrafia na kombinację wierzchołków niewystępującą w danych treningowych w zależności od wariantu modelu (zliczenia na Składnicy 0.5). Jak widać, wzbogacenie modelu powoduje szybki wzrost liczby niewidzianych obserwacji, co musi prowadzić do pogorszenia wyników. Aby temu zaradzić, trzeba by ostrożniej dobierać zestaw cech, uwzględniając je tylko tam, gdzie zysk z ich uwzględnienia przeważa nad stratą z rozrzedzenia danych.

Eksperyment 3: argumenty i modyfikatory

Jednym z elementów, które często są źródłem różnic w drzewach wybranych przez dwóch anotatorów, jest rozróżnienie między frazami wymaganymi i luźnymi w strukturze składniowej (por. p. 2.1.7). Można więc postawić pytanie, jak zmieniłaby się trudność zadania ujednoznaczniania analiz składniowych, gdyby zrezygnować z tego rozróżnienia.

Aby zbadać tę kwestię, drzewa składniowe Składnicy 0.5 przekształcono poprzez usunięcie ze struktury wierzchołków **fw** i **fl** i wprowadzenie zamiast nich ich jedynych potomków – fraz składnikowych odpowiednich typów.

Oto wyniki eksperymentu przeprowadzonego na danych Składnicy 0.5:

Tabela 7.5. Wyniki modelu PCFG na drzewach nie zawierających fraz wymaganych i luźnych

	F_N	F_A	ULAS
małpa	0,897	0,631	0,653
PCFG	0,960	0,922	0,890
PCFG+*RLOP*	0,943	0,925	0,859

Dla zmienionych drzew zmienia się punkt odniesienia, ponieważ małpa w Dendrarium losuje z innego zbioru możliwości. Zgodnie z oczekiwaniami wyniki małpy są wyższe, czyli zmienione zadanie jest istotnie prostsze.

Wynik zastosowania prostego modelu PCFG w tym wypadku jest lepszy o ok. 4 punkty procentowe F_N niż w wypadku oryginalnych drzew oraz o 20 pp. F_A. Prosty model PCFG bierze pod uwagę jedynie nazwy jednostek nieterminalnych, co sugeruje, że po usunięciu rozróżnienia fraz wymaganych i luźnych nazwy jednostek nieterminalnych w większym stopniu determinują wartości pozostałych atrybutów jednostek. Nie zmienia się miara ULAS dla modelu pracującego tylko na kategoriach składniowych, co nie jest zaskakujące, po-

nieważ rozróżnienie argumentów i modyfikatorów nie jest widoczne w strukturze drzew zależnościowych.

Ostatni wiersz tabeli 7.5 pokazuje wynik eksperymentu, w którym nazwy jednostek zostały wzbogacone o podstawowe cechy fleksyjne *RLOP*. Ponieważ w przekształconych drzewach nie ma atrybutu *Tfω*, w eksperymencie jako cechy wierzchołków wykorzystano rodzaj *R*, liczbę *L*, osobę *O* i przypadek *P*. Dodanie atrybutów pozwoliło odrobinę podwyższyć F_A (jest to najlepszy wynik dla przedstawionych eksperymentów z PCFG), ale jednocześnie zmalały miary F_N i ULAS. Zapewne znowu dochodzi do głosu problem rzadkości danych.

Przedstawiony tu eksperyment pokazuje, że zaniedbanie rozróżnienia atrybutów i okoliczników faktycznie znacząco poprawia wyniki ujednoznacznienia. Ponieważ rozróżnienie to jest słabo ugruntowane lingwistycznie, choć wszechobecne w teoriach i słownikach, może faktycznie można je zaniedbywać przy przetwarzaniu automatycznym poziomu składniowego.

W ramach badania błędów popełnianych przez algorytm poddano oglądowi wierzchołki z nieterminalem **zdanie**. Tabela 7.6 pokazuje, jak często algorytm wybierał podziały zdań na zbyt mało składników względem interpretacji wybranej przez anotatorów, a jak często – na zbyt wiele. Algorytm ma

Tabela 7.6. Błędy popełniane przez algorytm PCFG w podziale zdania na składniki

	zbyt mało składników	zbyt wiele składników
PCFG+*Tfω*	4,2%	15,0%
PCFG bez **fw**/**fl**	2,1%	26,3%

wyraźną tendencję do wybierania podziałów zbyt drobnych, która wzmaga się w wariancie zadania pomijającym argumenty i modyfikatory. Nasuwa się więc myśl, żeby w jakiś sposób wprowadzić do modelu czynnik preferujący interpretacje krótsze.

Zalety i wady metody

Zasadniczą zaletą algorytmu PCFG jest jego prostota. Co ciekawe, otrzymane za jego pomocą wyniki są zbliżone dla miary ULAS do tych uzyskiwanych przez statystyczne parsery zależnościowe. Wróblewska i Woliński (2012) podają jako najlepszy wynik 0,922 ULAS dla parserów trenowanych na wersji 0.5 Składnicy. Warto przy tym podkreślić, że model PCFG wybiera drzewo spośród zaakceptowanych przez parser regułowy, a więc zgodnych z jego regułami. Zabezpiecza to przed pojawieniem się w wynikach na przykład struktur z dwoma podmiotami, co zdarza się stochastycznym parserom zależnościowym. Jednak parser zależnościowy generuje jakąś odpowiedź dla każdego zdania, a parser regułowy dla części zdań po prostu zawodzi.

Wykorzystanie modelu PCFG jest możliwe dzięki temu, że najbardziej prawdopodobne drzewo da się wyznaczyć na podstawie upakowanego lasu

składniowego bez jego rozpakowywania na poszczególne drzewa. Dzieje się tak, ponieważ prawdopodobieństwo drzewa jest po prostu iloczynem warunkowych prawdopodobieństw reguł użytych do jego wygenerowania. Największe prawdopodobieństwo dla danego wierzchołka dadzą największe czynniki uzyskane dla jego składników. Stosując proste programowanie dynamiczne tablicujące iloczyny prawdopodobieństw dla poddrzew, można wyznaczyć najbardziej prawdopodobne drzewo (Woliński i Rogozińska 2016). Jak powiedziano wcześniej (p. 6.3), liczba drzew w upakowanym lesie może być wykładnicza względem długości zdania, uniknięcie rozpakowania lasu jest więc zasadniczym warunkiem stosowalności algorytmu.

Wyraźną wadą modelu PCFG jest to, że zakłada się w nim probabilistyczną niezależność zastosowania poszczególnych reguł w obrębie wywodu zdania (co pozwala wyrazić prawdopodobieństwo wywodu jako iloczyn prawdopodobieństw reguł). Intuicja podpowiada, że w rzeczywistości między regułami zachodzą zależności i interakcje.

Siła wyrazu modelu PCFG jest też ograniczona przez to, że do konkretnego węzła w drzewie może się odnosić tylko jedna ocena modelu. Nie da się więc skomponować oceny węzła z osobnych składowych (np. wystąpienie danego typu potomka przy danym nadrzędniku; współwystąpienie danych podrzędników w danej jednostce; liczba wszystkich składników lub składników danego typu itd.). Również selekcja cech uwzględnianych w modelu musiałaby dokonać się niejako globalnie: w najlepszym razie dla każdego typu wierzchołka trzeba by ustalić jego uwzględniane cechy. Ograniczenia te sprawiają, że przy wzbogacaniu modelu zliczane obserwacje szybko robią się bardzo rzadkie. Nasuwa się więc pytanie, czy można polepszyć wyniki za pomocą bardziej wyrafinowanego modelu statystycznego.

7.3. Modelowanie maksimum entropii

Model PCFG przypisuje prawdopodobieństwa bezpośrednio regułom gramatyki, jest więc bardzo silnie związany z naturą zadania składnikowej analizy składniowej. W tym podrozdziale przedstawiona zostanie metoda oparta na dużo ogólniejszej zasadzie. Zakładana sytuacja jest następująca (por. Berger *et al.* 1996). Celem jest skonstruowanie modelu statystycznego, który będzie przewidywał zachowanie procesu losowego. Dana jest skończona próba zachowania owego procesu. Aby umożliwić ocenę, na ile model statystyczny dobrze odwzorowuje proces losowy, wprowadza się listę pewnych cech, których wystąpienia można zliczać w danych. Następnie poszukuje się modelu, który będzie generował takie same zliczenia cech jak stwierdzone w danych treningowych.

W wypadku ujednoznaczniania drzew składniowych modelowany proces losowy przypisuje drzewom prawdopodobieństwa warunkowe. Elementem

warunkującym prawdopodobieństwo drzewa t jest wygenerowany przez parser las h, z którego jest ono wyjmowane. Dla danego lasu idealny model przypisywałby prawdopodobieństwo 1 drzewu poprawnemu, a wszystkim pozostałym drzewom – prawdopodobieństwo 0. Próbę zachowania procesu stanowią oznaczone w korpusie składniowym drzewa poprawne na tle wszystkich drzew wygenerowanych przez analizator regułowy.

Formalne ujęcie wymaga skojarzenia z każdą cechą i funkcji f_i, $1 \leqslant i \leqslant K$, gdzie K jest liczbą zdefiniowanych cech[2]:

$$f_i(t,h) = \begin{cases} 1 & \text{jeżeli } i\text{-ta cecha przysługuje drzewu } t \text{ w lesie } h \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Poszukiwany jest model M prawdopodobieństwa warunkowego drzewa t, gdy dany jest las h, o postaci:

$$p_M(t|h) = \frac{\prod\limits_{1 \leqslant i \leqslant K} \alpha_i^{f_i(t,h)}}{\sum\limits_{t' \in \tau(h)} \prod\limits_{1 \leqslant i \leqslant K} \alpha_i^{f_i(t',h)}}$$

gdzie α_i są parametrami modelu, które trzeba wyznaczyć w procesie uczenia, a $\tau(h)$ jest zbiorem drzew, które można wypakować z lasu h. Obecność normalizującego wyrażenia w mianowniku zapewnia, że wartości $p_M(t|h)$ spełniają definicję prawdopodobieństwa. Jeżeli potrzebny jest jedynie względny ranking drzew, porównanie można oprzeć na wartościach wyrażenia w liczniku.

Wartość oczekiwaną i-tej cechy ze względu na rozkład empiryczny $\tilde{p}(t,h)$ (a więc rozkład oparty na częstościach w korpusie treningowym C) można określić jako

$$\tilde{p}(f_i) = \sum_{t,h \in C} \tilde{p}(t,h) f_i(t,h)$$

gdzie C jest korpusem składniowym, rozumianym jako zbiór poprawnych drzew t oznaczonych w lasach h.

Dopasowanie modelu do danych treningowych dokonuje się poprzez nałożenie warunków, aby zliczenia i-tej cechy przewidziane przez model były zgodne z empiryczną wartością oczekiwaną:

$$\sum_{t,h \in C} \tilde{p}(h) p_M(t|h) f_i(t,h) = \tilde{p}(f_i) \qquad (\star)$$

gdzie $\tilde{p}(h)$ jest rozkładem prawdopodobieństwa poszczególnych lasów w danych treningowych, zwykle zakłada się, że jest to rozkład jednorodny.

Zgodność modelu p_M z korpusem wyraża się jako warunek, aby równanie $(\star)$ zachodziło dla wszystkich i, $1 \leqslant i \leqslant K$. Modelowanie maksimum

[2] W ogólnym przypadku funkcje cech nie muszą być binarne, mogą przypisywać obserwacjom jakieś oceny rzeczywiste. W przedstawionych eksperymentach były jednak stosowane funkcje binarne.

entropii polega na wyborze spośród modeli spełniających ten zbiór ograniczeń tego, który charakteryzuje się największą wartością entropii

$$H(t|h) = -\sum_{t,h \in C} \tilde{p}(h) p_M(t|h) \log p_M(t|h)$$

Można argumentować (por. Berger *et al.* 1996), że odpowiada to wyborowi modelu, który jest najbardziej jednorodny, a przez to najmniej zakłada o niewiadomych czynnikach wpływających na obserwowany rozkład. Model ten jednocześnie maksymalizuje wiarogodność korpusu treningowego.

Modelowanie maksimum entropii można było zastosować do problemu ujednoznaczniania drzew składniowych dzięki temu, że optymalizacja modelu jest możliwa do wykonania na upakowanych lasach drzew. Szczegóły algorytmu obliczania współczynników modelu w takiej sytuacji przedstawili Miyao i Tsujii (2002, 2008). W przedstawionych w tym podrozdziale eksperymentach do optymalizacji parametrów modelu wykorzystano gotowe narzędzie o nazwie Amis, które opracował Yoshida (2006)[3].

Eksperyment 1: Jednostki nieterminalne

Pierwszy eksperyment z modelowaniem maksimum entropii miał na celu zbadanie, czy bardziej wyrafinowana metoda statystyczna (w szczególności niezakładająca niezależności obserwacji) da lepsze wyniki niż PCFG na podstawie tych samych informacji wejściowych.

Zdefiniowano więc cechy dla modelu odpowiadające dokładnie „regułom" podstawowej wersji modelu PCFG, a więc obserwowanymi cechami drzew składniowych było pojawianie się danej jednostki nieterminalnej rozpisanej na dany ciąg jednostek składniowych (tylko nazwy jednostek).

Tabela 7.7. Porównanie wyników podstawowych wersji dwóch metod modelowania

	F_N	F_A	ULAS
PCFG	0,933	0,856	0,902
MaxEnt	0,947	0,879	0,924

Efekty dla Składnicy 18.06 w dziesięciokrotnej walidacji krzyżowej przedstawiono w tabeli 7.7. Miara F_N wzrosła o 1,4 pp., co odpowiada poprawieniu przez model maksimum entropii 21% błędów popełnianych przez model PCFG. W wypadku miary F_A jest to polepszenie o 2,3 pp., a więc poprawa 16% popełnianych uprzednio błędów. Wszystkie trzy oceny w tym eksperymencie są wyższe od not uzyskiwanych przez zbadane w poprzednim punkcie warianty algorytmu PCFG. Zastosowanie modelowania maksimum entropii przyniosło więc wyraźną poprawę wyników już przez zmianę typu stosowanego modelu statystycznego.

[3] Pierwsze eksperymenty wykonała Dominika Rogozińska w ramach pracy magisterskiej (Rogozińska 2016), przygotowanej pod kierunkiem autora.

Eksperyment 2: Wzbogacenie atrybutów

W następnym eksperymencie poprzedni zestaw obserwacji (oznaczony con od *constituents*) porównano z wynikiem uzyskanym dzięki wzbogaceniu informacji o składnikach o typ fraz wymaganych (contfw). Porównanie to przedstawiają dwa pierwsze wiersze tabeli 7.8. Wyniki modelu wzbogaconego są nieco lepsze od podstawowego, jednak ciekawsze jest to, że oba zestawy cech da się połączyć (dzięki temu, że można wiązać wiele obserwacji z jednym wierzchołkiem drzewa). Model z oboma zestawami cech uzyskuje jeszcze wyższe noty – por. trzeci wiersz, con+contfw.

Efekt ten jest związany z interesującą cechą modelu maksimum entropii polegającą na tym, że proces trenowania modelu wykonuje automatyczną selekcję cech: pewnym cechom przypisywana jest waga 1, co oznacza, że zaobserwowanie ani niezaobserwowanie tej cechy w drzewie nie wpływa na jego ocenę.

Oznacza to również, że modelowanie maksimum entropii jest odporne na rozrzedzanie danych w tym sensie, że gdy zastosowano jednocześnie cechy bardziej szczegółowe (contfw) i mniej szczegółowe (con), algorytm był w stanie skorzystać z obu rodzajów.

Inaczej są też traktowane dane testowe, które „nie przypominają" danych treningowych. W modelu PCFG pojawienie się kombinacji składników, która nie pojawiła się w danych treningowych, powodowało przypisanie jej małego prawdopodobieństwa wygładzającego, co z kolei prowadziło do potraktowania drzewa zawierającego taki węzeł jako bardzo mało prawdopodobnego. Analogiczna sytuacja w modelu maksimum entropii oznacza, że do danego węzła drzewa nie stosuje się żadna cecha, a więc że węzeł taki nie wpływa na ocenę całości drzewa. Przy odpowiednim nasyceniu cechami ocenianych drzew jest to podejście bardzo intuicyjne. Odpowiada ono stwierdzeniu niemożności oceny danego węzła drzewa i wykonaniu oceny na podstawie węzłów, o których daje się coś powiedzieć w oparciu o wcześniej widziane dane treningowe.

Zdefiniowanie zbyt dużego zbioru cech dla modelu maksimum entropii spowolni proces trenowania modelu, jednak nie spowoduje negatywnych skutków rozrzedzenia danych obserwowanych przy PCFG. W tabeli 7.8 widać, że wyniki dla danego zbioru cech są zawsze nie gorsze od uzyskiwanych dla podzbiorów tego zbioru cech, co w szczególności oznacza, że tworzone modele nie wykazują oznak przeuczenia.

Eksperyment 3: Zmiana zasady budowania cech

Ponieważ w modelowaniu maksimum entropii można z jednym węzłem drzewa powiązać wiele cech, nasuwa się koncepcja odejścia od cech wzorowanych na PCFG, które charakteryzowały wierzchołek z uwzględnieniem własności wszystkich składników. Zamiast tego można z każdym wierzchołkiem

Tabela 7.8. Warianty zestawów obserwowanych cech i odpowiadające im wyniki modelu maksymalizacji entropii (Składnica 18.06; 10-krotna walidacja krzyżowa)

	F_N	F_A	ULAS	liczba cech	z wagą 1
con	0,947	0,879	0,924	12643	26,3%
contfw	0,950	0,894	0,926	67590	17,0%
con+contfw	0,953	0,899	0,934	80233	22,8%
dep	0,945	0,884	0,915	1090	4,3%
dep+headlen	0,950	0,895	0,928	1645	5,8%
dep+2gram	0,951	0,892	0,930	3675	2,8%
dep+head+headlen	0,951	0,896	0,928	2070	5,9%
dep+head+headlen +2gram+svo+rhslen	0,956	0,905	0,937	4803	4,4%
dep+head+headlen+2gram +svo+rhslen+con+contfw	0,956	0,906	0,937	85036	31,3%

związać wiele cech, charakteryzujących różne aspekty opisu tego wierzchołka. W następnym eksperymencie podjęto próbę wybrania cech, które miałyby większą gęstość w danych treningowych, a przez to stanowiły lepsze źródło informacji statystycznych o opisywanych danych.

Oto pełna lista typów cech stosowanych w eksperymentach zestawionych w tabeli 7.8. Przy każdym typie podano przykłady cech danej klasy opisujących przedstawiony na rysunku 7.1 wierzchołek **zdanie** z centrum *powiedział* (każda nazwa cechy niesie na początku czteroliterowe oznaczenie jej przynależności do typu, właściwa charakterystyka znajduje się po znaku podkreślenia):

con jednostka nieterminalna i ciąg jej składników bezpośrednich (same nazwy jednostek); zaznaczane jest centrum składniowe (symbol @):

fcon_zdanie→fw.fl.@ff.fw.fw

contfw jak con, ale dla fraz wymaganych podawany również typ frazy wymaganej:

ftfw_zdanie→fw(subj(np(nom))).fl.ff.fw(np(dat)).fw(cp(że))

head jednostka nieterminalna i jej centrum składniowe:

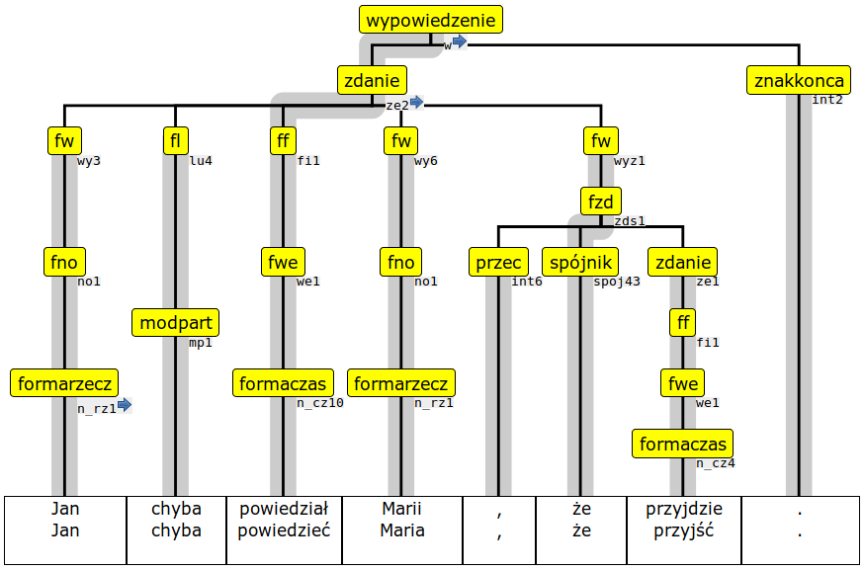
fhd_zdanie→ff

headlen jak head z dodaniem liczby składników bezpośrednich jednostki:

fhd2_zdanie→ff=5

dep trójka złożona z jednostki, jej centrum składniowego oraz jednego z pozostałych składników z informacją, czy wystąpił on przed czy za centrum składniowym; notowane są nazwy jednostek nieterminalnych i typy fraz wymaganych:

fdep_zdanie→fw(subj(np(nom)))..@ff
fdep_zdanie→fl..@ff
fdep_zdanie→@ff..fw(np(dat))
fdep_zdanie→@ff..fw(cp(że))



Rysunek 7.1. Drzewo dla zdania *Jan chyba powiedział Marii, że przyjdzie.*

2gram bigram jednostek występujących jako składniki bezpośrednie jakiejś jednostki; jeżeli jedna z nich jest centrum składniowym jest oznaczana; notowane są nazwy jednostek nieterminalnych i typy fraz wymaganych:

f2gt_fw(subj(np(nom))).fl
f2gt_fl.@ff
f2gt_@ff.fw(np(dat))
f2gt_fw(np(dat)).fw(cp(że))

svo układ składników zdania kodowany analogicznie do badania pokazanego w punkcie 6.9: S – fraza wymagana podmiotowa, V – fraza werbalna, O – inna fraza wymagana; pozostałe składniki są pomijane; powtórzenia symboli są kompresowane do jednego:

fsvo_SVO

rhslen obserwacją jest etykieta wierzchołka i liczba składników, na które jest on rozpisywany:

flen_zdanie=5

Jak widać z tabeli 7.8, można skonstruować zestawy cech nieodwołujące się do pełnych zestawów składników bezpośrednich, które dają lepszą jakość ujednoznacznienia. Wydaje się więc, że gramatyka Świgra 2 zawiera w sobie wystarczającą informację o możliwych pełnych ciągach składników, moduł ujednoznaczniający zaś może korzystać z informacji bardziej wybiórczej, dotyczącej kombinacji nadrzędników z podrzędnikami i współwystąpień podrzędników. Już sam zestaw cech dep, które przypominają informację zawartą

w krawędziach drzewa zależnościowego, daje porównywalne wyniki jak zestaw con. Dodanie informacji o centrach składniowych (zestaw dep+head+headlen) pozwala pobić pod względem wszystkich trzech miar zarówno zestaw con, jak i contfw (używane osobno). Ciekawe jest jednak to, że zestaw dep+head+headlen odpowiada jedynie 2070 cechom zaobserwowanym w danych Składnicy, podczas gdy gorszy od niego zestaw contfw liczy 67 590 cech (przedostatnia kolumna w tabeli 7.8). W ostatniej kolumnie tabeli podano procent cech w danym eksperymencie, którym algorytm konstruujący model przypisał wagę 1, a więc uznał je za praktycznie nieużyteczne do oceny drzew (zliczono wagi dokładnie równe 1, co wystarcza do uzyskania danych poglądowych; gdyby chcieć odfiltrować nieistotne cechy, zapewne należałoby wprowadzić jakiś niewielki przedział wokół wartości 1). Jak widać, „lekkie" cechy zaproponowane w tym podpunkcie są w dużo większym procencie użyteczne dla modelu, można przypuszczać, że odpowiedzialna za to jest ich większa gęstość w danych treningowych.

Generowanie cech opisujących dany las składniowy jest główną częścią czasu pracy algorytmu znajdującego najlepsze drzewo. Modele operujące mniejszą liczbą cech są więc korzystniejsze, ponieważ działają one zauważalnie szybciej[4].

Czego model się nauczył

Bardzo ciekawą cechą modelu maksymalizacji entropii jest to, że przynajmniej niektóre z wag przypisanych cechom mają bardzo intuicyjną interpretację. Tabela 7.9 pokazuje fragment listy cech z zestawu contfw i przypisanych im wag posortowanej według malejącej wagi.

Najwyższą wagę (zob. pierwszy wiersz tabeli) otrzymała obserwacja, w której fraza wymagana podmiotowa jest realizowana jako fraza nominalna. Ponieważ wagi nie są w tym modelu prawdopodobieństwami warunkowymi pod warunkiem lewej strony reguły, waga ta nie mówi, „jeżeli wystąpił podmiot, to powinien on być realizowany przez frazę nominalną". Wysoka waga mówi tylko, że pożądane jest pojawienie się konfiguracji węzłów opisanej tą cechą. Można ją więc interpretować jako: „dobrze jest, gdy w zdaniu jest podmiot", względnie: „preferuj drzewa zawierające podmiot względem niezawierających go".

Informacja o preferencji dla obecności podmiotu pojawia się również w innych modelach. Na przykład w zestawie cech dep obecność podmiotu promują następujące dwie:

5	fdep_zdanie→fw(subj(np(nom)))..@ff	8,91866
17	fdep_zdanie→@ff..fw(subj(np(nom)))	2,72852

[4] Między modelami opisanymi w ostatnich dwóch wierszach tabeli 7.8 różnica w czasie działania jest trzydziestokrotna na korzyść modelu mniejszego. Należy jednak zastrzec, że program nie był optymalizowany pod kątem szybkości działania.

Tabela 7.9. Wagi cech obliczone przez algorytm maksymalizacji entropii dla modelu contfw

	cecha	α_i
1	ftfw_fw(subj(np(nom)))→fno	16,50689
2	ftfw_fno→flicz.fno	6,96087
3	ftfw_fno→fpt.fno.fno	4,80178
4	ftfw_fw(np(accgen))→fno	4,50726
5	ftfw_fw(prepnp(o.loc))→fpm	3,98303
6	ftfw_ff→fwe	3,80818
7	ftfw_fw(prepnp(do.gen))→fpm	3,55386
8	ftfw_fw(prepnp(na.acc))→fpm	3,19388
...	...	...
3128	ftfw_fno→fno.fw(xp(caus)).fno.fno.fno.fw(np(gen)).fpm	1,00000
3129	ftfw_fwe→fl.fw(xp(adl)).fl.fl.fl.fw(subj(np(nom))).fwe	1,00000
3130	ftfw_zdanie→fl.fl.fw(subj(np(nom))).fw(xp(perl)).fl.fw(np(accgen))	1,00000
...	...	...
34367	ftfw_fno→fno.fpt.fno.fpt.fpm.fpm.fno	1,00000
34368	ftfw_zdanie→fw(subj(np(nom))).fw(xp(mod)).ff.fw(xp(adl)).fl.fl.fl	1,00000
34369	ftfw_zdanie→fw(xp(adl)).fl.ff.fw(prepnp(z.inst)).fl.fl	1,00000
...	...	...
67586	ftfw_fno→fpm.fno	0,13908
67587	ftfw_fno→flicz	0,13024
67588	ftfw_fl→fno	0,11157
67589	ftfw_fl→fpt	0,01967
67590	ftfw_fno→fpt	0,01293

Pierwsza z nich opisuje sytuację, gdy zdanie ma jako centrum frazę finitywną **ff**, a podmiot nominalny poprzedza to centrum. Opisana w drugiej cesze sytuacja, gdy podmiot stoi za centrum finitywnym otrzymała niższą notę, ale również wyraźnie promującą taką sytuację. (W tym zestawie cech na pierwszą pozycję wysunęła się cecha opisująca obecność podrzędnika przymiotnikowego przed nominalnym centrum frazy nominalnej: fdep_fno→fpt..@fno z wagą 41,64737).

Wiersz 4 w tabeli 7.9 wyraża analogiczną obserwację dla najbardziej typowego dopełnienia w języku polskim, a więc dla frazy nominalnej w bierniku. Siła preferencji dla takiego argumentu nie jest jednak tak wielka jak dla obserwacji dotyczącej podmiotu. Wiersz drugi tabeli opisuje typowy preferowany skład frazy liczebnikowo-nominalnej: dwa składniki, z których pierwszym jest składnik liczebnikowy, a drugim nominalny. Wiersze 5, 7 i 8 pokazują typowe wymagane frazy przyimkowo-nominalne. Jak się okazuje, algorytm ustalił, że ich obecność jest przesłanką do wysokiej oceny danego drzewa.

Od wiersza 3128 do 34 369 ciągnie się blok obserwacji, którym przypisano wagę 1. Dominują tu skomplikowane ciągi składników, które okazały się nie wnosić nic do procesu oceny drzewa.

Ostatni wiersz tabeli mówi o możliwości nietypowej realizacji frazy nominalnej przez frazę przymiotnikową, przypisując jej najniższą możliwą wagę.

Ta cecha jest również bardzo intuicyjna: taka realizacja frazy nominalnej jest „realizacją ostatniej szansy", preferowane powinny być drzewa, w których frazy nominalne mają centra nominalne.

W poprzednich wierszach można także znaleźć nominalne i przymiotnikowe realizacje fraz luźnych (co zapewne oznacza, że frazy takie są przez model preferowane jako wymagane). W wierszu 67 587 pojawia się „kara" za realizację frazy nominalnej przez samotną frazę liczebnikową, a w poprzednim – informacja, że czymś nietypowym jest podrzędnik przyimkowy we frazie nominalnej poprzedzający jej centrum nominalne. Ta ostatnia obserwacja jest również bardzo intuicyjna: fraza *na wzgórzu dom* jest nietypowa w porównaniu z *dom na wzgórzu*. Oczywiście dla większości cech ze środkowej części listy trudno byłoby podać równie przejrzyste interpretacje. Ponadto jakość ocen dokonywanych przez model kryje się w odpowiednim dopasowaniu do siebie całego zestawu wag.

7.4. Podsumowanie

W rozdziale tym pokazano, że problem ujednoznaczniania składnikowych drzew składniowych poddaje się rozwiązaniu algorytmicznemu. Referowane tu eksperymenty stanowią pierwszą próbę w zakresie statystycznego ujednoznaczniania składnikowych drzew składniowych dla języka polskiego.

Wyniki eksperymentów wskazują wyraźnie na przewagę metody maksymalizacji entropii nad prostszym modelem probabilistycznych gramatyk bezkontekstowych. Cenna dla zastosowań praktycznych jest także możliwość użycia małych i przez to szybkich modeli opartych na cechach modelujących zależności między nadrzędnikami a podrzędnikami. Proste cechy pozwalają wzbogacać model statystyczny bez jego rozgęszczenia i bez przeuczenia.

Istotnym aspektem przedstawionych eksperymentów jest to, że analiza składniowa została przeprowadzona na ujednoznacznionych danych fleksyjnych. Potrzebne więc są dalsze eksperymenty, w których zostanie zbadana interakcja ujednoznacznienia fleksyjnego z ujednoznacznieniem struktur składniowych.

Nie można też powiedzieć, że zostały wyczerpane możliwości eksploracji możliwych cech. Zbadania wymagałoby, które z atrybutów jednostek nieterminalnych mogą wnieść informację użyteczną dla ujednoznacznienia składniowego. Można by również spróbować użyć takich informacji jak długości fraz (w segmentach), wysokości poddrzew lub głębokości wierzchołków w drzewie. W dotychczasowych eksperymentach nie wykorzystano też możliwości budowania cech pokazujących dane drzewo na tle wygenerowanego przez analizator lasu. Mogłyby to być cechy sygnalizujące że dla danej konstrukcji istnieje lub nie istnieje pewna interpretacja alternatywna, a więc na przykład że pewna fraza luźna może być też interpretowana jako wymagana.

Zakończenie

W pracy przedstawione zostały spójne ze sobą powierzchniowe opisy fleksji i składni polskiej. Pierwszy zawdzięcza swoje istnienie pracom Saloniego, których zwieńczeniem jest *Słownik gramatyczny języka polskiego*. Danych tego słownika używa przedstawiony w pracy analizator fleksyjny Morfeusz SGJP. Ciekawym jego aspektem jest reprezentacja analiz fleksyjnych w postaci acyklicznych grafów interpretacji.

Przedstawiony opis składni wywodzi się z gramatyki Świdzińskiego (GFJP), prezentuje on jednak istotnie nowy etap rozwoju. Przede wszystkim znacznie ograniczony został repertuar jednostek nieterminalnych wyróżnianych przez gramatykę. W związku z tym nowy opis operuje dużo prostszymi, czytelniejszymi i bardziej intuicyjnymi drzewami rozbioru. Opis składniowy rozbudowano o wiele typów konstrukcji, w szczególności skoordynowanych fraz różnych typów. Uwzględniono pewne typy konstrukcji niezdaniowych i nieciągłych. Szczegółowa lista owych rozszerzeń jest zawarta w punkcie 2.16.

Proponowana postać drzew składniowych łączy prostotę drzew bliską drzewom zależnościowym (dzięki konsekwentnemu odejściu od strukturyzacji binarnych) z możliwością naturalnego reprezentowania konstrukcji współrzędnych i leksykalnych jednostek wieloczłonowych, które są piętą achillesową opisów zależnościowych. Informacja jest wystarczająco bogata, żeby zapewnić konwersję na drzewa zależnościowe, co oznacza użyteczność wyników również dla badaczy posługujących się innym formalizmem (por. Nivre 2003).

Bardzo istotnym aspektem nowej gramatyki jest przeniesienie obiektu zainteresowania z wypisywania reguł, pozwalających zaakceptować dany zbiór zdań polskich, na myślenie o kształtach struktur, które powinny dane zdania reprezentować. Ich adekwatność jest bowiem kluczowa z punktu widzenia językoznawczego. Struktury takie są również danymi dla kolejnych etapów przetwarzania, np. tworzenia reprezentacji semantycznej. W przedstawionej analizie składniowej wykorzystywany jest najbardziej wyrafinowany z dostępnych słowników walencyjnych – Walenty (rozdział 3).

Zaprezentowany opis ma charakter regułowy, co oznacza, że korzystający z niego analizator składniowy generuje dla danego wypowiedzenia wszystkie możliwe interpretacje składniowe. Ze względu na niejednoznaczność języka naturalnego oznacza to, że dla praktycznie każdego wypowiedzenia generowany jest las składniowy zawierający więcej niż jedno drzewo. Elementem

bardzo ważnym ze względu na możliwość praktycznego wykorzystania analizatora Świgra 2 jest dostępność modułu ujednoznaczniania statystycznego drzew analizy (rozdział 7).

Adekwatność przedstawionego opisu potwierdza korpus składniowy Składnica (rozdział 6). Co prawda analiza korpusu nie może potwierdzić pełności opisu, jednak jej istotna rola polega na tym, że może ujawnić jego luki. Składnica jest pierwszym korpusem składniowym dla języka polskiego, w którym pełne drzewa składniowe zostały wygenerowane automatycznie w zgodzie z gramatyką formalną, a następnie zweryfikowane przez ekspertów. W obecnej wersji korpusu odsetek wypowiedzeń, dla których udało się to zrobić, wynosi około 60%. Liczba ta może nie wydawać się bardzo duża, jednak Składnica przewyższa poprzednie próby budowy korpusu składniowego wielkością. Jest przez to pierwszym korpusem składniowym, który może być przydatny w badaniach kwantytatywnych.

W ramach referowanych prac zostało opracowane kompletne środowisko służące do ujednoznaczniania i weryfikacji składnikowych drzew składniowych (system Dendrarium). Bardzo ważnym jego elementem jest algorytm aktualizujący zatwierdzone przez ekspertów drzewa składniowe, pozwalający utrzymać zgodność drzew z ewoluującą gramatyką formalną. Opracowano także wyszukiwarkę korpusową pozwalającą na znalezienie struktur składniowych na podstawie cech wierzchołków drzew składniowych i relacji między wierzchołkami. Jej interesującym elementem jest możliwość odwołania się w zapytaniu zarówno do drzewa wybranego przez ekspertów, jak i do wszystkich alternatyw zaproponowanych przez analizator regułowy.

Efektywność przetwarzania struktur składniowych jest zapewniona przez to, że wszystkie narzędzia pracują na upakowanych lasach składniowych. Dotyczy to analizatora Świgra 2, systemu Dendrarium, statystycznych algorytmów ujednoznaczniających i wreszcie wyszukiwarki drzew.

Wszystkie elementy opisu mają postać działających, publicznie dostępnych narzędzi, które pozwalają na samodzielne eksperymenty. Dotyczy to analizatora Morfeusz SGJP (http://sgjp.pl/morfeusz/demo), analizatora składniowego Świgra 2 (http://swigra.nlp.ipipan.waw.pl/) i wyszukiwarki drzew składniowych w korpusie Składnica (http://treebank.nlp.ipipan.waw.pl/).

Kierunki rozwoju

Dostępność przedstawionych zasobów otwiera możliwości dalszych badań. Opis można rozbudowywać zarówno wszerz, przez rozbudowę reguł gramatyki, aby objąć więcej konstrukcji składniowych, jak i wgłąb – na przykład poprzez wysubtelnienie wykorzystania informacji ze słownika Walenty (informacji o kontroli składniowej czy pełniejsze wykorzystanie opisów konstrukcji frazeologicznych) lub przejście do reprezentacji poziomu semantycznego.

Osobnym zadaniem mogłoby być przystosowanie narzędzi do przetwarzania tekstów niestarannych, „zaszumionych", co pozwoliłoby analizować nieprzebrane zasoby tekstów obecnych w internecie.

Opisane prace były prowadzone na wstępnie zanalizowanym fleksyjnie korpusie NKJP1M. Zbadania wymaga więc optymalna strategia postępowania w wypadku przetwarzania tekstu bez takich informacji. Można wyobrazić sobie analizę składniową na nieujednoznacznionych wynikach analizy fleksyjnej (generowanych przez Morfeusza, tak działa webowa wersja Świgry 2) albo wplecenie w łańcuch przetwarzania narzędzia ujednoznaczniającego opis fleksyjny – tagera. Pierwsza droga ma wadę w postaci złożoności przetwarzania składniowego zwiększonej przez mnogość interpretacji fleksyjnych. W wypadku drugiej błędna decyzja tagera eliminuje możliwość uzyskania właściwego rozbioru składniowego. Ciekawym kierunkiem byłoby zastosowanie tagera „ostrożnego", który ujednoznaczniałby tylko częściowo dla zmniejszenia liczby wariantów poddawanych analizie składniowej, ale ograniczał się do decyzji, które są w miarę pewne.

Ciekawym wykorzystaniem danych Składnicy (w wariancie Składnicy zależnościowej) jest trenowanie analizatorów zależnościowych. Interesująca byłaby również próba wytrenowania stochastycznego analizatora składnikowego. Budowa takiego efektywnego analizatora zwiększyłaby możliwości wykorzystania przedstawionego opisu w zastosowaniach praktycznych.

Bibliografia

Abney *et al.* (1991): Steven Abney, Dan Flickenger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Frederick Jelinek, Judith Klavans, Mark Liberman, Mitch Marcus, Salim Roukos, Beatrice Santorini i Tomek Strzalkowski, *Procedure for quantitatively comparing the syntactic coverage of English grammars*, w: *Proceedings of the workshop on Speech and Natural Language*, red. E. Black, HLT '91, Association for Computational Linguistics, 1991, s. 306–311, DOI: 10.3115/112405.112467.

Acedański (2010): Szymon Acedański, *A Morphosyntactic Brill Tagger with Lexical Rules for Inflectional Languages*, w: *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010*, Springer, 2010, s. 3–14.

Andrzejczuk (2011): Anna Andrzejczuk, *Dwoje urodzin to brzmi dziwnie. Norma językowa dotycząca połączeń rzeczowników PT z liczebnikami a jej realizacja w tekstach Narodowego Korpusu Języka Polskiego i w tekstach internetowych*, „Język Polski" XCI (4), 2011, s. 273–283.

Bartosiak (2017): Tomasz Bartosiak, *Shared Forest Representation of Predicate-Argument Structures for Shared Syntactic Forests*, w: *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, red. Zygmunt Vetulani *et al.*, Fundacja UAM, Poznań, 2017, s. 410–414.

Bartosiak i Woliński (2015): Tomasz Bartosiak i Marcin Woliński, *On Genitive Clusters, Kleene Star, and an Exploding Parser*, w: *Proceedings of the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, red. Zygmunt Vetulani *et al.*, Poznań, 2015, s. 509–513.

Berger *et al.* (1996): Adam L. Berger, Stephen A. Della Pietra i Vincent J. Della Pietra, *A Maximum Entropy Approach to Natural Language Processing*, „Computational Linguistics" 22 (1), 1996, s. 39–71.

Bień (1991): Janusz S. Bień, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa UW, 1991.

——— (1996): Janusz S. Bień, *Komputerowa weryfikacja opisu składni polskiej*, raport techniczny 96-06 (227), Instytut Informatyki UW, Warszawa, 1996.

——— (1997): Janusz S. Bień, *Komputerowa weryfikacja formalnej gramatyki Świdzińskiego*, „Biuletyn Polskiego Towarzystwa Językoznawczego" LII, 1997, s. 147–164.

——— (2009): Janusz S. Bień, *Problemy formalnego opisu składni polskiej*, BEL Studio, Warszawa, 2009.

Bień *et al.* (1973): Janusz S. Bień, Witold Łukaszewicz i Stanisław Szpakowicz, *Opis systemu MARYSIA. I. Zasady pisania scenariusza i scenopisu*, Sprawozdania Instytutu Maszyn Matematycznych i Zakładu Obliczeń Numerycznych Uniwersytetu Warszawskiego nr 41, Wydawnictwa UW, 1973.

Bień *et al.* (2001): Janusz S. Bień, Krzysztof Szafran i Marcin Woliński, *Experimental Parsers of Polish*, w: *Current Issues in Formal Slavic Linguistics*, red. Gerhild Zybatow *et al.*, Linguistik International 5, Peter Lang, Frankfurt am Main, 2001, s. 185–190.

Bień i Saloni (1982): Janusz S. Bień i Zygmunt Saloni, *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*, „Prace Filologiczne" XXXI, 1982, s. 31–45.

Bień i Woliński (2003): Janusz S. Bień i Marcin Woliński, *Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej"*, w: *Prace językoznawcze dedykowane Profesor Jadwidze Sambor*, red. Jadwiga Linde-Usiekniewicz *et al.*, Wydział Polonistyki UW, Warszawa, 2003, s. 6–10.

Billot i Lang (1989): Sylvie Billot i Bernard Lang, *The Structure of Shared Forests in Ambiguous Parsing*, w: *Meeting of the Association for Computational Linguistics*, 1989, s. 143–151.

Böhmová *et al.* (2003): Alena Böhmová, Jan Hajič, Eva Hajičová i Barbora Hladká, *The Prague Dependency Treebank: A 3-Level Annotation Scenario*, w: *Treebanks. Building and Using Parsed Corpora*, red. Anne Abeillé, Kluwer Academic Publishers, 2003, s. 103–127.

Booth i Thompson (1973): Taylor R. Booth i Richard A. Thompson, *Applying Probability Measures to Abstract Languages*, „IEEE Transactions on Computers" 22, 1973, s. 442–450.

Branco (2009): António Branco, *LogicalFormBanks, the Next Generation of Semantically Annotated Corpora: Key Issues in Construction Methodology*, w: *Recent Advances in Intelligent Information Systems*, red. Mieczysław A. Kłopotek *et al.*, EXIT, Warszawa, 2009, s. 3–11.

Brants *et al.* (2003): Thorsten Brants, Wojciech Skut i Hans Uszkoreit, *Syntactic Annotation of a German Newspaper Corpus*, w: *Treebanks: Building and Using Parsed Corpora*, red. Anne Abeillé, Springer, 2003, s. 73–87.

Collins (1997): Michael Collins, *Three generative, lexicalised models for statistical parsing*, w: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL'98, Association for Computational Linguistics, 1997, s. 16–23, DOI: 10.3115/976909.979620.

Colmerauer (1978): Alain Colmerauer, *Metamorphosis grammar*, w: *Natural Language Communication with Computers*, red. Leonard Bolc, Lecture Notes in Computer Science 63, Springer, 1978, s. 133–189.

Dalrymple (2001): Mary Dalrymple, *Lexical Functional Grammar*, Academic Press, 2001.

Derwojedowa (2000): Magdalena Derwojedowa, *Porządek linearny składników zdania elementarnego w języku polskim*, Elipsa, Warszawa, 2000.

——— (2011): Magdalena Derwojedowa, *Grupy liczebnikowe we współczesnym języku polskim*, Wydział Polonistyki UW, Warszawa, 2011.

Derwojedowa i Rudolf (2003): Magdalena Derwojedowa i Michał Rudolf, *Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu*, „Poradnik Językowy" (3), 2003.

Dębowski (2004): Łukasz Dębowski, *Trigram morphosyntactic tagger for Polish*, w: *Intelligent Information Processing and Web Mining. Proceedings of the International IIS:IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004*, red. Mieczysław A. Kłopotek *et al.*, Springer, 2004, s. 409–413.

——— (2009): Łukasz Dębowski, *Valence extraction using EM selection and co-occurrence matrices*, „Language Resources and Evaluation" 43, 2009, s. 301–327.

Dębowski i Woliński (2007): Łukasz Dębowski i Marcin Woliński, *Argument co-occurrence matrix as a description of verb valence*, w: *Proceedings of the 3rd Language & Technology Conference*, red. Zygmunt Vetulani, Poznań, 2007, s. 260–264.

Doroszewski (1958–1969): Witold Doroszewski, red., *Słownik języka polskiego PAN*, Wiedza Powszechna – PWN, Warszawa, 1958–1969.

Francez i Wintner (2011): Nissim Francez i Shuly Wintner, *Unification Grammars*, Cambridge University Press, 2011, DOI: 10.1017/CBO9781139013574.

Gruszczyński (1989): Włodzimierz Gruszczyński, *Fleksja rzeczowników pospolitych we współczesnej polszczyźnie pisanej*, Prace językoznawcze 122, Ossolineum, Wrocław, 1989.

——— (2001): Włodzimierz Gruszczyński, *Rzeczowniki w słowniku gramatycznym współczesnego języka polskiego*, w: *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15 000 dni pracy naukowej*, red. Włodzimierz Gruszczyński *et al.*, Wydawnictwo Uniwersytetu w Białymstoku, 2001, s. 99–116.

Gruszczyński i Saloni (2006): Włodzimierz Gruszczyński i Zygmunt Saloni, *Notowanie informacji o odmianie rzeczowników w projektowanym Słowniku gramatycznym języka polskiego*, w: *Od fonemu do tekstu. Prace dedykowane Profesorowi Romanowi Laskowskiemu*, red. Ireneusz Bobrowski *et al.*, Lexis, Kraków, 2006, s. 203–213.

Grzegorczykowa i Puzynina (1973): Renata Grzegorczykowa i Jadwiga Puzynina, red., *Indeks a tergo do Słownika języka polskiego pod redakcją Witolda Doroszewskiego*, PWN, Warszawa, 1973.

Hajič *et al.* (2001): Jan Hajič, Barbora Vidová-Hladká i Petr Pajas, *The Prague Dependency Treebank: Annotation Structure and Support*, w: *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, USA, 2001, s. 105–114.

Hajnicz (2011): Elżbieta Hajnicz, *Automatyczne tworzenie semantycznego słownika walencyjnego*, EXIT, Warszawa, 2011.

Hajnicz *et al.* (2015): Elżbieta Hajnicz, Bartłomiej Nitoń, Agnieszka Patejuk, Adam Przepiórkowski i Marcin Woliński, *Internetowy słownik walencyjny języka polskiego oparty na danych korpusowych*, „Prace Filologiczne" LXV, 2015, s. 95–110.

Hajnicz *et al.* (2016a): Elżbieta Hajnicz, Anna Andrzejczuk i Tomasz Bartosiak, *Semantic Layer of the Valence Dictionary of Polish Walenty*, w: *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, red. Nicoletta Calzolari *et al.*, ELRA, European Language Resources Association (ELRA), 2016, s. 2625–2632.

Hajnicz *et al.* (2016b): Elżbieta Hajnicz, Agnieszka Patejuk, Adam Przepiórkowski i Marcin Woliński, *Walenty: słownik walencyjny języka polskiego z bogatym komponentem frazeologicznym*, w: *Výzkum slovesné valence ve slovanských zemích*, red. Karolina Skwarska *et al.*, Slovanský ústav AV ČR, Praga, 2016, s. 71–102.

Hajnicz i Andrzejczuk (2018): Elżbieta Hajnicz i Anna Andrzejczuk, *Poziom semantyczny słownika walencyjnego Walenty*, złożone do publikacji, 2018.

Jackendoff (1977): Ray Jackendoff, *X-bar Syntax: A Study of Phrase Structure*, Linguistic Inquiry Monographs 2, MIT Press, 1977.

Jadacka (2005): Hanna Jadacka, *Kultura języka polskiego. Fleksja, słowotwórstwo, składnia*, PWN, Warszawa, 2005.

Jassem (2006): Krzysztof Jassem, *Przetwarzanie tekstów polskich w systemie tłumaczenia automatycznego POLENG*, Wydawnictwo Naukowe UAM, Poznań, 2006.

Kallas (1978): Krystyna Kallas, *Struktura syntaktyczna polskich konstrukcji apozycyjnych*, „Slavia Orientalis" XXVII (3), 1978, s. 345–350.

——— (1980): Krystyna Kallas, *Grupy apozycyjne we współczesnym języku polskim*, Wydawnictwo UMK, Toruń, 1980.

Karpowicz (1994): Tomasz Karpowicz, *Próba określenia normy składniowej dotyczącej użycia liczebników zbiorowych*, w: *Polszczyzna a/i Polacy u schyłku XX wieku: zbiór studiów*, red. Kwiryna Handke *et al.*, Slawistyczny Ośrodek Wydawniczy, Warszawa, 1994, s. 113–122.

Kieraś *et al.* (2017): Witold Kieraś, Dorota Komosińska, Emanuel Modrzejewski i Marcin Woliński, *Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish*, w: *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, red. Kamil Ekštein *et al.*, Lecture Notes in Computer Science 10415, Springer, 2017, s. 308–316.

Kieraś i Woliński (2018): Witold Kieraś i Marcin Woliński, *Manually Annotated Corpus of Polish Texts Published between 1830 and 1918*, w: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, red. Nicoletta Calzolari *et al.*, European Language Resources Association (ELRA), 2018, s. 3854–3859.

Klemensiewicz (1969): Zenon Klemensiewicz, *Zarys składni polskiej*, PWN, Warszawa, 1969.

Kobyliński (2014): Łukasz Kobyliński, *PoliTa: A multitagger for Polish*, w: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, red. Nicoletta Calzolari *et al.*, ELRA, 2014, s. 2949–2954.

Kobyliński i Kieraś (2016): Łukasz Kobyliński i Witold Kieraś, *Part of Speech Tagging for Polish: State of the Art and Future Perspectives*, w: *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, 2016.

König *et al.* (2003): Esther König, Wolfgang Lezius i Holger Voormann, *TIGERSearch 2.1 User's Manual*, raport techniczny, Universität Stuttgart, 2003.

König i Lezius (2003): Esther König i Wolfgang Lezius, *The TIGER language – A Description Language for Syntax Graphs, Formal Definition*, raport techniczny, Universität Stuttgart, 2003.

Krasnowska *et al.* (2012): Katarzyna Krasnowska, Witold Kieraś, Marcin Woliński i Adam Przepiórkowski, *Using Tree Transducers for Detecting Errors in a Treebank of Polish*, w: *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic*, red. Petr Sojka *et al.*, Lecture Notes in Artificial Intelligence 7499, Springer, 2012, s. 119–126.

Krasnowska-Kieraś (2017): Katarzyna Krasnowska-Kieraś, *Morphosyntactic Disambiguation for Polish with Bi-LSTM Neural Networks*, w: *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, red. Zygmunt Vetulani *et al.*, Fundacja UAM, Poznań, 2017, s. 367–371.

Krasnowska-Kieraś i Patejuk (2015): Katarzyna Krasnowska-Kieraś i Agnieszka Patejuk, *Integrating Polish LFG with External Morphology*, w: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*, red. Markus Dickinson *et al.*, IPI PAN, Warszawa, 2015, s. 134–147.

Kupść (2000): Anna Kupść, *An HPSG Grammar of Polish Clitics*, praca doktorska, IPI PAN, i Université Paris 7, Warszawa, 2000.

Kurcz *et al.* (1990): Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran i Jerzy Woronczak, *Słownik frekwencyjny polszczyzny współczesnej*, Wydawnictwo Instytutu Języka Polskiego PAN, Kraków, 1990.

Lewicki (1976): Andrzej Maria Lewicki, *Wprowadzenie do frazeologii syntaktycznej. Teoria zwrotu frazeologicznego*, Prace naukowe Uniwersytetu Śląskiego w Katowicach 116, 1976.

Lezius (2002): Wolfgang Lezius, *TIGERSearch – Ein Suchwerkzeug für Baumbanken*, w: *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, red. Stephan Busemann, Saarbrücken, 2002.

Maier i Lichte (2011): Wolfgang Maier i Timm Lichte, *Characterizing Discontinuity in Constituent Treebanks*, w: *Formal Grammar: 14th International Conference, FG 2009, Bordeaux, France, July 25-26, 2009, Revised Selected Papers*, red. Philippe Groote *et al.*, Springer, 2011, s. 167–182.

Mańczak (1956): Witold Mańczak, *Ile rodzajów jest w polskim?*, „Język Polski" XXXVI (2), 1956, s. 116–121, http://mbc.malopolska.pl/Content/17856/index.djvu.

Marciniak (2001): Małgorzata Marciniak, *Algorytmy implementacyjne syntaktycznych reguł koreferencji zaimków dla języka polskiego w terminach HPSG*, praca doktorska, IPI PAN, Warszawa, 2001.

Marciniak *et al.* (2000): Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski i Anna Kupść, *An HPSG-Annotated Test Suite for Polish*, w: *Proceedings of the Linguistic Resources and Evaluation Conference*, Athens, 2000.

Marciniak *et al.* (2011): Małgorzata Marciniak, Agata Savary, Piotr Sikora i Marcin Woliński, *Toposław – A Lexicographic Framework for Multi-word Units*, w: *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6–8, 2009, Revised Selected Papers*, red. Zygmunt Vetulani, Lecture Notes in Artificial Intelligence 6562, Springer, 2011, s. 139–150.

Marcus *et al.* (1993): Mitchell P. Marcus, Mary Ann Marcinkiewicz i Beatrice Santorini, *Building a Large Annotated Corpus of English: The Penn Treebank*, „Computational Linguistics" 19 (2), 1993, s. 313–330.

Marek *et al.* (2008): Torsten Marek, Joakim Lundborg i Martin Volk, *Extending the TIGER query language with universal quantification*, w: *KONVENS 2008: 9. Konferenz zur Verarbeitung natürlicher Sprache*, 2008, s. 5–17.

Meľčuk (1988): Igor Meľčuk, *Dependency Syntax: Theory and Practice*, State University of New York Press, 1988.

Mikolov *et al.* (2013): Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado i Jeffrey Dean, *Distributed Representations of Words and Phrases and their Compositionality*, „CoRR" abs/1310.4546, 2013, http://arxiv.org/abs/1310.4546.

Miyao i Tsujii (2002): Yusuke Miyao i Jun'ichi Tsujii, *Maximum Entropy Estimation for Feature Forests*, w: *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, Morgan Kaufmann, 2002, s. 292–297.

——— (2008): Yusuke Miyao i Jun'ichi Tsujii, *Feature Forest Models for Probabilistic HPSG Parsing*, „Computational Linguistics" 34 (1), 2008, s. 35–80, DOI: 10.1162/coli.2008.34.1.35.

Mykowiecka (1999): Agnieszka Mykowiecka, *Opis składniowy polskich konstrukcji względnych w formalizmie HPSG*, praca doktorska, IPI PAN, Warszawa, 1999.

Nivre (2003): Joakim Nivre, *Theory-Supporting Treebanks*, w: *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, 2003, s. 117–128.

Nivre *et al.* (2016): Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira *et al.*, *Universal Dependencies v1: A Multilingual Treebank Collection*, w: *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, red. Nicoletta Calzolari *et al.*, ELRA, European Language Resources Association (ELRA), 2016, s. 1659–1666.

Obrębski (2002): Tomasz Obrębski, *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej*, praca doktorska, IPI PAN, Warszawa, 2002.

Ogrodniczuk (2003a): Maciej Ogrodniczuk, *Nowa edycja wzbogaconego korpusu słownika frekwencyjnego*, w: *Językoznawstwo w Polsce. Stan i perspektywy*, red. Stanisław Gajda, Komitet Językoznawstwa PAN oraz Instytut Filologii Polskiej, Uniwersytet Opolski, Opole, 2003, s. 181–190.

——— (2003b): Maciej Ogrodniczuk, *Rozszerzenie opisów morfologicznych w tekstach korpusu „Słownika frekwencyjnego polszczyzny współczesnej"*, w: *Prace lingwistyczne dedykowane prof. Jadwidze Sambor*, red. Romuald Huszcza *et al.*, Wydział Polonistyki UW, Warszawa, 2003, s. 164–168.

——— (2005): Maciej Ogrodniczuk, *An extension of Świdziński's grammar of Polish*, „Archives of Control Sciences" 15 (3), 2005, s. 393–402.

——— (2006): Maciej Ogrodniczuk, *Weryfikacja korpusu wypowiedników polskich (z wykorzystaniem gramatyki formalnej Świdzińskiego)*, praca doktorska, Wydział Neofilologii UW, Warszawa, 2006.

Patejuk (2015): Agnieszka Patejuk, *Unlike coordination in Polish: an LFG account*, praca doktorska, Instytut Języka Polskiego PAN, Kraków, 2015.

——— (2017): Agnieszka Patejuk, *A Gapping Analysis of Lexicalised Comparative Constructions*, w: *The Proceedings of the LFG'17 Conference*, red. Miriam Butt *et al.*, CSLI Publications, Stanford, CA, 2017, s. 306–326.

Patejuk i Przepiórkowski (2012): Agnieszka Patejuk i Adam Przepiórkowski, *Towards an LFG Parser for Polish: An Exercise in Parasitic Grammar Development*, w: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, red. Nicoletta Calzolari *et al.*, ELRA, 2012, s. 3849–3852.

——— (2014): Agnieszka Patejuk i Adam Przepiórkowski, *Synergistic Development of Grammatical Resources: A Valence Dictionary, an LFG Grammar, and an LFG Structure Bank for Polish*, w: *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*, red. Verena Henrich *et al.*, Department of Linguistics, University of Tübingen, 2014, s. 113–126.

——— (2015): Agnieszka Patejuk i Adam Przepiórkowski, *Parallel Development of Linguistic Resources: Towards a Structure Bank of Polish*, „Prace Filologiczne" LXV, 2015, s. 255–270.

——— (2016): Agnieszka Patejuk i Adam Przepiórkowski, *Reducing Grammatical Functions in Lexical Functional Grammar*, w: *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, red. Doug Arnold *et al.*, CSLI Publications, Stanford, CA, 2016, s. 541–559.

Pereira i Warren (1980): Fernando Pereira i David H. D. Warren, *Definite clause grammars for language analysis – a survey of the formalism and a comparison with augmented transition networks*, „Artificial Intelligence" 13, 1980, s. 231–278.

Piasecki (2007): Maciej Piasecki, *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*, „Task Quarterly" 11 (1–2), 2007, s. 151–167.

Piasecki *et al.* (2009): Maciej Piasecki, Stanisław Szpakowicz i Bartosz Broda, *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej, 2009.

Piskorski *et al.* (2004): Jakub Piskorski, Peter Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski i Marcin Woliński, *Information Extraction for Polish Using the SProUT Platform*, w: *Intelligent Information Processing and Web Mining*, red. Mieczysław Kłopotek *et al.*, Advances in Soft Computing, Springer, 2004, s. 227–236.

Polański (1980–1992): Kazimierz Polański, red., *Słownik syntaktyczno-generatywny czasowników polskich*, t. I–V, Ossolineum, Wrocław, 1980–1992.

Pollard i Sag (1987): Carl Pollard i Ivan A. Sag, *Information-Based Syntax and Semantics. Vol. I: Fundamentals*, CSLI Publications, 1987.

——— (1994): Carl Pollard i Ivan A. Sag, *Head-driven Phrase Structure Grammar*, Chicago University Press / CSLI Publications, 1994.

Przepiórkowski (1999): Adam Przepiórkowski, *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*, praca doktorska, Universität Tübingen, 1999.

——— (2003a): Adam Przepiórkowski, *A Hierarchy of Polish Genders*, w: *Generative Linguistics in Poland: Morphosyntactic Investigations*, red. Piotr Bański *et al.*, IPI PAN, Warszawa, 2003, s. 109–122.

——— (2003b): Adam Przepiórkowski, *Składniowe uwarunkowania znakowania morfosyntaktycznego w Korpusie IPI PAN*, „Polonica" XXII–XXIII, 2003, s. 57–76.

——— (2004a): Adam Przepiórkowski, *Korpus IPI PAN. Wersja wstępna*, IPI PAN, Warszawa, 2004.

——— (2004b): Adam Przepiórkowski, *O wartości przypadka podmiotów liczebnikowych*, „Biuletyn Polskiego Towarzystwa Językoznawczego" LX, 2004, s. 133–143.

Przepiórkowski (2006): Adam Przepiórkowski, *Poliqarp: Przeszukiwarka korpusowa dla lingwistów*, w: *Korpusy w angielsko-polskim językoznawstwie kontrastywnym*, red. Anna Duszak *et al.*, Universitas, Kraków, 2006, s. 398–426.

——— (2008): Adam Przepiórkowski, *Powierzchniowe przetwarzanie języka polskiego*, EXIT, Warszawa, 2008.

——— (2009): Adam Przepiórkowski, *A comparison of two morphosyntactic tagsets of Polish*, w: *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, red. Violetta Koseska-Toszewa *et al.*, Warszawa, 2009, s. 138–144.

——— (2016): Adam Przepiórkowski, *Against the Argument–Adjunct Distinction in Functional Generative Description*, „The Prague Bulletin of Mathematical Linguistics" 106, 2016, s. 5–20.

——— (2017): Adam Przepiórkowski, *Argumenty i modyfikatory w gramatyce i w słowniku*, Wydawnictwa UW, Warszawa, 2017.

Przepiórkowski *et al.* (2002): Adam Przepiórkowski, Anna Kupść, Małgorzata Marciniak i Agnieszka Mykowiecka, *Formalny opis języka polskiego: Teoria i implementacja*, EXIT, Warszawa, 2002.

Przepiórkowski *et al.* (2003): Adam Przepiórkowski, Piotr Bański, Łukasz Dębowski, Elżbieta Hajnicz i Marcin Woliński, *Konstrukcja korpusu IPI PAN*, „Polonica" XXII–XXIII, 2003, s. 33–38.

Przepiórkowski *et al.* (2012): Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk, red., *Narodowy Korpus Języka Polskiego*, PWN, Warszawa, 2012.

Przepiórkowski *et al.* (2014a): Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk i Marcin Woliński, *Extended phraseological information in a valence dictionary for NLP applications*, w: *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, Association for Computational Linguistics i Dublin City University, 2014, s. 83–91.

Przepiórkowski *et al.* (2014b): Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski i Marek Świdziński, *Walenty: Towards a comprehensive valence dictionary of Polish*, w: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, red. Nicoletta Calzolari *et al.*, ELRA, 2014, s. 2785–2792.

Przepiórkowski *et al.* (2014c): Adam Przepiórkowski, Filip Skwarski, Elżbieta Hajnicz, Agnieszka Patejuk, Marek Świdziński i Marcin Woliński, *Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego*, „Polonica" XXXIII, 2014, s. 159–178.

Przepiórkowski *et al.* (2017): Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz i Zdeňka Urešová, *Phraseology in two Slavic Valency Dictionaries: Limitations and Perspectives*, „International Journal of Lexicography" 30 (1), 2017, s. 1–38.

Przepiórkowski i Świdziński (1997): Adam Przepiórkowski i Marek Świdziński, *Polish Verbal Negation Revisited: A Metamorphosis vs. HPSG Account*, raport techniczny 829, IPI PAN, Warszawa, 1997.

Przepiórkowski i Woliński (2003a): Adam Przepiórkowski i Marcin Woliński, *A Flexemic Tagset for Polish*, w: *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*, 2003, s. 33–40.

Przepiórkowski i Woliński (2003b): Adam Przepiórkowski i Marcin Woliński, *A Morphosyntactic Tagset for Polish*, w: *Investigations into Formal Slavic Linguistics (Contributions of the Fourth European Conference on Formal Description on Slavic Languages)*, red. Peter Kosta *et al.*, 2003, s. 349–362.

——— (2003c): Adam Przepiórkowski i Marcin Woliński, *The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish*, w: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, 2003, s. 109–116.

Radziszewski (2013): Adam Radziszewski, *A Tiered CRF Tagger for Polish*, w: *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, red. Robert Bembenik *et al.*, Springer, 2013, s. 215–230, DOI: 10.1007/978-3-642-35647-6_16.

Radziszewski i Śniatowski (2011): Adam Radziszewski i Tomasz Śniatowski, *A Memory-Based Tagger for Polish*, w: *Proceedings of the 5th Language & Technology Conference*, red. Zygmunt Vetulani, Poznań, 2011, s. 556–560.

Rogozińska (2016): Dominika Rogozińska, *Automatyczne metody ujednoznaczniania drzew rozbioru wypowiedzeń w języku polskim jako ostatnia faza przetwarzania parsera Świgra*, praca magisterska, Uniwersytet Warszawski, 2016.

Rosén *et al.* (2006): Victoria Rosén, Koenraad de Smedt i Paul Meurer, *Towards a Toolkit Linking Treebanking to Grammar Development*, w: *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, red. Jan Hajič *et al.*, 2006, s. 55–66.

Rosén *et al.* (2007): Victoria Rosén, Paul Meurer, Koenraad De Smedt, Miriam Butt i Tracy Holloway King, *Designing and implementing discriminants for LFG grammars*, w: *The proceedings of the LFG'07 conference*, 2007, s. 397–417.

Saloni (1974a): Zygmunt Saloni, *Klasyfikacja gramatyczna leksemów polskich (cz. 1)*, „Język Polski" LIV (1), 1974, s. 3–13, http://mbc.malopolska.pl/Content/57081/.

——— (1974b): Zygmunt Saloni, *Klasyfikacja gramatyczna leksemów polskich (cz. 2)*, „Język Polski" LIV (2), 1974, s. 93–101, http://mbc.malopolska.pl/Content/57082/.

——— (1976): Zygmunt Saloni, *Kategoria rodzaju we współczesnym języku polskim*, w: *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, Ossolineum, Wrocław, 1976, s. 41–75, http://rcin.org.pl/dlibra/doccontent?id=2040.

——— (1977): Zygmunt Saloni, *Kategorie gramatyczne liczebników we współczesnym języku polskim*, w: *Studia gramatyczne I*, Wrocław, 1977, s. 145–173.

——— (1981): Zygmunt Saloni, *Uwagi o opisie fleksyjnym tzw. zaimków rzeczownych*, w: *Acta Universitatis Lodziensis. Folia Linguistica 2*, Wydawnictwo UŁ, 1981, s. 265–271.

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

——— (2001): Zygmunt Saloni, *Czasownik polski. Odmiana, słownik*, Wiedza Powszechna, Warszawa, 2001.

——— (2004): Zygmunt Saloni, *O podrzędnym mianowniku przyrzeczownikowym w języku polskim*, w: *Studia z gramatyki i semantyki języka polskiego*, red. Marek Wiśniewski *et al.*, UMK, Toruń, 2004, s. 55–65.

——— (2005): Zygmunt Saloni, *O przypadkach w języku polskim (na marginesie artykułu Adama Przepiórkowskiego)*, „Biuletyn Polskiego Towarzystwa Językoznawczego" LXI, 2005, s. 27–48.

Saloni (2007): Zygmunt Saloni, *Czasownik polski. Odmiana, słownik*, wydanie 3, Wiedza Powszechna, Warszawa, 2007.

——— (2016): Zygmunt Saloni, *Systematyzacja wzorów fleksyjnych dla „Słownika gramatycznego języka polskiego"*, „Polonica" XXXVI, 2016, s. 5–18.

Saloni *et al.* (2007a): Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński i Robert Wołosz, *Grammatical dictionary of Polish. Presentation by the authors*, „Studies in Polish Linguistics" 4, 2007, s. 5–25.

——— (2007b): Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński i Robert Wołosz, *Słownik gramatyczny języka polskiego*, Wiedza Powszechna, Warszawa, 2007, s. 177 + CD.

Saloni *et al.* (2012): Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński i Danuta Skowrońska, *Słownik gramatyczny języka polskiego*, wydanie 2, Warszawa, 2012.

——— (2015): Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński i Danuta Skowrońska, *Słownik gramatyczny języka polskiego*, wydanie 3, 2015, on–line, http://sgjp.pl.

Saloni i Świdziński (2001): Zygmunt Saloni i Marek Świdziński, *Składnia współczesnego języka polskiego*, wydanie 5, PWN, Warszawa, 2001.

Saloni i Woliński (2003): Zygmunt Saloni i Marcin Woliński, *A Computerized Description of Polish Conjugation*, w: *Investigations into Formal Slavic Linguistics (Contributions of the Fourth European Conference on Formal Description on Slavic Languages)*, red. Peter Kosta *et al.*, 2003, s. 373–384.

——— (2004): Zygmunt Saloni i Marcin Woliński, *Jak pracowaliśmy na książką „Czasownik polski"*, „Biuletyn Polskiego Towarzystwa Językoznawczego" 60, 2004, s. 145–156.

——— (2005): Zygmunt Saloni i Marcin Woliński, *O planowanej postaci Słownika gramatycznego języka polskiego*, „Poradnik Językowy" 9, 2005, s. 74–81.

Savary (2005): Agata Savary, *MULTIFLEX. User's Manual and Technical Documentation. Version 1.0*, raport techniczny 285, François Rabelais University of Tours, 2005.

Seddah *et al.* (2013): Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Yuval Marton, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński i Alina Wróblewska, *Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages*, w: *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, Association for Computational Linguistics, 2013, s. 146–182.

Szafran (1993): Krzysztof Szafran, *Automatyczna analiza fleksyjna tekstu polskiego (na podstawie „Schematycznego indeksu a tergo" Jana Tokarskiego)*, praca doktorska, Wydział Polonistyki UW, 1993.

——— (1996): Krzysztof Szafran, *Analizator morfologiczny SAM-95: opis użytkowy*, raport techniczny 96-05 (226), Instytut Informatyki UW, Warszawa, 1996.

Szpakowicz (1978): Stanisław Szpakowicz, *Automatyczna analiza składniowa polskich zdań pisanych*, praca doktorska, Instytut Informatyki UW, 1978.

——— (1983): Stanisław Szpakowicz, *Formalny opis składniowy zdań polskich*, Wydawnictwa UW, Warszawa, 1983.

Szpakowicz i Świdziński (1981): Stanisław Szpakowicz i Marek Świdziński, *Zarys klasyfikacji schematów zdaniowych we współczesnej polszczyźnie pisanej*, „Polonica" VII, 1981, s. 5–35.

——— (1990): Stanisław Szpakowicz i Marek Świdziński, *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej*, „Studia Gramatyczne" IX, 1990, s. 9–54.

Szupryczyńska (1996): Maria Szupryczyńska, *Problem pozycji składniowej*, w: *Polonistyka toruńska Uniwersytetowi w 50. rocznicę utworzenia UMK. Językoznawstwo*, red. Krystyna Kallas, Wydawnictwo UMK, Toruń, 1996, s. 135–144.

Świdziński (1981): Marek Świdziński, *O spójnikach i partykułach odmiennych przez osobę*, w: *Acta Universitatis Lodziensis. Folia Linguistica 2*, Wydawnictwo UŁ, 1981, s. 273–284, http://hdl.handle.net/11089/14672.

——— (1987): Marek Świdziński, *Formalny opis składniowy polskich zdań o składniku zdaniowym*, (maszynopis powielony), praca habilitacyjna, Wydział Polonistyki UW, Warszawa, 1987.

——— (1989): Marek Świdziński, *A Dependency Syntax of Polish*, w: *Metataxis in Practice. Dependency Syntax for Multilingual Machine Translation*, Foris, 1989, s. 69–87.

——— (1992): Marek Świdziński, *Gramatyka formalna języka polskiego*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa UW, Warszawa, 1992.

——— (1994): Marek Świdziński, *Syntactic Dictionary of Polish Verbs*, maszynopis, Uniwersytet Warszawski i Universiteit van Amsterdam, 1994.

——— (1996): Marek Świdziński, *Własności składniowe wypowiedników polskich*, Elipsa, Warszawa, 1996.

——— (2001): Marek Świdziński, *Transmisja oddolna i odgórna negacji w zdaniu polskim: konstrukcje ze spójnikiem negatywnym*, w: *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15 000 dni pracy naukowej*, Wydawnictwo Uniwersytetu w Białymstoku, 2001, s. 275–285.

——— (2005): Marek Świdziński, *Proper treatment of some peculiarities of Polish numerals: metamorphosis approach*, w: *Human Language Technologies as a Challenge for Computer Science and Linguistics. 2nd Language and Technology Conference*, Poznań, 2005, s. 182–186.

Świdziński *et al.* (2002): Marek Świdziński, Magdalena Derwojedowa i Michał Rudolf, *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich*, „Biuletyn Polskiego Towarzystwa Językoznawczego" LVIII, 2002, s. 187–199.

Świdziński *et al.* (2013): Marek Świdziński, Marcin Woliński i Katarzyna Głowińska, *A new version of the formal grammar of Polish: corpus-backed improvements and corrections*, w: *Travaux de slavistique, Actes du VIème congrès de la Slavic Linguistic Society*, red. Irina Kor Chahine *et al.*, Université de Provence, Aix-en-Provence, 2013, s. 299–309.

Świdziński i Woliński (2007): Marek Świdziński i Marcin Woliński, *Towards a new version of the formal grammar of Polish: the NP redefined*, w: *Formal Description of Slavic Languages FDSL-7. University of Leipzig, 30 November – 2 December, 2007. Book of Abstracts*, Leipzig, 2007, s. 101–103.

Świdziński i Woliński (2009): Marek Świdziński i Marcin Woliński, *A new formal definition of Polish nominal phrases*, w: *Aspects of Natural Language Processing. Essays dedicated to Leonard Bolc on the Occasion of His 75th Birthday*, red. Małgorzata Marciniak *et al.*, Lecture Notes in Computer Science 5070, Springer, 2009, s. 143–162.

——— (2010): Marek Świdziński i Marcin Woliński, *Towards a Bank of Constituent Parse Trees for Polish*, w: *Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, September 2010, Proceedings*, red. Petr Sojka, Lecture Notes in Artificial Intelligence 6231, Springer, 2010, s. 197–204.

Tokarski (1951): Jan Tokarski, *Czasowniki polskie. Formy, typy, wyjątki. Słownik*, Warszawa, 1951.

——— (1973): Jan Tokarski, *Fleksja polska*, PWN, Warszawa, 1973.

——— (1993): Jan Tokarski, *Schematyczny indeks a tergo polskich form wyrazowych*, red. Zygmunt Saloni, PWN, Warszawa, 1993.

Vetulani (2004): Zygmunt Vetulani, *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej*, EXIT, Warszawa, 2004.

Vetulani *et al.* (2010): Zygmunt Vetulani, Jacek Marciniak, Tomasz Obrębski, Grażyna Vetulani, Adam Dąbrowski, Marek Kubis, Jędrzej Osiński, Justyna Walkowska, Piotr Kubacki i Krzysztof Witalewski, *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego*, Wydawnictwo Naukowe UAM, Poznań, 2010.

Waszczuk (2012): Jakub Waszczuk, *Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language*, w: *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, 2012, s. 2789–2804.

Wells (1947): Rulon S. Wells, *Immediate Constituents*, „Language" 23 (2), 1947, s. 81–117.

Woliński (2001): Marcin Woliński, *Rodzajów w polszczyźnie jest osiem*, w: *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15 000 dni pracy naukowej*, Wydawnictwo Uniwersytetu w Białymstoku, 2001, s. 303–305.

——— (2003): Marcin Woliński, *System znaczników morfosyntaktycznych w korpusie IPI PAN*, „Polonica" XXII–XXIII, 2003, s. 39–55.

——— (2004): Marcin Woliński, *Komputerowa weryfikacja gramatyki Świdzińskiego*, praca doktorska, IPI PAN, Warszawa, 2004.

——— (2005): Marcin Woliński, *An efficient implementation of a large grammar of Polish*, „Archives of Control Sciences" 15(LI) (3), 2005, s. 251–258.

——— (2006a): Marcin Woliński, *Jak się nie zgubić w lesie, czyli o wynikach analizy składniowej według gramatyki Świdzińskiego*, „Poradnik Językowy" 9, 2006, s. 102–114.

——— (2006b): Marcin Woliński, *Morfeusz – a Practical Tool for the Morphological Analysis of Polish*, w: *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, red. Mieczysław Kłopotek *et al.*, Springer, 2006, s. 503–512.

——— (2009): Marcin Woliński, *A Relational Model of Polish Inflection in Grammatical Dictionary of Polish*, w: *Human Language Technology: Challenges of the Information Society*, red. Zygmunt Vetulani *et al.*, Lecture Notes in Artificial Intelligence 5603, Springer, 2009, s. 96–106.

Woliński (2010): Marcin Woliński, *Dendrarium – an Open Source Tool for Treebank Building*, w: *Intelligent Information Systems*, red. Mieczysław A. Kłopotek *et al.*, Siedlce, 2010, s. 193–204.

——— (2014): Marcin Woliński, *Morfeusz Reloaded*, w: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, red. Nicoletta Calzolari *et al.*, ELRA, 2014, s. 1106–1111.

——— (2015): Marcin Woliński, *Deploying the New Valency Dictionary Walenty in a DCG Parser of Polish*, w: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*, red. Markus Dickinson *et al.*, IPI PAN, Warszawa, 2015, s. 221–229.

Woliński *et al.* (2011): Marcin Woliński, Katarzyna Głowińska i Marek Świdziński, *A Preliminary Version of Składnica – a Treebank of Polish*, w: *Proceedings of the 5th Language & Technology Conference*, red. Zygmunt Vetulani, Poznań, 2011, s. 299–303.

Woliński *et al.* (2012): Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk, Adam Przepiórkowski i Łukasz Szałkiewicz, *PoliMorf: a (not so) new open morphological dictionary for Polish*, w: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, red. Nicoletta Calzolari *et al.*, ELRA, 2012, s. 860–864.

Woliński *et al.* (2018): Marcin Woliński, Elżbieta Hajnicz i Tomasz Bartosiak, *A New Version of the Składnica Treebank of Polish Harmonised with the Walenty Valency Dictionary*, w: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, red. Nicoletta Calzolari *et al.*, European Language Resources Association (ELRA), 2018, s. 1839–1844.

Woliński i Kieraś (2016): Marcin Woliński i Witold Kieraś, *The On-Line Version of Grammatical Dictionary of Polish*, w: *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, red. Nicoletta Calzolari *et al.*, ELRA, European Language Resources Association (ELRA), 2016, s. 2589–2594.

Woliński i Przepiórkowski (2001): Marcin Woliński i Adam Przepiórkowski, *Projekt anotacji morfosynktaktycznej korpusu języka polskiego*, Prace IPI PAN 938, IPI PAN, 2001.

Woliński i Rogozińska (2013): Marcin Woliński i Dominika Rogozińska, *First Experiments in PCFG-like Disambiguation of Constituency Parse Forests for Polish*, w: *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, red. Zygmunt Vetulani, Fundacja UAM, Poznań, 2013, s. 343–347.

——— (2016): Marcin Woliński i Dominika Rogozińska, *Experiments in PCFG-like Disambiguation of Constituency Parse Forests for Polish*, w: *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*, red. Zygmunt Vetulani *et al.*, Lecture Notes in Artificial Intelligence 9561, Springer, 2016, s. 146–158.

Woliński i Zaborowski (2012): Marcin Woliński i Andrzej Zaborowski, *An Ambiguity Aware Treebank Search Tool*, w: *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic*, red. Petr Sojka *et al.*, Lecture Notes in Artificial Intelligence 7499, Springer, 2012, s. 88–94.

Wołosz (2005): Robert Wołosz, *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*, EXIT, Warszawa, 2005.

Wróblewska (2012): Alina Wróblewska, *Polish Dependency Bank*, „Linguistic Issues in Language Technology" 7 (1), 2012.

——— (2014): Alina Wróblewska, *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*, praca doktorska, IPI PAN, Warszawa, 2014.

Wróblewska i Wieczorek (2018): Alina Wróblewska i Aleksandra Wieczorek, *Status morfoskładniowy wyrazu jako we współczesnej polszczyźnie*, „Język Polski" XCVIII (3), 2018, s. 16–30.

Wróblewska i Woliński (2012): Alina Wróblewska i Marcin Woliński, *Preliminary Experiments in Polish Dependency Parsing*, w: *Security and Intelligent Information Systems: International Joint Conference, SIIS 2011, Warszawa, Poland, June 13-14, 2011, Revised Selected Papers*, red. Pascal Bouvry *et al.*, Lecture Notes in Computer Science 7053, Springer, 2012, s. 279–292.

Yoshida (2006): Kazuhiro Yoshida, *Amis – A maximum entropy estimator for feature forests*, online, 2006, http://nactem.ac.uk/amis/.

Zaborowski (2010): Andrzej Zaborowski, *Wizualizacja lasów drzew rozbioru składniowego zdań*, praca magisterska, Uniwersytet Warszawski, 2010, http://bc.klf.uw.edu.pl/222/.

Zalizniak (1977): Andrei Zalizniak, *Grammaticheskij slovar' russkogo jazyka*, wydanie 1, Russkij jazyk, Moskwa, 1977.

Żmigrodzki (2013–2018): Piotr Żmigrodzki, red., *Wielki słownik języka polskiego PAN*, Instytut Języka Polskiego PAN, 2013–2018, http://wsjp.pl.