Marcin Woliński
Warszawa

Jak się nie zgubić w lesie, czyli o wynikach analizy składniowej według gramatyki Świdzińskiego

Niniejszy tekst stanowi skromny przyczynek do lingwistyki dendrologicznej. Zdaję sobie sprawę, że lingwistycznie nie mogę zaoferować Jubilatowi nic, oprócz tego, co już ma, czyli jego własnej gramatyki formalnej w nowych szatach. Mam jednak nadzieję, że mój tekst dobrze wpisze się w badania prowadzone w zespole Jubilata dzięki opatrzeniu go dziwacznym tytułem (por. [3, 4]).

Tematem artykułu są obszerne nieraz kolekcje drzew rozbioru składniowego, jakie uzyskuje się, analizując wypowiedzenia polskie zgodnie z *Gramatyką formalną języka polskiego* ([6], dalej GFJP) autorstwa Jubilata. Opracowany przeze mnie analizator składniowy *Świgra*[1] pozwala automatycznie uzyskiwać zbiory drzew analizy zdań polskich według GFJP. Zbiory takie (nazywane w literaturze lasami analiz) są czasami zaskakująco liczne. Przedstawię ulepszoną postać prezentacyjną drzew analizy, która pozwala dość łatwo przeanalizować źródła niejednoznaczności danego wypowiedzenia.

Program *Świgra* jest dostępny na licencji GNU GPL, a więc wszyscy chętni mogą swobodnie go używać. Program i przykłady wyników analizy są dostępne w Internecie pod adresem `http://nlp.ipipan.waw.pl/~wolinski/swigra`.

1. Gramatyka formalna języka polskiego

Mimo iż GFJP zapisana jest w sposób bardzo formalny, możliwość jej realizacji komputerowej nie była celem jej Autora. Gramatyka ta jest tworem bardzo złożonym. Sądzę, że czytelnik zaznajomiony z formalizmem stosowanym przez Świdzińskiego jest w stanie osiągnąć taki poziom zrozumienia, żeby przewidywać, jak „powinny" być interpretowane w zgodzie z GFJP dane konstrukcje. Jednak stwierdzenie przez człowieka, czy faktycznie wszystkie warunki występujące w regułach są spełnione, jest praktycznie niemożliwe.

Można więc chyba powiedzieć, że GFJP, mimo że nie tworzona z myślą o maszynie, może być dogłębnie zrozumiana przez człowieka tylko z pomocą maszyny (por. [1]). Taką możliwość stwarza program *Świgra*. Program ten zawiera interpretację GFJP na tyle wierną oryginałowi, na ile tylko było to możliwe (większość wprowadzonych zmian miała charakter techniczny).

2. Gramatyka metamorficzna

Techniką opisu stosowaną przez Świdzińskiego jest analiza wypowiedzenia na składniki bezpośrednie (por.[5, rozdz. III]). Podstawowym krokiem takiej analizy jest konsta-

[1] System *Świgra* jest częścią mojej pracy doktorskiej [7], opracowanej w Instytucie Podstaw Informatyki PAN pod kierunkiem dr. hab. Janusza S. Bienia, prof. UW.

tacja, że dany fragment wypowiedzenia, interpretowany jako pewna jednostka składniowa, może być rozłożony na pewien ciąg składników (występujących w ustalonej kolejności). Krok taki można opisać regułą gramatyczną, składającą się z lewej strony, która charakteryzuje jednostkę rozkładaną, i prawej strony, którą stanowi ciąg składników tej jednostki.

Formalizm stosowany przez Świdzińskiego, zwany gramatyką metamorficzną (por. [2]), pozwala dodatkowo wyposażać jednostki składniowe w parametry, opisujące cechy gramatyczne tych jednostek.

Program komputerowy realizujący gramatykę sprawdza, czy wypowiedzenie (dane w formie napisu) jest akceptowane przez tę gramatykę, i określa jego strukturę. W wypadku niejednoznaczności interpretacji, zostają określone wszystkie struktury, mogące odpowiadać danemu wypowiedzeniu.

Oto pierwsza reguła gramatyki Świdzińskiego (reguła (w1)), głosząca, że jednostka **wypowiedzenie** może zostać rozłożona na występujące po sobie jednostki **zr** (zdanie równorzędne) i **znakkonca**[2]:

wypowiedzenie $\longrightarrow$ (w1)
zr(*Wf*, *A*, *C*, *T*, *Rl*, *O*, *Neg*, *I*, *Z***)**,
znakkonca(*Z***)**.

Jednostka **znakkonca** ma jeden parametr — parametr zależności *Z*. Jednostka **zr** ma więcej parametrów. Wszystkie jej parametry, występujące w tej regule, są zmiennymi (nazwy pisane wielką literą), tak więc w tej regule dopuszczalne jest zdanie równorzędne o dowolnych wartościach parametrów. Źródłem jedynego ograniczenia na wartości parametrów jest powtórzenie się zmiennej *Z*. Oznacza ono, że wartości zależności dwóch jednostek występujących po prawej stronie muszą być równe.

Ponadto w regułach pojawiają się elementy terminalne, czyli obiekty uznawane za elementarne dla tej gramatyki i nieopisywane w jej ramach. Takie elementy są w regułach ujmowane w nawiasy kwadratowe. Oto przykład reguły zawierającej element terminalny:

znakkonca(p**)** $\longrightarrow$ (int1)
['?'**]**.

Wyraża ona fakt, że jednostka **znakkonca** może być realizowana przez znak zapytania, reprezentowany tutaj przez prologowy symbol '?'. W regule tej wartością parametru zależności jednostki **znakkonca** jest stała p — wartość pytajna.

Oczywiście elementy terminalne są „podane" na wejście gramatyki przez jakiś mechanizm zewnętrzny względem niej. W wypadku programu *Świgra* elementami terminalnymi są wyniki interpretacji fleksyjnej słów tekstu przez analizator morfologiczny *Morfeusz* autorstwa Marcina Wolińskiego i Zygmunta Saloniego (por. [8]).

Ostatnim elementem, który może pojawić się po prawej stronie reguły, są warunki ujęte w nawiasy klamrowe. Oto przykład:

zaimrzecz(*F*, *P*, *Rl***)** $\longrightarrow$ (jel5)
[*F***]**,
{ slow(*F*, zaimrzecz, *P.Rl*) }.

Według tej reguły zaimek rzeczowny **zaimrzecz** o pewnym opisie może być realizowany przez dowolny element terminalny *F* pod warunkiem, że dla tego elementu spełniony

[2] Reguły GFJP przytaczam w wariancie notacji stosowanym w obecnie używanych realizacjach języka programowania Prolog, w którym są zapisywane wyniki programu *Świgra*. Różni się on nieco od notacji stosowaniej oryginalnie w książce [6].

jest warunek slow(*F*, zaimrzecz, *P.Rl*), czyli jeżeli słowo *F* jest w słowniku opisane jako realizacja zaimka rzeczownego o przypadku *P* i rodzaju-liczbie *Rl*.[3]

3. Drzewa analizy

Ponieważ techniką pracy stosowaną w gramatyce metamorficznej jest analiza na składniki bezpośrednie, jej wynikiem jest drzewo składników bezpośrednich danego wypowiedzenia. Przykłady drzew analizy dla GFJP przedstawiono na rysunkach 1–4 na końcu artykułu.

Liście drzewa odpowiadają elementom terminalnym i pustym realizacjom jednostek nieterminalnych, wierzchołki wewnętrzne odpowiadają zastosowanym regułom gramatycznym. W każdym liściu drzewa znajduje się element terminalny ujęty w ramkę. Podawane jest słowo (kursywą) i identyfikator leksemu (kapitalikami). Interpretację morfosyntaktyczną można wywnioskować z opisu wierzchołka nieterminalnego znajdującego się nad danym terminalnym.

W każdym z wierzchołków wewnętrznych zapisywana jest jednostka nieterminalna stojąca po lewej stronie zastosowanej reguły wraz z przypisanymi jej wartościami parametrów. Bezpośrednimi potomkami wierzchołka są elementy prawej strony reguły. Kolejności od lewej do prawej w regule odpowiada kolejność z góry do dołu w drzewie.

Korzeń drzewa znajduje się na górze diagramu. Kreski ciągłe łączą każdy z wierzchołków wewnętrznych z jego bezpośrednimi potomkami w drzewie. Do jednostek stanowiących prawą stronę jednej reguły (współskładników) dochodzą kreski poziome wychodzące ze wspólnej kreski pionowej.

Rysunki są zapisem skróconym, pokazującym tylko istotne poziomy hierarchii. Pomijane są nierozwidlające się fragmenty gałęzi drzewa. Kreski przerywane stosowane są tam, gdzie wierzchołek jest łączony ze swoim niebezpośrednim potomkiem.

Numery reguł GFJP użytych do utworzenia poszczególnych wierzchołków drzewa podano obok opisu tych wierzchołków, po prawej stronie diagramu.

Nieustalone wartości parametrów są zapisywane w formie litery *X* i liczby (np. *X0*). Wszystkie wystąpienia tej samej wartości nieustalonej są oznaczone tak samo.

Bardziej szczegółowy opis postaci drzew można znaleźć w pracy [7].

Na rysunku 1 w korzeniu drzewa znajduje się jednostka **wypowiedzenie**. Odpowiada ona zastosowaniu przytoczonej uprzednio reguły w1. Potomkami tego wierzchołka są jednostki **zr** i **znakkonca**. Jednostka **zr** ma następujące wartości parametrów: wyróżnik fleksyjny — osobowy os, aspekt — niedokonany nd, czas — teraźniejszy ter, tryb — oznajmujący ozn, rodzaj-liczba — nieustalony-pojedyncza *X0*/poj, osoba — pierwsza 1, negacja — niezanegowana tak, inkorporacja — nieinkorporacyjna ni, zależność — niepytajna np; ostatni, pomocniczy parametr o wartości 2 nie występuje w gramatyce oryginalnej (w szczególności w przytoczonej regule w1). Zgodnie z regułą w1, wartości zależności jednostki **zr** i **znakkonca** są równe.

4. Niejednoznaczność wyników

Niejednoznaczność interpretacyjna w gramatyce metamorficznej pojawia się, gdy dana jednostka składniowa o danych wartościach parametrów i o danym składzie leksykalnym (powiedzmy ostrożniej: reprezentowana przez dany ciąg słów) może być zanalizowana na kilka sposobów.

[3] W programie *Świgra* reguła ta ma nieco inną postać ze względu na użycie analizatora morfologicznego.

Są dwa źródła niejednoznaczności. Po pierwsze może się zdarzyć, że występujące w gramatyce reguły, które mają taką samą lewą stronę i różne strony prawe, pozwalają na przypisanie temu samemu fragmentowi tekstu różnych struktur. Po drugie niejednoznaczne mogą być same dane wejściowe, czyli niejednoznaczność może mieć pochodzenie fleksyjne. Ten drugi wypadek sprowadza się do pierwszego, jeżeli pamiętamy, że jednostki nieterminalne o różnych wartościach paramterów traktujemy jako różne.

Nazwijmy wierzchołkami wyboru wierzchołki drzewa analizy zawierające jednostkę nieterminalną, którą udało się zanalizować za pomocą więcej niż jednej reguły. Zauważmy, że wszelkie niejednoznaczności przejawiają się w formie wierzchołków wyboru.

Prezentowana tutaj koncepcja prezentacji niejednoznaczności interpretacyjnych polega na oznaczeniu w drzewie wszystkich wierzchołków wyboru. Każdy z nich zawiera etykietkę postaci: ◁**2/3**▷ **132** (por rys. 2). Zapis 2/3 wskazuje, że dana interpretacja została uzyskana za pomocą drugiej z trzech reguł, które można było zastosować w tym miejscu. Dalej następuje liczba wszystkich poddrzew, jakimi można opisać jednostkę nieterminalną w danym wierzchołku, uwzględniająca wszystkie reguły możliwe do zastosowania w danym wierzchołku, a dla każdej z nich liczbę możliwości w podrzędnych wierzchołkach wyboru.

W postaci elektronicznej drzew strzałki ◁ i ▷ są elementem aktywnym. Kliknięcie jednej z nich powoduje wyświetlenie drzewa uzyskanego za pomocą odpowiednio poprzedniej lub następnej z reguł, których można użyć w tym wierzchołku wyboru. Wyświetlane jest drzewo odpowiadające wyborowi pierwszej możliwości we wszystkich wierzchołkach wyboru potomnych względem danego.

I tak, kliknięcie strzałki w lewo w opisie ◁**2/3**▷ **132** na rys. 2 spowoduje wyświetlenie drzewa z rys. 1. Kliknięcie strzałki w prawo prowadzi do drzewa z rys. 3.

Drzewa generowane przez poprzednią wersję programu *Świgra* nie zawierały takich oznaczeń. Biorąc pod uwagę, że GFJP przypisuje pewnym przykładom setki, a nawet tysiące drzew, trudno było wśród nich odnaleźć konkretną interpretację. Aktywne wierzchołki wyboru są, jak sądzę, prostym a jednocześnie wygodnym środkiem pozwalającym poruszać się w lesie analiz.

5. Analiza jednego przykładu

Rozważmy dla przykładu następujące zdanie, stanowiące w aneksie do książki Świdzińskiego ilustrację dla reguły no35:

(1) *Pytam was,* $\underbrace{\textit{jaką}}_{\textbf{fpt}}$ $\underbrace{\textit{książkę}}_{\textbf{knoink}}$ *czytano.*

Fragment *jaką książkę* ma stanowić przykład konstrukcji nominalnej z atrybutem **knoatr** opisanej w regule no35, a złożonej z frazy przymiotnikowej **fpt** *jaką* i z konstrukcji nominalnej z inkorporacją **knoink** *książkę*.

Dla tego zdania program *Świgra* generuje 132 drzewa rozbioru. Na końcu artykułu zamieszczono kilka spośród tych drzew. Zachęcam Czytelnika do zapoznania się z plikiem PDF ze wszystkimi drzewami dostępnym pod adresem `http://nlp.ipipan.waw.pl/~wolinski/swigra`.

Drzewa na rysunkach 1, 2 i 3 obrazują możliwości interpretacyjne w wierzchołku wyboru najbliższym korzenia.

Omawiane zdanie, podobnie jak wszystkie inne realizacje jednostki **wypowiedzenie**, jest analizowane jako zdanie równorzędne **zr**, po którym następuje znak końca wypowiedzenia **znakonca**. Wierzchołek wyboru znajduje się przy jednostce **wypowiedzenie**, ponieważ uzyskane tu realizacje jednostki **zr** różnią się ostatnim parametrem.

Zdanie równorzędne *Pytam was, jaką książkę czytano* może strukturalnie być zdaniem jednorodnym **zj**, szeregowym **zsz** lub prostym **zp** (por. odpowiednio rysunki 1, 2, 3):

(2) **zj**: $\underbrace{\textit{Pytam was}}_{\textbf{zp}}\ \underbrace{,}_{\textbf{spoj}}\ \underbrace{\textit{jaką książkę czytano}}_{\textbf{zj}}$

(3) **zsz**: $\underbrace{\textit{Pytam was}}_{\textbf{zj}}\ \underbrace{,}_{\textbf{spoj}}\ \underbrace{\textit{jaką książkę czytano}}_{\textbf{zp}}$

(4) **zp**: $\underbrace{\textit{Pytam was, jaką książkę czytano}}_{\textbf{ze}}$

Zaskakujące interpretacje (2) i (3) odpowiadają strukturze, jaką bylibyśmy skłonni przypisać zdaniu:

(5) *Pytam was, odpowiada mi cisza.*

Pojawiają się one, ponieważ fragment *jaką książkę czytano* może zostać zinterpretowany jako zdanie o neutralnej wartości np parametru zależności. Zdanie takie składa się mianowicie z frazy finitywnej i frazy luźnej:

(6) **ze**: $\underbrace{\textit{jaką książkę}}_{\textbf{fl}}\ \underbrace{\textit{czytano}}_{\textbf{ff}}$

Problem wyeliminowania tej interpretacji w opisie czysto składniowym jest trudny. Nie można bowiem wykluczyć zdania o postaci:

(7) **ze**: $\underbrace{\textit{czytano}}_{\textbf{ff}}\ \underbrace{\textit{jakąś godzinę}}_{\textbf{fl}}$

Tego rodzaju frazy luźne bardzo znacząco przyczyniają się do zwiększenia liczby interpretacji. Ich wyeliminowanie wymagałoby wprowadzenia warunków leksykalnych. Wydaje się, że praktycznie jedynym sensownym rodzajem fraz luźnych rzeczownikowych są określenia czasu. Niestety oprócz nazw jednostek czasu mogą one zawierać wiele innych rzeczowników.

Na rysunku 3 warto także zwrócić uwagę na dziwną interpretację następującego fragmentu:

(8) **ze**: $\underbrace{\textit{jaką}}_{\textbf{fl}}\ \underbrace{\textit{książkę czytano}}_{\textbf{zr}}$

Interpretacja ta została uzyskana za pomocą reguły e16, która służy do opisu zdań, w których fraza luźna jest wspólna dla kilku składowych zdania złożonego równorzędnie, np.:

(9) **ze**: $\underbrace{\textit{kiedy}}_{\textbf{fl}}\ \underbrace{\textit{ty tam zostałeś, a on poszedł}}_{\textbf{zr}}$

Możliwość zastosowania jej do zdania, które nie jest faktycznie równorzędnie złożone jest chyba nadmiarowa.

Oczekiwaną interpretacją badanego zdania jest (4). Niestety w wariancie przedstawionym na rysunku 3 zdanie elementarne **ze** składa się z frazy finitywnej i jednej wymaganej frazy zdaniowej , *jaką książkę czytano*. Element *was* został tu zinterpretowany jako fraza luźna (na 3 sposoby wymienione w wierzchołku wyboru przy jednostce **fl1**: wedle GFJP możemy tu mieć do czynienia z formą dopełniaczową, biernikową lub miejscownikową). Szczęśliwie wierzchołek wyboru ponad tym zdaniem elementarnym

zawiera również interpretację, w której zdanie elementarne składa się z frazy finitywnej i dwóch fraz wymaganych — nominalnej i zdaniowej (por rys. 4):

(10) **ze**: $\underbrace{Pytam}_{\textbf{ff}}\ \underbrace{was,\ jaką\ książkę\ czytano}_{\textbf{fw}\qquad\qquad\textbf{fw}}$

W obrębie zdania elementarnego realizującego frazę zdaniową możliwości interpretacyjne są następujące (w sumie 21 drzew):

(11) **ze**: $\underbrace{jaką}_{\textbf{fl}}\ \underbrace{książkę\ czytano}_{\textbf{zr}}$

(12) **ze**: $\underbrace{jaką\ książkę}_{\textbf{fl}}\ \underbrace{czytano}_{\textbf{zr}}$

(13) **ze**: $\underbrace{jaką\ książkę}_{\textbf{fl}}\ \underbrace{czytano}_{\textbf{ff}}$

(14) **ze**: $\underbrace{jaką}_{\textbf{fl}}\ \underbrace{książkę}_{\textbf{fw}}\ \underbrace{czytano}_{\textbf{ff}}$

(15) **ze**: $\underbrace{jaką\ książkę}_{\textbf{fw}}\ \underbrace{czytano}_{\textbf{ff}}$

Poszukiwana przez nas interpretacja fragmentu *jaką książkę* jako jednostki **knoink** pojawia się w wariancie interpretacyjnym (12), (13) i (15). Na rysunku 4 pokazano interpretację (15). Jak można przypuszczać, drzewo to w pełni odpowiada intencji Autora gramatyki.

6. Podsumowanie

Przedstawiona analiza jednego przykładowego zdania pokazuje ogromną nadmiarowość drzew generowanych według GFJP. Uwidocznione zostało kilka typowych źródeł nadmiarowych analiz. Część z nich (jak np. (8)) wskazuje na konieczność uzupełnienia warunków w regułach gramatycznych. Inne (np. (6)) są skutkiem ograniczeń metody opisu abstrahującej od semantyki.

Mam nadzieję, że udało mi się pokazać, że dzięki ulepszeniu sposobu prezentacji drzew analizy, stało się możliwe łatwe poruszanie się w obszernym lesie wynikowym. Może to pozwolić to na dokładną analizę niezmiernie interesującego obiektu lingwistycznego, jakim jest bez wątpienia GFJP.

Literatura

[1] Janusz Stanisław Bień, Komputerowa weryfikacja formalnej gramatyki Świdzińskiego, *Biuletyn PTJ*, LII, str. 147–164, 1997.

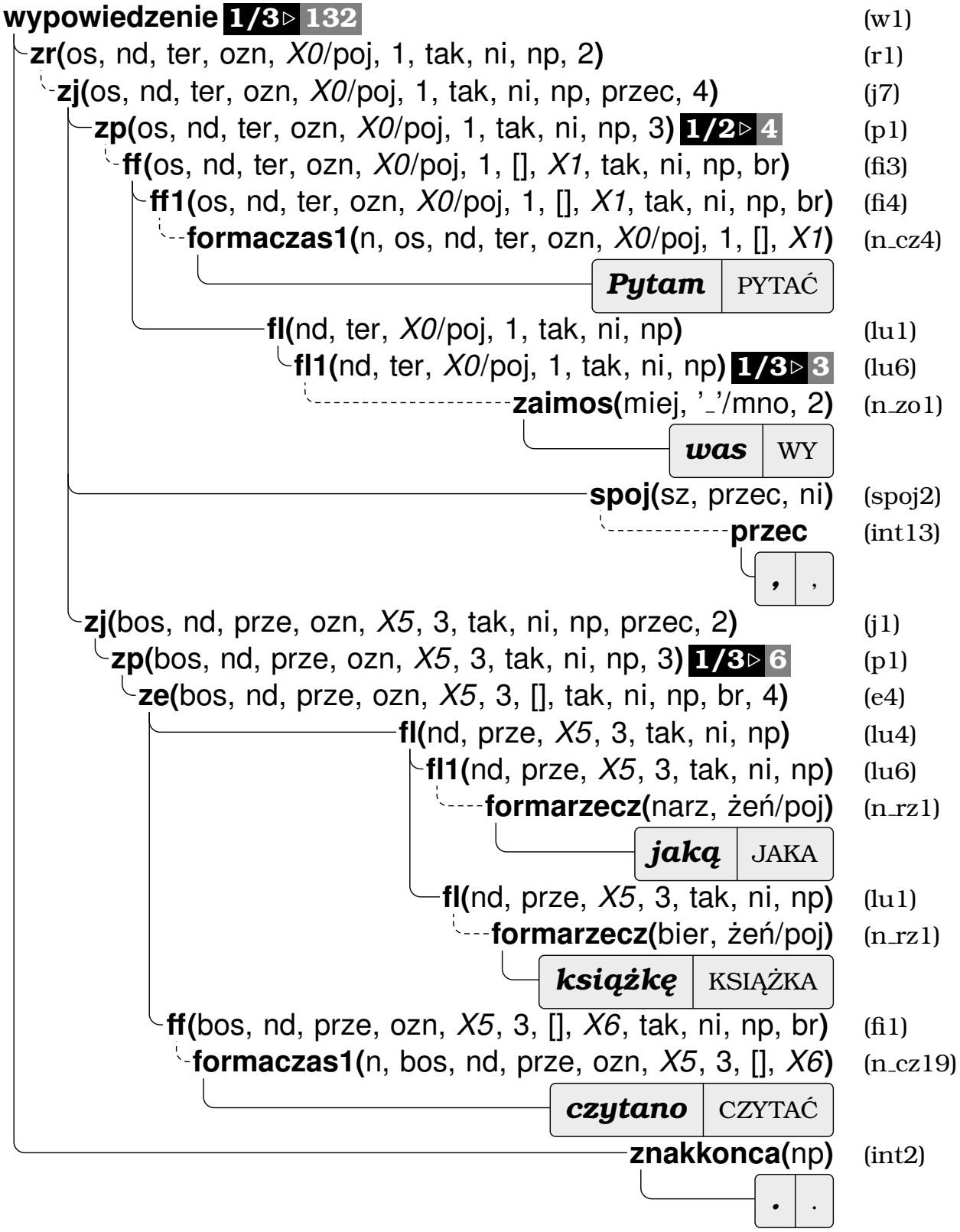
[2] Alain Colmerauer, Metamorphosis grammar, w *Natural Language Communication with Computers*, red. Leonard Bolc, Lecture Notes in Computer Science 63, str. 133–189, Springer-Verlag, 1978.

[3] 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.

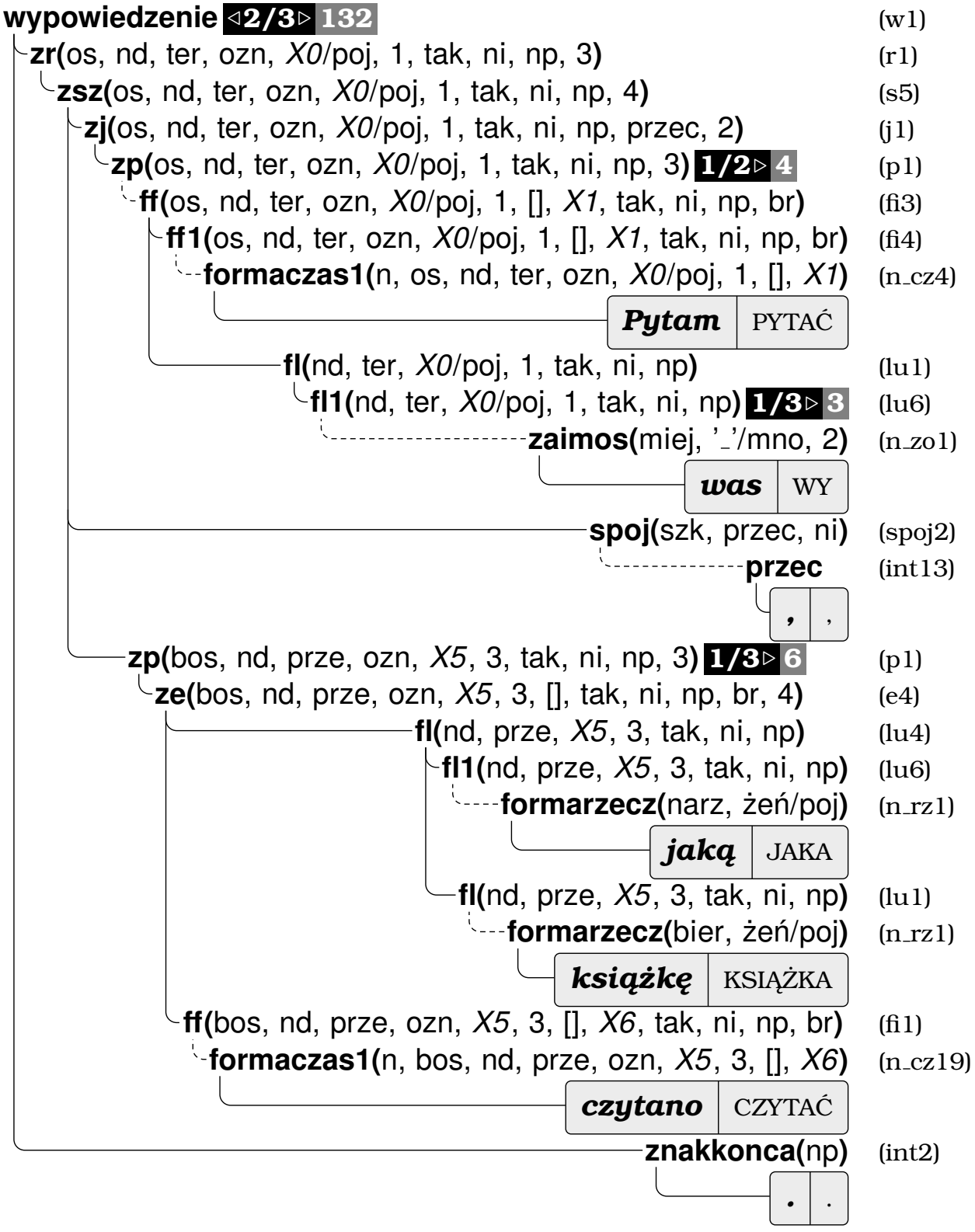
[4] Magdalena Derwojedowa i Marek Świdziński, Idiosynkrazja na przecięciu idiosynkrazyj, czyli o poprzyimkowości i liczebnikach, w *Studia z gramatyki i semantyki języka polskiego*, red. A. Moroz i M. Wiśniewski, str. 33–42, Toruń, 2004.

[5] Zygmunt Saloni i Marek Świdziński, *Składnia współczesnego języka polskiego*, Wydawnictwo Naukowe PWN, Warszawa, 5 wyd., 2001.

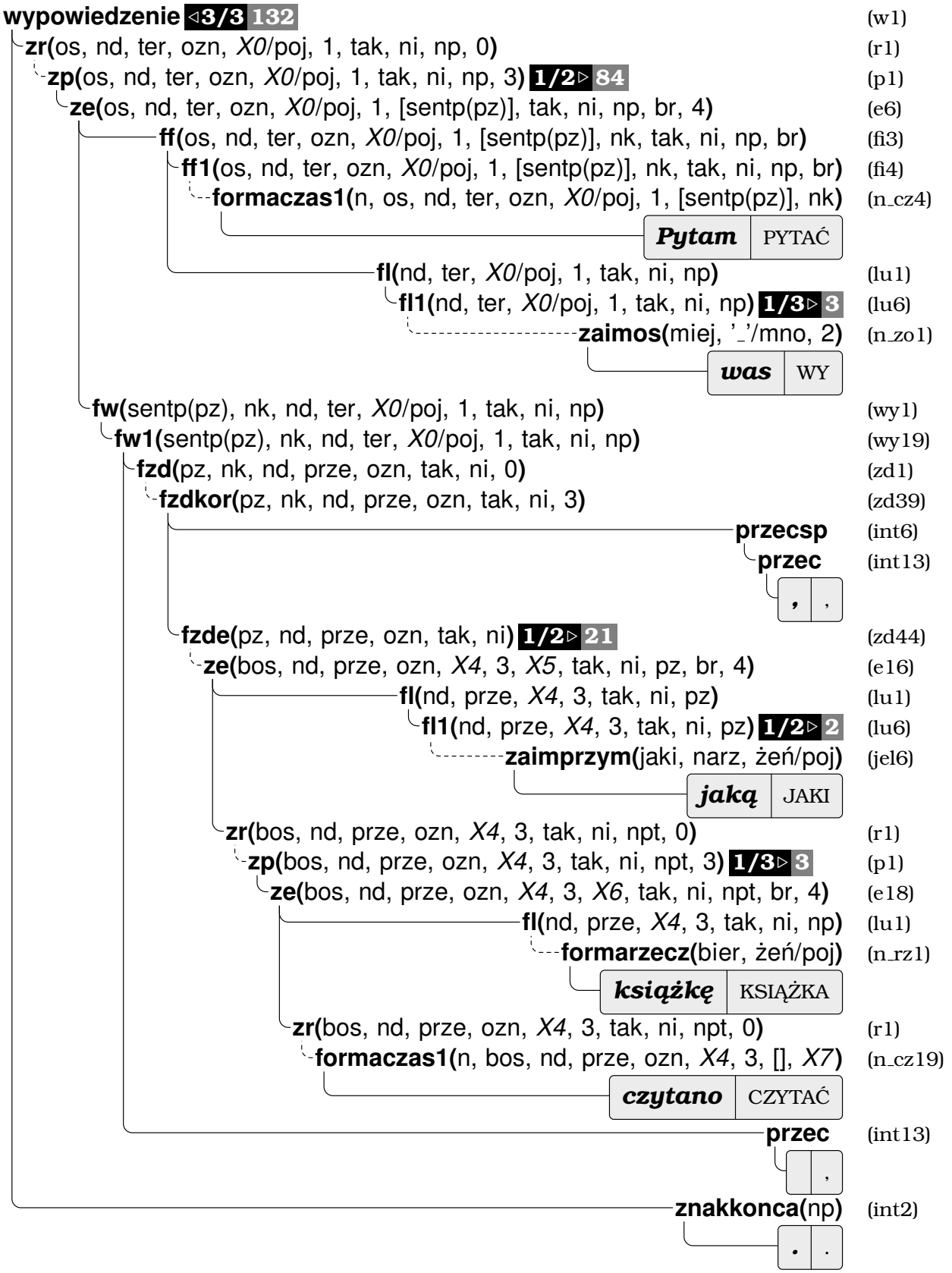
[6] Marek Świdziński, *Gramatyka formalna języka polskiego*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 1992.
[7] Marcin Woliński, *Komputerowa weryfikacja gramatyki Świdzińskiego*, Rozprawa doktorska, Instytut Podstaw Informatyki PAN, Warszawa, 2004.
[8] Marcin 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, Sławomir Wierzchoń i Krzysztof Trojanowski, str. 503–512, Springer, 2006.



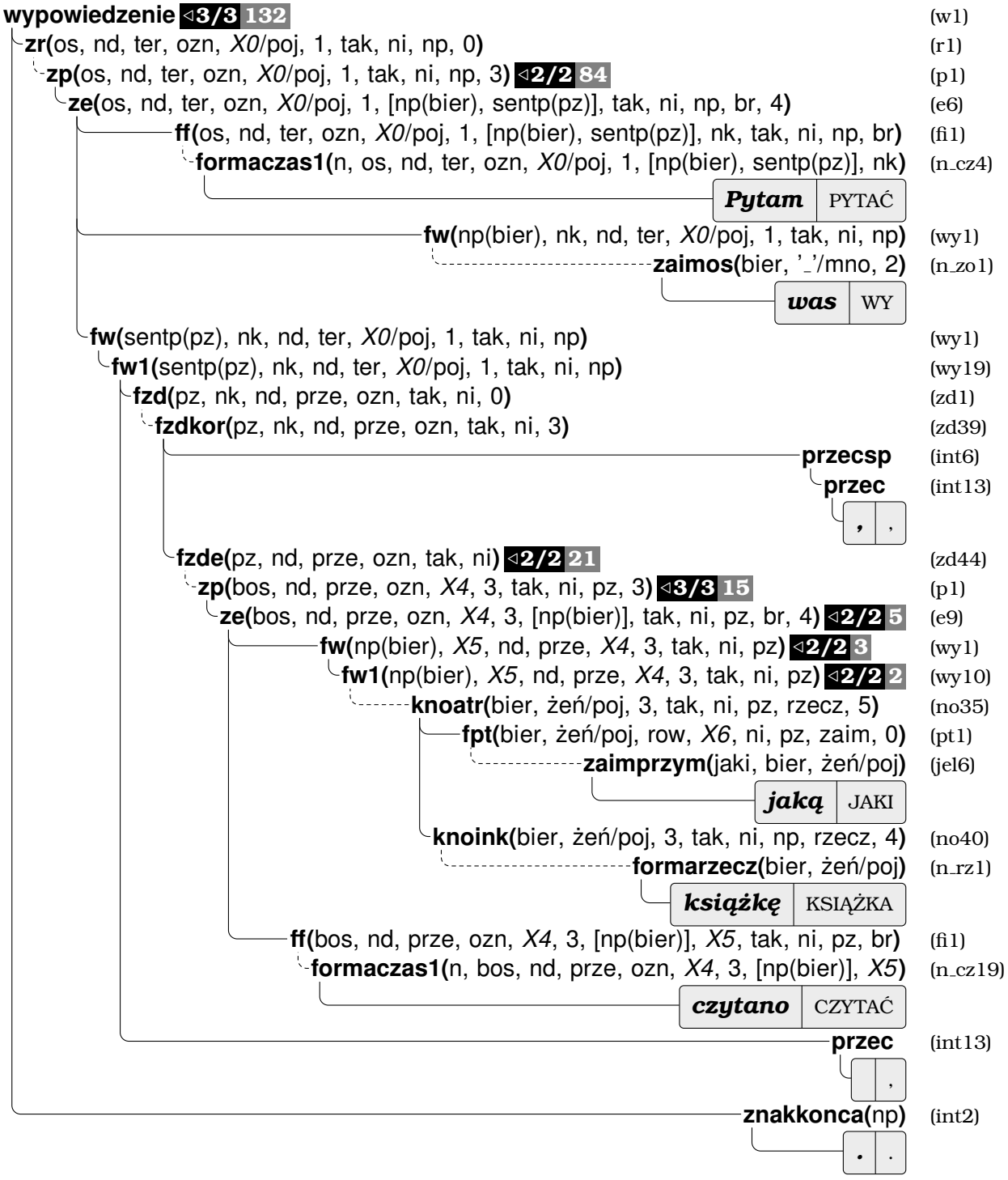
Rys. 1. Pierwsze ze 132 drzew analizy zdania *Pytam was, jaką książkę czytano.*



Rys. 2. Drzewo 25 dla tego samego przykładu



Rys. 3. Drzewo 49



Rys. 4. Drzewo 132 zawierające poszukiwaną interpretację według reguły no35