Elżbieta Hajnicz

Znakowanie semantyczne *Składnicy frazowej*

Założenia ogólne, nazwy własne, aktualizacja

Nr 1025

Warszawa, grudzień 2012

Streszczenie

Niniejszy raport omawia zasady znakowania leksykalno-semantycznego banku drzew *Składnica* jednostkami leksykalnymi pochodzącymi ze *Słowosieci*. Ponadto prezentuje metodę przeniesienia znakowania nazw własnych z NKJP do *Składnicy* (wraz z ewaluacją). Wszystkie trzy wspomniane zasoby zostały pokrótce opisane. Na koniec przedstawiona została metoda aktualizacji uzyskanego znakowania do zmian zachodzących zarówno w *Słowosieci*, jak i w *Składnicy*.

Słowa kluczowe: lingwistyka komputerowa, korpusy tekstów, banki drzew, znakowanie korpusów, semantyka leksykalna, wordnet

Abstract

Semantic Annotation of Treebank *Składnica*

Fundamentals, named entities, actualisation

The present report discusses the principles of lexical-semantic annotation of treebank *Składnica* by means of *Słowosieć (PlWordNet)* lexical units. Moreover, it presents a method of mapping NKJP named entities annotation to *Składnica* (including evaluation). All three resources mentioned above are shortly described. Finally, a method of updating the annotation to changes appearing both in *Słowosieć* and *Składnica*.

Keywords: computational linguistics, text corpora, treebanks, corpora annotation, lexical semantics, wordnet

Spis treści

1. Wstęp . . . 5

2. Korpus NKJP . . . 6

2.1. Nazwy własne w NKJP . . . 10

3. Składnica frazowa . . . 13

3.1. Informacje ogólne . . . 13

3.2. Wierzchołki . . . 14

4. Słowosieć . . . 17

4.1. Reprezentacja *Słowosieci* . . . 18

4.2. Nazwy własne w Słowosieci . . . 18

4.2.1. Lista nazw własnych . . . 19

4.2.2. Imiona i nazwiska . . . 19

4.2.3. Sztuczne synsety przymiotnikowe . . . 20

5. Reprezentacja nazw własnych w Składnicy . . . 21

5.1. Składnia xml-owa . . . 21

5.2. Podstawy automatycznej konwersji nazw własnych z NKJP do *Składnicy* . . . 23

5.2.1. Wyznaczanie wierzchołka . . . 26

5.2.2. Wyznaczanie interpretacji semantycznej . . . 27

5.2.3. Dyzambiguacja . . . 32

5.3. Ręczna korekta znakowania nazw własnych . . . 34

5.4. Ocena automatycznej konwersji nazw własnych . . . 42

5.5. Źródła błędów konwersji . . . 48

6. Znakowanie semantyczne wyrazów pospolitych . . . 51

6.1. Składnia xml-owa . . . 51

6.1.1. Informacje ogólne . . . 52

6.1.2. Znakowanie segmentu . . . 52

6.2. Zasady znakowania semantycznego segmentów w *Składnicy Semantycznej* . . . 53

6.2.1. Segmenty specjalnej troski . . . 56

6.3. Elipsy . . . 64

6.4. Składnia xml-owa . . . 64

6.5. Zasady znakowania elips . . . 64

7. Przenoszenie znakowania semantycznego na nowe wersje *Składnicy* 67

8. Aktualizacja znakowania semantycznego względem kolejnych wersji *Słowosieci* . . . 69

8.1. Identyfikacja zmian wprowadzonych w *Słowosieci* . . . 69

8.2. Przeniesienie zmian wykrytych w *Słowosieci* do tabeli nazw . . . 71

8.3. Wprowadzenie zmian wykrytych w *Słowosieci* do *Składnicy* . . . 73

8.3.1. Analiza wyników . . . 75

9. Podsumowanie . . . 77

1. Wstęp

Ręczne tworzenie zasobów językowych, choć pracochłonne, jest niezmiernie cenne. Nie tylko służy do trenowania i weryfikacji metod automatycznego ich tworzenia, lecz stanowi najbardziej wiarygodne źródło wiedzy dla użytkowników. Podstawowym rodzajem takich zasobów są korpusy tekstów, a ważnym poziomem ich znakowania jest znakowanie semantyczne i pragmatyczne, w tym nazw własnych. Jako że poziom składniowy i semantyczny języka są w sposób dość wyraźny powiązane, wydaje się uzasadnione semantyczne znakowanie korpusów oznakowanych składniowo, czyli *banków drzew*.

Niniejszy raport opisuje zasady rozszerzania *Składnicy frazowej* (por. rozdz. 3), banku drzew zdań polskich, o znakowanie semantyczne wyrazów występujących w zdaniach zgromadzonych w *Składnicy* za pomocą jednostek leksykalnych polskiego wordnetu zwanego *Słowosiecią* (por. rozdz. 4). Znakowaniu podlegają wyłącznie zdania, dla których udało się znaleźć poprawne drzewo rozbioru.

Proces właściwego znakowania semantycznego *Składnicy* poprzedzony został oznaczeniem w drzewach nazw własnych. Taka potrzeba wynika z odrębności nazw własnych, które nie posiadają właściwej interpretacji semantycznej i w związku z tym są reprezentowane w *Słowosieci* w stopniu ograniczonym. Dodatkową motywacją była możliwość wykorzystania istniejącego znakowania nazw własnych w milionowym podkorpusie NKJP (por. punkt 2.1).

Zarówno *Słowosieć*, jak i *Składnica* są zasobami podlegającymi intensywnemu rozwojowi. Nie jest to bynajmniej rozwój przyrostowy, w obu zasobach korygowane są wykryte błędy, zmienia się koncepcja reprezentacji. Jednostki leksykalne w *Słowosieci* mogą być usuwane lub przesuwane do innych synsetów, synsety mogą być dzielone lub usuwane. Zmianom podlegają też łączące je relacje. Rozwijana i modyfikowana jest gramatyka GFJP (i wykorzystujący ją parser *Świgra*) będąca podstawą *Składnicy*. W rezultacie zestaw drzew rozbioru danego zdania generowany przez *Świgrę* może ulec zmianie; dotyczy to w szczególności drzewa wybranego jako poprawne. Zmiany te mogą i będą występować w czasie realizacji znakowania semantycznego. Niezbędne więc stało się opracowanie metody modyfikowania zrealizowanego już znakowania w sposób zachowujący spójność z kolejnymi wersjami *Świgry* i *Słowosieci* przy jednoczesnym ograniczeniu ingerencji ludzkiej do niezbędnego minimum.

Aktualna wersja *Składnicy frazowej* składa się z akapitów zawierających 20 tys. zdań, z czego 8210 posiada poprawne drzewo rozbioru. W podkorpusie milionowym NKJP zaznaczone zostało 75359 nazw własnych, z czego 4473 znajduje się w zdaniach mających poprawne drzewo rozbioru w *Składnicy* (jest 2595 takich zdań).

Składnica semantyczna będąca rezultatem znakowania ma być podstawą do wspomaganego automatycznie tworzenia semantycznego słownika walencyjnego (Hajnicz, 2011). Fakt ręcznego wyznaczania zarówno drzew rozbioru, jak i słowosieciowych interpretacji semantycznych uwierzytelnia jakość tego wspomagania.

2. Korpus NKJP

Narodowy Korpus Języka Polskiego (NKJP Przepiórkowski *et al.*, 2012) jest obecnie największym korpusem języka polskiego, posiadającym najwyższe standardy znakowania realizowanego na wielu niezależnych poziomach. NKJP jako całość zawiera ok. 1,8 mld. segmentów, jego podkorpus zrównoważony zawiera ok 300 mln. segmentów, jest to więc korpus naprawdę duży.

Z punktu widzenia niniejszego opracowania najważniejszy jest zrównoważony podkorpus zawierający ok. miliona tradycyjnie rozumianych słów, zwany w związku z tym (pod)korpusem milionowym (Przepiórkowski *et al.*, 2012, s. 54). Składa się on z próbek zawierających od 40 do 70 słów. Każda próbka obejmuje (przynajmniej) jeden pełny akapit (korpus w rezultacie składa się z 18473 akapitów). Korpus ten został ręcznie oznakowany.

Korpusy przechowywane są w sposób jednorodny w postaci drzewa katalogów, w którego liściach znajdują się poszczególne dokumenty (lub ich „agregaty"). Jeżeli więc do korpusu milionowego zostało wylosowane kilka akapitów z tego samego dokumentu, są one przechowywane łącznie. W wypadku korpusu milionowego struktura drzewa katalogów jest płaska.

Ręczne znakowanie korpusu odbywało się na kilku poziomach: morfosyntaktycznym, powierzchniowym składniowym, semantycznym (dla ograniczonego zestawu wyrazów wieloznacznych) i nazw własnych (por. punkt 2.1). Każdy poziom znakowania przechowywany jest w odrębnym pliku XML. Sposób znakowania jest zgodny ze standardami TEI P5 (Przepiórkowski i Bański, 2009). Tak więc, metadane znajdują się w pliku `header.xml`, tekst źródłowy w pliku `text.xml`, plik `ann_segmentation.xml` zawiera podział na akapity, zdania i segmenty. Następnie w pliku `ann_morphosyntax.xml` przechowywane jest znakowanie morfosyntaktyczne, w pliku `ann_senses.xml` znakowanie semantyczne, a w pliku `ann_named.xml` identyfikacja nazw własnych. Natomiast znakowanie syntaktyczne (Głowińska i Przepiórkowski, 2010) ma charakter powierzchniowy i obejmuje dwa poziomy: wyrazów syntaktycznych i grup (fraz) syntaktycznych, w związku z czym przechowywane jest na dwóch plikach: `ann_words.xml` i `ann_groups.xml`, odpowiednio.

Tagset zastosowany do znakowania morfosyntaktycznego niewiele różni się od wykorzystanego w Korpusie IPI PAN (por. Przepiórkowski, 2009). Na rys. 1 przedstawiony jest fragment znakowania dla zdania (1) zawierający frazę *ulic Gdyńskiej i Piaskowej*. Jak widać, sposób znakowania jest bardzo rozbudowany; rysunek zawiera jedynie ciąg czterech segmentów. W szczególności, przechowywane są wszystkie możliwe lematyzacje, a dla każdej z nich wszystkie interpretacje morfosyntaktyczne segmentów znane analizatorowi morfosyntaktycznemu[1].

[1] Do znakowania NKJP wykorzystany został analizator morfosyntaktyczny *Morfeusz* (Woliński, 2006).

```
<seg corresp="ann_segmentation.xml#segm_1.22-seg" xml:id="morph_1.22-seg">
 <fs type="morph">
  <f name="orth">
   <string>ulic</string>
  </f>
  <!-- ulic [143,4] -->
  <f name="interps">
   <fs type="lex" xml:id="morph_1.22.1-lex">
    <f name="base">
     <string>ulica</string>
    </f>
    <f name="ctag">
     <symbol value="subst"/>
    </f>
    <f name="msd">
     <symbol value="pl:gen:f" xml:id="morph_1.22.1.1-msd"/>
    </f>
   </fs>
  </f>
  <f name="disamb">
   <fs feats="#an8003" type="tool_report">
    <f fVal="#morph_1.22.1.1-msd" name="choice"/>
    <f name="interpretation">
     <string>ulica:subst:pl:gen:f</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
<seg corresp="ann_segmentation.xml#segm_1.23-seg" xml:id="morph_1.23-seg">
 <fs type="morph">
  <f name="orth">
   <string>Gdyńskiej</string>
  </f>
  <!-- Gdyńskiej [148,9] -->
  <f name="interps">
   <fs type="lex" xml:id="morph_1.23.1-lex">
    <f name="base">
     <string>gdyński</string>
    </f>
    <f name="ctag">
     <symbol value="adj"/>
    </f>
    <f name="msd">
     <vAlt>
      <symbol value="sg:gen:f:pos" xml:id="morph_1.23.1.1-msd"/>
      <symbol value="sg:dat:f:pos" xml:id="morph_1.23.1.2-msd"/>
      <symbol value="sg:loc:f:pos" xml:id="morph_1.23.1.3-msd"/>
     </vAlt>
    </f>
   </fs>
   <fs type="lex" xml:id="morph_1.23.2-lex">
    <f name="base">
     <string>Gdyński</string>
    </f>
    <f name="ctag">
     <symbol value="adj"/>
    </f>
```

```
    <f name="msd">
     <symbol nkjp:manual="true" value="sg:gen:f:pos" xml:id="morph_1.23.2.1-msd"/>
    </f>
   </fs>
  </f>
  <f name="disamb">
   <fs feats="#an8003" type="tool_report">
    <f fVal="#morph_1.23.2.1-msd" name="choice"/>
    <f name="interpretation">
     <string>Gdyński:adj:sg:gen:f:pos</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
<seg corresp="ann_segmentation.xml#segm_1.24-seg" xml:id="morph_1.24-seg">
 <fs type="morph">
  <f name="orth">
   <string>i</string>
  </f>
  <!-- i [158,1] -->
  <f name="interps">
   <fs type="lex" xml:id="morph_1.24.1-lex">
    <f name="base">
     <string>i</string>
    </f>
    <f name="ctag">
     <symbol value="conj"/>
    </f>
    <f name="msd">
     <symbol value="" xml:id="morph_1.24.1.1-msd"/>
    </f>
   </fs>
   <fs type="lex" xml:id="morph_1.24.2-lex">
    <f name="base">
     <string>i</string>
    </f>
    <f name="ctag">
     <symbol value="interj"/>
    </f>
    <f name="msd">
     <symbol value="" xml:id="morph_1.24.2.1-msd"/>
    </f>
   </fs>
   <fs type="lex" xml:id="morph_1.24.3-lex">
    <f name="base">
     <string>i</string>
    </f>
    <f name="ctag">
     <symbol value="qub"/>
    </f>
    <f name="msd">
     <symbol value="" xml:id="morph_1.24.3.1-msd"/>
    </f>
   </fs>
  </f>
```

```
  <f name="disamb">
   <fs feats="#an8003" type="tool_report">
    <f fVal="#morph_1.24.1.1-msd" name="choice"/>
    <f name="interpretation">
     <string>i:conj</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
<seg corresp="ann_segmentation.xml#segm_1.25-seg" xml:id="morph_1.25-seg">
 <fs type="morph">
  <f name="orth">
   <string>Piaskowej</string>
  </f>
  <!-- Piaskowej [160,9] -->
  <f name="interps">
   <fs type="lex" xml:id="morph_1.25.1-lex">
    <f name="base">
     <string>piaskowy</string>
    </f>
    <f name="ctag">
     <symbol value="adj"/>
    </f>
    <f name="msd">
     <vAlt>
      <symbol value="sg:gen:f:pos" xml:id="morph_1.25.1.1-msd"/>
      <symbol value="sg:dat:f:pos" xml:id="morph_1.25.1.2-msd"/>
      <symbol value="sg:loc:f:pos" xml:id="morph_1.25.1.3-msd"/>
     </vAlt>
    </f>
   </fs>
   <fs type="lex" xml:id="morph_1.25.2-lex">
    <f name="base">
     <string>Piaskowy</string>
    </f>
    <f name="ctag">
     <symbol value="adj"/>
    </f>
    <f name="msd">
     <symbol nkjp:manual="true" value="sg:gen:f:pos" xml:id="morph_1.25.2.1-msd"/>
    </f>
   </fs>
  </f>
  <f name="disamb">
   <fs feats="#an8003" type="tool_report">
    <f fVal="#morph_1.25.2.1-msd" name="choice"/>
    <f name="interpretation">
     <string>Piaskowy:adj:sg:gen:f:pos</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
```

Rysunek 1. Fragment reprezentacji zdania w notacji NKJP

Ponadto w wypadku korpusu milionowego anatatorzy mieli możliwość dodania własnych interpretacji morfosyntaktycznych (i własnych lematów), co było oznaczane przez `nkjp:manual="true"`. Każdy segment ma przyporządkowany identyfikator rozpoczynający się od napisu `morph` wskazującego, że mamy do czynienia z morfosyntaktycznym poziomem znakowania, zaś kończy się napisem `seg` wskazującym, że jest to identyfikator segmentu. Właściwą część identyfikatora stanowią dwie liczby rozdzielone kropką. Pierwsza z nich determinuje numer akapitu, a druga kolejny numer segmentu w ramach akapitu.

(1) *Pierwsza z wymienionych sygnalizacji pojawi się u zbiegu ulic Gdyńskiej i Piaskowej w Koziegłowach.*
(2) *Kolejna sygnalizacja pojawi się u zbiegu ulic Okrężnej i Gdyńskiej i ułatwi wyjazd z osiedla 40-lecia w Czerwonaku.*

Pozostałe poziomy znakowania są zupełną nowością na gruncie polskiej lingwistyki korpusowej. Znakowanie na danym poziomie odwołuje się do znakowania na poziomie niższym za pomocą atrybutu `corresp`. Znakowanie na poziomie morfosyntaktycznym zawiera odwołania do pliku `ann_segmentation.xml`, zaś pozostałe poziomy znakowania odwołują się do pliku `ann_morphosyntax.xml`. Jedynym wyjątkiem jest tu znakowanie syntaktyczne na poziomie grup, które zawiera odwołania do pliku `ann_words.xml`.

2.1. Nazwy własne w NKJP

Skupimy się teraz na najbardziej dla nas interesującym poziomie znakowania, czyli nazwach własnych. W NKJP zostały uwzględnione następujace typy jednostek nazewniczych.
— geogName,
— orgName,
— persName (surname, firstname, addName),
— placeName (bloc, country, district, region, settlement),
— time,
— date.

Wszelkie inne nazwy, na przykład nazwy wytworów (marka samochodu *Golf*, system operacyjny *Linux*), chorób (*Dawn*), zwierząt (kot *Dratewka*) itp. zostały zignorowane.

Reprezentacja nazw własnych z powyższego zdania znajduje się na rys. 2. Na tym poziomie notacji pojedyncze segmenty służą do opisu poszczególnych nazw własnych. Dla każdej nazwy podawany jest jej typ (opcjonalnie także podtyp), postać ortograficzna oraz lemat. Bardzo ważna jest lista elementów o nazwie `ptr` będąca listą składników danej nazwy. Każdy taki element posiada

```
<s xml:id="named_1.28-s" corresp="ann_morphosyntax.xml#morph_1.28-s">
  <seg xml:id="named_1.28-s_n4">
    <fs type="named">
      <f name="type">
        <symbol value="placeName"/>
      </f>
      <f name="subtype">
        <symbol value="settlement"/>
      </f>
      <f name="orth">
        <string>Koziegłowach</string>
      </f>
      <f name="base">
        <string>Koziegłowy</string>
      </f>
      <f name="certainty">
        <symbol value="high"/>
      </f>
    </fs>
    <ptr target="ann_morphosyntax.xml#morph_1.27-seg"/>
  </seg>
  <seg xml:id="named_1.28-s_n3">
    <fs type="named">
      <f name="type">
        <symbol value="geogName"/>
      </f>
      <f name="orth">
        <string>Piaskowej</string>
      </f>
      <f name="base">
        <string>ulica Piaskowa</string>
      </f>
      <f name="certainty">
        <symbol value="high"/>
      </f>
    </fs>
    <ptr target="ann_morphosyntax.xml#morph_1.25-seg"/>
    <ptr target="ann_morphosyntax.xml#morph_1.22-seg"/>
  </seg>
```

```
  <seg xml:id="named_1.28-s_n2">
    <fs type="named">
      <f name="type">
        <symbol value="geogName"/>
      </f>
      <f name="orth">
        <string>ulic Gdyńskiej</string>
      </f>
      <f name="base">
        <string>ulica Gdyńska</string>
      </f>
      <f name="certainty">
        <symbol value="high"/>
      </f>
    </fs>
    <ptr target="ann_morphosyntax.xml#morph_1.22-seg"/>
    <ptr target="named_1.28-s_n1"/>
  </seg>
  <seg xml:id="named_1.28-s_n1">
    <fs type="named">
      <f name="derived">
        <fs type="derivation">
          <f name="derivType">
            <symbol value="relAdj"/>
          </f>
          <f name="derivedFrom">
            <string>Gdynia</string>
          </f>
        </fs>
      </f>
      <f name="type">
        <symbol value="placeName"/>
      </f>
      <f name="subtype">
        <symbol value="settlement"/>
      </f>
      <f name="orth">
        <string>Gdyńskiej</string>
      </f>
      <f name="base">
        <string>gdyński</string>
      </f>
      <f name="certainty">
        <symbol value="high"/>
      </f>
    </fs>
    <ptr target="ann_morphosyntax.xml#morph_1.23-seg"/>
  </seg>
</s>
```

Rysunek 2. Reprezentacja nazw własnych w notacji NKJP

atrybut `target` wiążący nazwę z segmentem z pliku `ann_morphosyntax.xml` bądź z inną nazwą. Tak więc reprezentacja ta ma charakter rekurencyjny.

Zwróćmy jeszcze uwagę na dwa fakty. Po pierwsze, w NKJP jako nazwy własne reprezentowane są przymiotniki pochodne od zwykłych, rzeczownikowych nazw własnych. Jako przykład niech posłuży przymiotnik *gdyński* jako pochodny od nazwy miasta *Gdynia*. Informacja, że to przymiotnik, zaznaczana jest poprzez `<f name="derived"> ... <f name="derivType"> <symbol value="relAdj"/> </f> ... </f>`. Po drugie, reprezentowanie nazw jako list przynależnych do nich segmentów umożliwia reprezentację nieciągłości i częściowego zachodzenia nazw na siebie (por. Przepiórkowski *et al.*, 2012, s. 185). W naszym przykładzie nazwy *ulica Piaskowa* i *ulica Gdyńska* mają wspólną część *ulica*, a druga z nich jest nieciągła.

Podobny sposób reprezentacji przewidziano dla nazw mieszkańców państw, miejscowości i regionów (*Europejczyk, Polak, Mazowszanin, Warszawiak, Żoliborzanin*). W tym wypadku typ derywacyjny `derivType` ma wartość `persDeriv`.

3. Składnica frazowa

Składnica frazowa utworzona została na podstawie ręcznie znakowanego podkorpusu NKJP omówionego powyżej. Zasób ten wciąż jest rozwijany, i na razie obejmuje tylko część korpusu. Zdania wchodzące w jego skład są wpierw parsowane przez parser *Świgra* (Woliński, 2004), a następnie lingwiści zwani „dendrologami" wyszukują wśród wygenerowanych przezeń drzew poprawne za pomocą programu *Dendrarium*. Każde zdanie jest analizowane przez dwóch dendrologów, a w wypadku niezgodności trafia do weryfikacji przez „superdendrologa". Do pierwszej wersji *Składnicy* wylosowano ok. 20 tys. zdań z korpusu milionowego, ignorując przy tym transkrypcje tekstów mówionych oraz fora internetowe (Woliński, 2011). Dla 11 535 (57,7%) parser wygenerował zbiór drzew rozbioru (przechowywany w postaci upakowanego lasu), a dla 7841 z nich „dendrolodzy" znaleźli poprawne drzewo rozbioru.

3.1. Informacje ogólne

Elementem głównym pliku XML jest w *Składnicy frazowej* element `forest`. Element ten zawiera następujących potomków:

— `text`, którego wartością jest tekstowa postać zdania,
— `startnode`,
— `stats` zawierający dane o przebiegu procesu parsowania,
— `answer-data` zawierający opis wyniku doboru drzewa,
— `node` zawierający poszczególne wierzchołki drzew rozbioru.

Element `answer-data` posiada jednego potomka `base-answer` oraz dwóch potomków `extra-answer`. Każdy z nich posiada dwa atrybuty `type` i `username` oraz potomka tekstowego `comment`. Atrybut `type` posiada następujące wartości:
— FULL oznaczający zdanie posiadające poprawne drzewo rozbioru,
— MORPH oznaczający wystąpienie błędnego znakowania morfosyntaktycznego w zdaniu,
— NOT_SENTENCE oznaczający, że wypowiedzenie jest równoważnikiem zdania,
— NO_TREE oznaczający brak drzewa rozbioru,
— TOO_DIFFICULT oznaczający zdanie zbyt trudne dla parsera,
— WRONG_SENTENCE oznaczający zdanie niepoprawne składniowo.
Jeśli dwóch „dendrologów" dokonuje identycznej oceny zdania, a wypadku odpowiedzi „FULL" wybiera te same drzewa (te same wierzchołki mają przyporządkowany atrybut `chosen="true"`, patrz poniżej, element `base-answer` jest generowany automatycznie i dziedziczy atrybut `type`.

3.2. Wierzchołki

Pojedynczy wierzchołek drzewa jest reprezentowany za pomocą elementu `node`. Zgodnie z przyjętymi zasadami, w *Składnicy* przechowywany jest cały upakowany las. Oznacza to, że nie są tworzone duplikaty wierzchołków (terminalnych i nieterminalnych) drzew, za to pojedynczy wierzchołek może przynależeć do wielu drzew (może być potomkiem kilku różnych wierzchołków). Poprawne drzewo jest oznakowane za pomocą atrybutu `chosen` o wartości `true`. Atrybut ten przypisywany jest wierzchołkom, a w wypadku nieterminali także listom ich potomków (element `children`). Wynika to z faktu, że jeden nieterminal może mieć przypisanych kilka list potomków.

Na rys. 3 widnieje fragment reprezentacji zdania (3) dotyczący frazy *prawem fizyki* w formacie *Składnicy frazowej*.

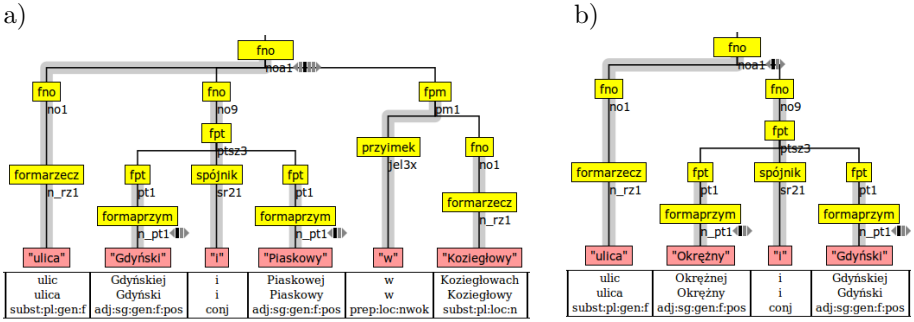
Zwróćmy uwagę, że element `child` będący potomkiem elementu `children` może być opatrzony atrybutem `head="true"`. Jest tak niezależnie od faktu, że *Składnica frazowa* ma reprezentację składnikową, a parser Świgra oparty jest na gramatyce DCG (Pereira i Warren, 1980), w której centrum frazy nie jest zaznaczane. Twórcom *Składnicy* zależało bowiem na możliwie jak największym uniezależnieniu tworzonego zasobu od konkretnego formalizmu. Dodanie informacji o centrum umożliwia m.in. przekształcenie wyników na formalizm zależnościowy (Wróblewska i Woliński, 2011).

Na rys. 4 a) przedstawione jest (pod)drzewo rozbioru frazy *ulic Gdyńskiej i Piaskowej w Koziegłowach*. Jak widać, ani *ulica Gdyńska*, ani tym bardziej *ulica Piaskowa* nie posiadają w rozbiorze tego zdania odpowiadającej im frazy. Także napis *ulic Gdyńskiej i Piaskowej* nie stanowi odrębnej frazy w tej

```
<forest sent_id="NKJP_1M_1305000000420/morph_1-p/morph_1.28-s" grammar_no="1305126214">
    <text>Pierwsza z wymienionych sygnalizacji pojawi się u zbiegu ulic
          Gdyńskiej i Piaskowej w Koziegłowach.</text>
    <startnode from="0" to="15">wypowiedzenie</startnode>
    <answer-data>
        <base-answer type="FULL" username="none">
            <comment>AUTO</comment>
        </base-answer>
    </answer-data>
    ....................................
    <node nid="70" from="7" to="9" subtrees="1" chosen="true">
        <nonterminal>
            <category>fno</category>
            <f type="przypadek">narz</f>
            <f type="rodzaj">nij</f>
            <f type="liczba">poj</f>
            <f type="osoba">3</f>
            <f type="rekcja">[]</f>
            <f type="klasa">rzecz</f>
            <f type="zap">bzap</f>
            <f type="poz">_</f>
            <f type="neg">tak</f>
            <f type="dest">neut</f>
            <f type="ink">ni</f>
        </nonterminal>
        <children rule="noa1" chosen="true">
            <child nid="71" from="7" to="8" head="true"/>
            <child nid="74" from="8" to="9" head="false"/>
        </children>
    </node>
    <node nid="71" from="7" to="8" subtrees="1" chosen="true">
        <nonterminal>
            <category>fno</category>
            <f type="przypadek">narz</f>
            <f type="rodzaj">nij</f>
            <f type="liczba">poj</f>
            <f type="osoba">3</f>
            <f type="rekcja">[]</f>
            <f type="klasa">rzecz</f>
            <f type="zap">bzap</f>
            <f type="poz">_</f>
            <f type="neg">tak</f>
            <f type="dest">neut</f>
            <f type="ink">ni</f>
        </nonterminal>
        <children rule="no1" chosen="true">
            <child nid="72" from="7" to="8" head="true"/>
        </children>
    </node>
    <node nid="72" from="7" to="8" subtrees="1" chosen="true">
        <nonterminal>
            <category>formarzecz</category>
            <f type="przypadek">narz</f>
            <f type="rodzaj">nij</f>
            <f type="liczba">poj</f>
            <f type="rekcja">[]</f>
        </nonterminal>
        <children rule="n_rz1" chosen="true">
            <child nid="73" from="7" to="8" head="true"/>
        </children>
    </node>
```

```
<node nid="73" from="7" to="8" subtrees="1" chosen="true">
    <terminal token_id="morph_2.49-seg" interp_id="morph_2.49.1.1-msd"
              disamb="true" nps="false">
        <orth>prawem</orth>
        <base>prawo</base>
        <f type="tag">subst:sg:inst:n</f>
    </terminal>
</node>
<node nid="74" from="8" to="9" subtrees="1" chosen="true">
    <nonterminal>
        <category>fno</category>
        <f type="przypadek">dop</f>
        <f type="rodzaj">żeń</f>
        <f type="liczba">poj</f>
        <f type="osoba">3</f>
        <f type="rekcja">[]</f>
        ......................
    </nonterminal>
    <children rule="no1" chosen="true">
        <child nid="75" from="8" to="9" head="true"/>
    </children>
</node>
<node nid="75" from="8" to="9" subtrees="1" chosen="true">
    <nonterminal>
        <category>formarzecz</category>
        <f type="przypadek">dop</f>
        <f type="rodzaj">żeń</f>
        <f type="liczba">poj</f>
        <f type="rekcja">[]</f>
    </nonterminal>
    <children rule="n_rz1" chosen="true">
        <child nid="76" from="8" to="9" head="true"/>
    </children>
</node>
<node nid="76" from="8" to="9" subtrees="1" chosen="true">
    <terminal token_id="morph_2.50-seg" interp_id="morph_2.50.2.1-msd"
              disamb="true" nps="false">
        <orth>fizyki</orth>
        <base>fizyka</base>
        <f type="tag">subst:sg:gen:f</f>
    </terminal>
</node>
<node nid="178" from="8" to="9" subtrees="1" chosen="false">
    <terminal token_id="morph_2.50-seg" interp_id="morph_2.50.2.3-msd"
              disamb="false" nps="false">
        <orth>fizyki</orth>
        <base>fizyka</base>
        <f type="tag">subst:pl:acc:f</f>
    </terminal>
</node>
```

Rysunek 3. Fragment reprezentacji zdania w *Składnicy frazowej*



Rysunek 4. Przykładowe poddrzewa rozbioru ze *Składnicy*

reprezentacji. Natomiast na rys. 4 b) widnieje (pod)drzewo rozbioru frazy *ulic Okrężnej i Gdyńskiej* ze zdania (2). Różnice w potraktowaniu podobnych fraz w tych dwóch rozbiorach wynikają z odmiennej budowy zdań.

Na rysunku ścieżka prowadząca od wierzchołka reprezentującego daną frazę do jej centrum składniowego zaznaczona jest pogrubiona („poszarzona").

4. Słowosieć

Słowosieć (Piasecki *et al.*, 2009) jest jednym z dwóch polskich wordnetów (drugim jest PolNet (Vetulani *et al.*, 2009)). Wordnet jest przykładem sieci semantycznej zbudowanej wokół idei *pojęcia* reprezentowanego poprzez zbiór synonimów; stąd powszechnie używany termin *synset*. Dlatego wordnet bywa niejednokrotnie postrzegany jako *leksykalna sieć semantyczna*. Najwcześniejszym a zarazem najbardziej znanym wordnetem jest bez wątpienia WordNet Prinstoński (ang. *Princeton WordNet*, PWN) (Fellbaum, 1998, `http://wordnet.princeton.edu/`).

W wypadku *Słowosieci* równie ważnym pojęciem co synset jest *jednostka leksykalna* (ang. *lexical unit*), które to pojęcie w PWN nie jest używane. Jako że wszystkie formy wyrazowe danego wzorca odmiany traktowane są jako całość, występuje tu powiązanie z pojęciem leksemu, w związku z tym jednostki leksykalne reprezentowane są w postaci trójki ⟨lemat, numer znaczenia, identyfikator numeryczny⟩. Zarówno para ⟨lemat, numer znaczenia⟩, jak i identyfikator numeryczny jednoznacznie identyfikują daną jednostkę. Pojęcie to poza wyrazami obejmuje także zleksykalizowane jednostki wieloczłonowe. Numery znaczeń (przy zadanym lemacie i kategorii gramatycznej) uporządkowane są od najczęstszych do najrzadszych na podstawie frekwencji w korpusie (począwszy od wersji 1.6).

```
124     aparycja  1                198     atrybut 3
136     apteka    1                199     atrybut 1
139     arbiter   2                18382   atrybut 2
19474   arbiter   1
```

Rysunek 5. Fragment pliku `jednostki.txt` dla rzeczowników ze *Słowosieci* 1.6

Także każdy synset posiada swój jednoznaczny identyfikator numeryczny.

Słowosieć obejmuje trzy klasy gramatyczne: rzeczowniki, czasowniki, przymiotniki; przysłówki mają zostać uwzględnione w przyszłości.

Zarówno jednostki leksykalne, jak i synsety są wiązane przez liczne relacje leksykalne. Synonimia wyrażana jest przez fakt przynależności do jednego synsetu. Podstawowymi relacjami pomiędzy synsetami są wzajemnie odwrotne hiponimia i hiperonimia oraz holonimia i meronimia. Relacjami pomiędzy jednostkami są m.in. antonimia, konwersja i derywacyjność.

4.1. Reprezentacja *Słowosieci*

Istnieją dwie zasadnicze reprezentacje *Słowosieci*: baza danych wykorzystywana przez programy służące do jej tworzenia i przeglądania oraz plik XML. Równoważna reprezentacja jest w postaci zestawu tabel, oddzielnie dla każdej z trzech rozważanych kategorii gramatycznych. Podstawowe dwa pliki to `jednostki.txt` zawierający trzykolumnową tabelę ⟨identyfikator jednostki, lemat, numer znaczenia⟩ oraz `synsety.txt` zawierający dwukolumnową tabelę ⟨identyfikator jednostki, identyfikator synsetu⟩. Pozostałe takie pliki są to `dziedziny.txt` zawierający dwukolumnową tabelę ⟨identyfikator jednostki, kategoria semantyczna⟩ oraz cały zestaw dwukolumnowych tabel relacji. Fragment pliku `jednostki.txt` widnieje na rys. 5.

4.2. Nazwy własne w Słowosieci

W *Słowosieci* została też uwzględniona, w sposób dość przypadkowy, pewna liczba nazw własnych. Są one reprezentowane za pomocą takich samych jednostek jak wyrazy pospolite. Jako że ten sam obiekt może posiadać kilka nazw choćby ze względu na występowanie skrótowców, jednostki reprezentujące nazwy własne także są grupowane w synsety. Tak więc z technicznego punktu widzenia nazwy własne są reprezentowane w *Słowosieci* w taki sam sposób jak pozostałe pojęcia. Wyróżnia je fakt, że są wiązane z reprezentowanymi przez siebie wyrazami pospolitymi za pomocą wzajemnie odwrotnych relacji *typu* i *egzemplarza* (relacje między synsetami). Jednak ponieważ relacje te zostały wprowadzone do *Słowosieci* dość późno, część wcześniej uwzględnionych nazw własnych jest wiązana ze swym pospolitym odpowiednikiem za pomocą relacji hiponimii, co wprowadza pewien bałagan.

4.2.1. Lista nazw własnych

Jako że w Słowosieci 1.0 nazwy własne reprezentowane były w stopniu nieomal śladowym, do celów automatycznego znakowania korpusu sensami wyrazów (Hajnicz, 2011) stworzyłam własny zestaw nazw. Ma on postać tabeli, w której lematom poszczególnych nazw własnych przypisywana jest lista ich typów ze *Słowosieci* w postaci liczbowych identyfikatorów odpowiednich synsetów. Podobnie jak jednostki w *Słowosieci*, są one uporządkowane wedle istotności znaczeń. Jednak przyjęty porządek jest arbitralny; nie prowadzono żadnych analiz statystycznych. Przyjęto konwencję, że miasto ma wyższy „priorytet" niż wieś (*Bolesławiec, Chełmno*) czy nazwisko (*Kalisz*), a planeta niż bóstwo (*Uran, Wenus*). Ustalenie jednak, czy „ważniejsza" jest rzeka czy jezioro (*Hańcza*, jest to także nazwisko) jest nieoczywiste.

Zestaw ten został znacząco rozbudowany na potrzeby znakowania *Składnicy* i obecnie liczy ponad 8300 lematów, z czego ok. 750 jest wieloznacznych (niecałe 10%).

4.2.2. Imiona i nazwiska

Rzecz jasna zarówno leksem imię, jak i nazwisko w *Słowosieci* występują. Wydaje się więc oczywiste, że imiona *Marek, Anna* i nazwiska *Olszewski, Wójcik* powinny mieć przypisany taki właśnie typ. Jednak oba są hiponimami nazwy własnej, a za jej pośrednictwem {leksemu, wyrazu, verbum} i znaku językowego. Jednostki te dobrze reprezentują więc użycie nazw własnych w wypowiedzeniach typu *Nazywam się Marek Olszewski*, *Mam na imię Marek*, *Mówią do niego Zdzichu*.

Rozważmy jednak zdanie *Marek Olszewski wszedł do restauracji*. W podobnym kontekście może pojawić się *Mężczyzna/Facet/Nieznajomy wszedł do restauracji*, nigdy jednak *Imię/Nazwisko weszło do restauracji* czy *Wyraz/Znak językowy wszedł do restauracji*. Tak więc z punktu widzenia testów na hiponimię/hiperonimię (por. Derwojedowa *et al.*, 2008; Piasecki *et al.*, 2009, Appendix A) nie jest to poprawna interpretacja tych nazw. Ktoś mógłby co prawda powiedzieć, że nazwy własne takim relacjom nie podlegają. W moim najgłębszym przekonaniu jednak zasady obowiązujące dla relacji typu powinny być podobne. Jako argument za takim rozwiązaniem niech posłuży fakt, że reprezentowane w *Słowosieci* Warszawa i Wrocław są typu miasto, a nie nazwa własna.

Dlatego proponuję wprowadzenie pomocniczych, sztucznych jednostek leksykalnych mężczyzna określany imieniem, kobieta określana imieniem, człowiek określany nazwiskiem oraz człowiek ze względu na miano, trzy pierwsze będące hiponimem ostatniego, ta z kolei będąca hiponimem człowiek, jednostka, osoba.

Nazwiska stanowią najliczniejszy podzbiór nazw w tabeli, jest ich 4178, więc stanowią ok 50% wszystkich nazw. Charakterystyczną cechą języka polskiego,

podobnie jak innych języków słowiańskich, jest występowanie odrębnych form żeńskich nazwisk, pochodnych od form męskich. Pierwszą grupę takich nazwisk stanowią formy żeńskie nazwisk o zakończeniach *-owa, -ówna* (*Boryczkowa, Florczakowa, Gawlikowa, Skowronówna*) pochodne od nazwisk zakończonych spółgłoskowo (*Florczak, Gawlik, Skowron*) lub na *-o* (*Boryczko*) występujące bezpośrednio po spółgłosce. Drugą grupę stanowią formy żeńskie nazwisk kończące się na *-cka,-dzka, -ska* (*Lipnicka, Gilowska*), pochodne od form męskich kończących się na *-cki,-dzki, -ski* (*Lipnicki, Gilowski*). Ze względu na regularność tworzenia takich form zrezygnowałam z osobnego reprezentowania ich w tabeli nazw. Zamiast tego stosowana jest prosta heurystyka przekształcania form żeńskich na ich męskie odpowiedniki[2]. Warto zauważyć, że formy żeńskie kończące się na *-owa, -ówna* są we współczesnej polszczyźnie w zaniku, więc nie warto inwestować w zaawansowane metody ich wykrywania. W rezultacie w tabeli nazw przechowywana jest niewielka liczba żeńskich form nazwisk, których nie obejmowały zastosowane heurystyki, a które wystąpiły w korpusie milionowym. Są to np. *Czapkowa, Fiołkowa, Guzianka, Kaczmarkowa, Kamzowa, Lichoniowa, Macioszkowa, Wacławkówna, Wiesioniowa.* Dotyczy to także żeńskich form nazwisk o odmianie przymiotnikowej (*Cicha, Dymna, Konieczna*).

4.2.3. Sztuczne synsety przymiotnikowe

Istnieje konieczność interpretacji licznych przymiotników pochodnych od nazw własnych, które w *Słowosieci* uwzględnione nie zostały (*algierski, czeczeński, europejski, hinduski, kaliningradzki, kawęczyński, legnicki, ogólnopolski, salwadorski, waszyngtoński*). Pierwotnie zamierzałam traktować takie nazwy jako egzemplarze przymiotników *kontynentalny, krajowy, narodowy, międzynarodowy, państwowy, okręgowy, dzielnicowy, miejski, wiejski* itp. Jednak po zastanowieniu uznałam, że *legnicki piekarz* to nie to samo co *miejski piekarz*, *biskup koszaliński* nie jest *biskupem miejskim*, nie mówiąc już o tym, że *naród indyjski* nie jest *narodem narodowym, państwowym* czy *krajowym.* Dlatego zdecydowałam się na wprowadzenie następujących sztucznych jednostek przymiotnikowych:

— pochodny od nazwy dzielnicy,

[2] W wersji skryptu wykorzystanej do przenoszenia nazw własnych z NKJP do *Składnicy Semantycznej* nie zostały uwzględnione przypadki zaniku *-e* przy tworzeniu takich form jak *Grabarkowa, Macioszkowa* od nazwisk *Grabarek, Macioszek* oraz zamiany miękkiej końcówki *-ć, ń, ś, ź* na *i* (*Lichoniowa, Mikuciówna* od *Lichoń, Mikuć*), lecz heurystyka ta może zostać bez trudu uzupełniona. Zwróćmy uwagę na takie nazwiska jak *Kopciówna* (od *Kopeć*), dla których należałoby zastosować obie te heurystyki. Nie została też wdrożona heurystyka tworzenia form o końcówkach *-yna/-ina, -(i)anka* dla nazwisk kończących się na *-a* oraz niektórych końcówek spółgłoskowych (*-g*), np. *Szelburżyna, Zarembina, Szelburżanka, Zarembianka* pochodnych od *Szelburg, Zaremba*; jednak w podkorpusie milionowym nie wystąpiła ani jedna forma tego typu.

— pochodny od nazwy miejscowości,
— pochodny od nazwy obszaru,
— pochodny od nazwy geograficznej.

5. Reprezentacja nazw własnych w Składnicy

Kiedy rozważamy problematykę znakowania semantycznego tekstów, bardzo istotny jest sposób potraktowania nazw własnych. Nazwy własne nie mają bowiem znaczenia w tradycyjnym rozumieniu tego pojęcia. *Warszawa, Wrocław, Paryż* nie oznaczają miast, one są miastami. Najprostszym sposobem potraktowania nazw własnych podczas znakowania semantycznego tekstu jest ich całkowite zignorowanie. Jednak takie podejście powoduje, że znacząca liczba zdań nie jest w pełni semantycznie zinterpretowana. Ma to swoje istotne konsekwencje na przykład w procesie automatycznego wykrywania preferencji selekcyjnych (Agirre i Martinez, 2001; Bergsma *et al.*, 2008; Brockmann i Lapata, 2003; Erk, 2007; Ribas, 1994) wydobywania walencji semantycznej z korpusu tekstów (Hajnicz, 2011) czy wykrywania alternacji (Lapata, 1999; McCarthy, 2001; Resnik, 1993). Można co prawda w ogóle zignorować zdania zawierające nazwy własne, jednak w ten sposób materiał językowy uległby wyraźnemu ograniczeniu.

Dlatego zdecydowałam się na oznakowanie nazw własnych występujących w *Składnicy*. Jednak znakowanie to dokonywane jest oddzielnie od właściwego znakowania semantycznego tekstu, jako proces wstępny. Tak więc lingwiści dokonujący ręcznego znakowania semantycznego *Składnicy* otrzymują zdania, w których nazwy już zostały oznakowane. Taka procedura została przyjęta z dwóch powodów. Po pierwsze, w *Słowosieci* uwzględniony jest jedynie niewielki podzbiór nazw występujących w korpusie. Po drugie, chciałam wykorzystać znakowanie nazw własnych podkorpusu milionowego NKJP, którego *Składnica* jest podzbiorem.

Nazwy własne zostały dodane do *Składnicy* w sposób półautomatyczny i są one reprezentowane nieco odmiennie niż wyrazy pospolite.

5.1. Składnia xml-owa

W celu reprezentowania nazw własnych została rozszerzona składnia plików xml-owych, w których przechowywana jest *Składnica*.

Nazwy własne reprezentowane będa jako element o nazwie `named`, którego jedynym atrybutem jest identyfikator `name_id`, a jego wartość w założeniu dziedziczona jest z NKJP. Wartość ta składa się z następujących elementów:

ustalonego prefiksu named[3], identyfikatora zdania złożonego z dwóch liczb rozdzielonych kropką: numeru akapitu i numeru zdania w akapicie wraz z sufiksem -s oraz identyfikatora nazwy w zdaniu: litery n, po której następuje numer kolejny nazwy (w kodzie szesnastkowym). Części te rozdzielone są znakiem podkreślenia (np. named_105.27-s_n6, por. rys. 6).

Element ten posiada następujących potomków: po pierwsze nambase reprezentujący lemat nazwy, następnie nkjp_type zawierający atrybuty type i subtype wskazujące na typ i podtyp nazwy przypisane im w znakowaniu NKJP. Najważniejszymi potomkami są plwn_units i plwn_types, stanowiące reprezentację nazwy w *Słowosieci*. Pierwsza z nich pojawia się w wypadku, gdy lemat nazwy (lub jej centrum, patrz poniżej) występuje w *Słowosieci*, druga z nich, gdy nazwa występuje w tabeli nazw (lub wcale). Podstawowymi atrybutami zarówno elementu plwn_units, jak i plwn_types, są part, case_agreement i polysemy. Atrybuty te wykorzystywane były przez skrypt dokonujący automatycznej konwersji nazw własnych z postaci występującej w NKJP na opartą na *Słowosieci* postać właściwą dla *Składnicy Semantycznej*, i zostaną dokładniej omówione w punkcie 5.2 opisującym tę procedurę.

Jako że przymiotniki pochodne od nazw własnych (np. *amerykański, lubuski*) znakowane były w NKJP jako nazwy własne, zostały one uwzględnione także w powyższym znakowaniu. Uzasadnieniem jest nie tylko chęć zachowania zgodności ze znakowaniem NKJP, ale także czynnik praktyczny: przymiotniki takie występuja w *Słowosieci* w stopniu równie ograniczonym co właściwe nazwy. Nazwy takie opatrzone zostały opcjonalnym atrybutem adjective o wartości true. Żeby nie mnożyć liczby atrybutów, zdecydowałam się w tej sytuacji ustawić atrybut type elementu nkjp_type na wartość typu derywacyjnego (który *de facto* jest typem nazwy), natomiast atrybut subtype pozostawić bez zmian, o ile istnieje (por. rys. 9, 10), a w przeciwnym razie przypisać mu oryginalną (a traconą) wartość atrybutu type (por. rys. 11).

Analogiczna procedura wykonywana jest dla typu derywacyjnego persDeriv („mieszkaniec").

Drugim specjalnym przypadkiem są formy żeńskie nazwisk omówione już w punkcie 4.2.2. Formy takie są lematyzowane do wersji męskiej i opatrywane atrybutem female o wartości true.

Element plwn_units posiada niepustą listę potomków o nazwie unit, zaś element plwn_types posiada niepustą listę potomków o nazwie type. Posiadają one identyfikatory w postaci atrybutów o nazwie luid i typid, odpowiednio. Identyfikatory te składają się z dwóch części rozdzielonych dywizem. Pierwszą stanowi ostatnia część identyfikatora nazwy, jednoznacznie identyfikująca ją

[3] Inne wartości tego prefiksu wynikające z różnic w metodzie znakowania nazw własnych w NKJP i *Składnicy Semantycznej* omówione zostaną w punkcie 5.3.

w ramach zdania. Druga składa się z liter `sv` lub `tv`, odpowiednio, po których następuje numer znaczenia. W wypadku elementu `unit` jest to numer znaczenia przypisany jednostce w *Słowosieci*, w przeciwnym razie liczba kolejna (por. `luid="n6-sv1"` na rys. 6 oraz `typid="n1-tv1"` na rys. 7).

Jeśli w *Słowosieci* wystepują jednostki, których lematy różnią się jedynie kasztą, element `plwn_units` będzie zawierał tylko te jednostki, dla których nastąpiło uzgodnienie kaszty (por. punkt 5.2.2). Jako że w *Słowosieci* numeracja znaczeń jest wspólna dla lematów różniących się jedynie kasztą (przykładem tego jest *kościół/Kościół*), identyfikatory elementów `unit` nie muszą być ciągłe pod względem numeracji.

Pozostałe atrybuty elementów `unit` i `type` to: `status` o wartościach `auto` (przypisany automatycznie), `modif` (zmodyfikowany ręcznie), `added` (element dodany ręcznie) oraz głównie dla przymiotników (oraz dla niektórych określeń czasu) `suggested`; `chosen` (opcjonalny) o wartościach `true, false` wskazujący rzeczywistą interpretację nazwy własnej w zdaniu (wartość `false` przypisywana jest jedynie w wypadku błędnego wyboru automatycznego) oraz, jedynie dla elementu `type`, opcjonalny `usage` o wartościach `metaph, metonymy`.

Element `type` ma dwóch potomków tekstowych: `nbase` określający lemat nazwy w tabeli oraz `plwntype` określający identyfikator synsetu *Słowosieci* będącego typem tej nazwy umieszczonym w tabeli nazw (por. punkt 4.2.1). Lista potomków tekstowych elementu `unit` jest dłuższa i zawiera: `lubase` określający lemat nazwy w *Słowosieci*, `lusense` określający numer znaczenia, `luident` określający identyfikator jednostki oraz `synset` określający identyfikator synsetu, do którego dana jednostka należy.

Odrębne przechowywanie lematu dla każdego elementu `unit/type` wynika z faktu, że lemat ten może być z różnych względów odmienny od właściwego lematu nazwy, a w szczególnych przypadkach mogą się one także różnić między sobą.

Przykład nazwy posiadającej interpretację bezpośrednio w *Słowosieci* zaprezentowany jest na rys. 6, zaś przykład nazwy własnej występującej wyłącznie w tabeli nazw widnieje na rys. 7. Są też przypadki, gdy nazwa występuje w obu zasobach (por. rys. 8).

Ze względu na zmiany w *Słowosieci*, zarówno element `unit`, jak i `type` może posiadać opcjonalny atrybut `update` wskazujący na zmiany dotyczące danej jednostki bądź synsetu (por. punkt 8.3).

5.2. Podstawy automatycznej konwersji nazw własnych z NKJP do *Składnicy*

Różnice w sposobie reprezentacji nazw własnych w NKJP i *Składnicy frazowej* są dość istotne i mają one zasadniczy wpływ na metodę i zakres konwersji. Podstawowa grupę różnic tworzy sam sposób przechowywania informacji.

```
<named name_id="named_105.27-s_n6">
    <nambase>Ministerstwo Spraw Wewnętrznych</nambase>
    <nkjp_type type="orgName"/>
    <plwn_units part="head" case_agreement="Full" polysemy="true">
        <unit luid="n6-sv1" status="auto" chosen="true">
            <lubase>ministerstwo</lubase>
            <lusense>1</lusense>
            <luident>3501</luident>
            <synset>2896</synset>
        </unit>
        <unit luid="n6-sv2">
            <lubase>ministerstwo</lubase>
            <lusense>2</lusense>
            <luident>3502</luident>
            <synset>8257</synset>
        </unit>
        <unit luid="n6-sv3">
            <lubase>ministerstwo</lubase>
            <lusense>3</lusense>
            <luident>3503</luident>
            <synset>1663</synset>
        </unit>
    </plwn_units>
</named>
```

Rysunek 6. Przykład nazwy własnej uwzględnionej w *Słowosieci*

```
<named name_id="named_86.19-s_n1">
    <nambase>UW</nambase>
    <nkjp_type type="orgName"/>
    <plwn_types part="whole" case_agreement="Full" polysemy="true">
        <type typid="n1-tv1" status="auto" chosen="false">
            <nbase>UW</nbase>
            <plwntype>3631</plwntype>
        </type>
        <type typid="n1-tv2" status="modif" chosen="true">
            <nbase>UW</nbase>
            <plwntype>7692</plwntype>
        </type>
        <type typid="n1-tv2">
            <nbase>UW</nbase>
            <plwntype>6423</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 7. Przykład nazwy własnej uwzględnionej wyłącznie w tabeli nazw

```
<named name_id="named_404.50-s_n1">
    <nambase>Wielkopolska</nambase>
    <nkjp_type type="geogName"/>
    <plwn_units part="whole" case_agreement="Full" polysemy="false">
        <unit luid="n1-sv1" status="auto" chosen="true">
            <lubase>Wielkopolska</lubase>
            <lusense>1</lusense>
            <luident>63313</luident>
            <synset>43106</synset>
        </unit>
    </plwn_units>
    <plwn_types part="whole" case_agreement="Full" polysemy="false">
        <type typid="n1-tv1" status="auto" chosen="true">
            <nbase>Wielkopolska</nbase>
            <plwntype>1397</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 8. Przykład nazwy własnej uwzględnionej w obu zbiorach danych

W NKJP pojedynczy plik XML zawiera cały dokument podzielony na akapity i zdania (w wypadku podkorpusu oznakowanego ręcznie ograniczony do wybranych akapitów), w *Składnicy* każde zdanie reprezentowane jest w oddzielnym pliku. Reprezentacja NKJP ma charakter linearny, tzn. zdanie jest ciągiem segmentów, *Składnica* zawiera drzewa rozbioru. Z drugiej strony NKJP zawiera mechanizmy reprezentacji zjawisk nieciągłych (także na poziomie nazw własnych), które w *Składnicy* są ignorowane[4].

Automatyczna konwersja nazw własnych z NKJP do *Składnicy* składa się z dwóch niezależnych faz: wyznaczenia wierzchołka w drzewie, który ma zostać oznakowany jako nazwa własna oraz przypisanie nazwie właściwej interpretacji opartej na *Słowosieci*.

5.2.1. Wyznaczanie wierzchołka

Jako że nazwy własne bardzo często są jednostkami wieloczłonowymi, których interpretacja często nie jest kompozycyjna, przyjęta została zasada, że elementy reprezentujące nazwy własne nie są dodawane do wierzchołków terminalnych, lecz do najniżej położonych wierzchołków nieterminalnych reprezentujących daną nazwę. Tak więc w wypadku jednoczłonowych nazw rzeczownikowych jest to wierzchołek kategorii formarzecz, zaś w wypadku nazw przymiotnikowych formaprzym.

Każdy wierzchołek w *Składnicy* posiada atrybuty `from` oraz `to` wskazujące granicy frazy. Jak łatwo zauważyć dla terminali, nie jest to numeracja segmentów, lecz pozycji przed (`from="0"`) i pomiędzy segmentami. Z drugiej strony każdy terminal posiada atrybut `token_id` wiążący go z reprezentowanym przezeń segmentem z pliku `ann_morphosyntax.xml`, podobnie jak nazwy własne. W rezultacie dla każdego wierzchołka można wyznaczyć pierwszy i ostatni segment frazy, którą wierzchołek ten opisuje. Wśród wierzchołków opisujących tę sama frazę (o identycznej wartości atrybutów `from` i `to`) wybierany jest wierzchołek najniżej położony w drzewie.

Z kolei dla każdej nazwy reprezentowanej w pliku `ann_named.xml` można rekurencyjnie wyznaczyć zbiór segmentów wchodzących w jej skład. Jako że segmenty te posiadają identyfikatory dające się uporządkować numerycznie, możliwe jest bezpośrednie wyznaczenie pierwszego i ostatniego elementu w tym zbiorze. Umożliwia to powiązanie każdej nazwy własnej z odpowiadającym jej wierzchołkiem drzewa rozbioru, o ile nazwa ta jest reprezentowana w drzewie jako odrębna fraza.

Z powyższego wynika jednoznacznie, że nie wszystkie nazwy własne udaje się bezpośrednio przenieść z NKJP do *Składnicy*. Spośród nazw występujących w zdaniu (2) jedynie *Koziegłowy* oraz przymiotnik *gdyński* dają się bezpośrednio reprezentować w *Składnicy*.

[4] Co czasem prowadzi do braku poprawnego drzewa rozbioru zdania.

Warto także podkreślić, że metoda ta ignoruje przypadek fraz nieciągłych. Trudno uznać to za błąd, gdyż frazy nieciągłe nie są obsługiwane przez parser *Świgra*. Jednak zdania zawierające pewnego typu nieciągłości w *Składnicy* się znajdują; dotyczy to w szczególności rozważanych powyżej przykładów. Powyższa metoda nie potrafi rzecz jasna obsłużyć takich przypadków poprawnie. W rezultacie na przykład dla zdania (2) automatyczna konwersja zostanie dokonana nie tylko dla nazw *Czerwonak* i *gdyński*, lecz także dla nazwy *ulica Gdyńska* — element `named` zostanie przypisany do wierzchołka reprezentującego frazę *ulic Okrężnej i Gdyńskiej*. Wynika to z faktu, że pierwszym segmentem zarówno frazy, jak i nazwy, jest segment *ulic*, zaś ostatnim — segment *Gdyńskiej*. Takie są właśnie konsekwencje występowania w znakowaniu podkorpusu milionowego nazw nieciągłych. Jak zobaczymy poniżej, nie jest to sytuacja całkowicie błędna. Natomiast nazwy *ulica Okrężna* oraz *osiedle 40-lecia*[5] zostaną w procesie automatycznej konwersji zignorowane i będą musiały zostać dodane ręcznie.

5.2.2. Wyznaczanie interpretacji semantycznej

Zasadniczą kwestią podczas konwersji nazw własnych z milionowego do *Składnicy* jest przypisanie nazwom ich typów semantycznych w oparciu o *Słowosieć* (por. rozdz. 4). Istnieją dwa źródła tej informacji: właściwa *Słowosieć* oraz tabela nazw (por. punkt 4.2.1). W założeniu źródła te powinny być rozłączne. Dlatego podczas automatycznej konwersji analiza tych dwóch źródeł oraz tworzenie na ich podstawie elementów `plwn_units` i `plwn_types` (por. punkt 5.1) dokonywane było całkowicie niezależnie, co mogło prowadzić do błędu w wypadku ich nierozłączności (por. punkt 5.5), gdyż w każdym z nich niezależnie heurystycznie wybierana była najbardziej prawdopodobna interpretacja nazwy.

Pierwszym krokiem w tym procesie było dopasowanie lematu nazwy z korpusu do *Słowosieci* (bądź tabeli). W procesie konwersji przyjęto założenie, że nie musi być to postać identyczna. Dopasowywanie nazwy dokonywane było na kilku poziomach. Co ważne, kolejny poziom realizowany był jedynie wówczas, gdy dopasowywanie lematów na wyższym poziomie zakończyło się porażką (modyfikowany był lemat nazwy pobrany z korpusu):

— pełna zgodność nazwy (`part="Full"`);
— usunięcie cudzysłowów (`part="Quote"`);
— zmiana kaszty (w nawiasach znajduje się odpowiednia wartość atrybutu `case_aggreement`):

[5] W zdaniu tym frazę stanowi *osiedla 40-lecia w Czerwonaku*. Co ciekawe, w znakowaniu milionowego kwestia, która z tych fraz powinna być znakowana jako nazwa własna nie jest potraktowana zbyt konsekwentnie.

  — pierwszych liter wyrazów składowych na duże (`FirstUpper`),
  — pierwszych liter wyrazów składowych na małe (`FirstLower`),
  — wszystkich liter na duże (`AllUpper`),
  — wszystkich liter na małe (`AllLower`),
  — wszystkich liter poza pierwszymi na małe (`AllLower`);
— podmiana żeńskiej formy nazwiska (`female="true"`);
— zastosowanie heurystyki centrum (`part="head"`).

Ponieważ bieżąca wersja *Świgry* nie obsługuje cudzysłowów, heurystyka usuwania cudzysłowów jest nadmiarowa. Została wpisana, ponieważ w znakowaniu milionowego cudzysłowy zostały w lematach zachowane (*Związek Zawodowy „Solidarność", NSZZ „Solidarność", „S"* itd.) i to bez ujednolicania pisowni. Heurystyka zamiany wszystkich liter poza pierwszymi na małe służy do obsługi sytuacji, w których lematy nazwy z wypowiedzeń pisanych w całości dużymi literami zostały zapisane bez zmiany kaszty (uzyskujemy zamianę *JAWORSKI* na *Jaworski*, nie *jaworski*).

Kwestia żeńskich form nazwisk została omówiona w punkcie 4.2.2. Warto zauważyć, że jeśli występują nazwiska różniące się jedynie wystąpieniem *-o* po spółgłosce (*Curył/Curyło, Gromyk/Gromyko*), forma żeńska lematyzowana jest do nazwiska kończącego się na spółgłoskę, gdyż taki test przeprowadzany jest jako pierwszy. Odwrotnie jest tylko w wypadku, gdy forma o końcówce spółgłoskowej nie zostanie wykryta ze względu na wymianę wewnątrztematową (*Gwarkowa, Orzeszkowa, Raczkówna* zostaną zlematyzowane do *Gwarko, Orzeszko, Raczko*, nie *Gwarek, Orzeszek, Raczek*).

Najważniejszą heurystyką dotycząca poszukiwania lematu, który zostanie następnie poddany interpretacji semantycznej, jest heurystyka centrum. Zarówno w *Słowosieci* jako takiej, jak i przede wszystkim w tabeli nazw, znajduje się niewiele wieloczłonowych jednostek nazewniczych. I są to w większości nazwy niekompozycyjne, głównie geograficzne (miasta *Góra Kalwaria, Grodzisk Mazowiecki, New Jersey, Nowy Jork, Rio de Janeiro, Ruda Śląska, San Francisco*, państwa *Nowa Zelandia, Trynidad i Tobago*, rzeki *Rio Grande*), ale i nazwiska (*ben Laden*), partie polityczne (*Platforma Obywatelska*) czy firmy (*Canal Plus*). Jednak wiele nazw, przede wszystkim nazw instytucji, ma charakter kompozycyjny. Wiele firm ma w nazwie termin *Fabryka, Przedsiębiorstwo, Zakład, Zakłady* (*Warszawska Fabryka Dźwigów,Państwowe Przedsiębiorstwo Instalacyjne, Przedsiębiorstwo Budownictwa Przemysłowego, Państwowy Zakład Ubezpieczeń, Zakład Przetwórstwa Owocowo-Warzywnego, Zakłady Tłuszczowe SA, Zakłady Cegielskiego*)[6], większość szkół ma w nazwie *szkoła (podstawowa, średnia), liceum, technikum* (*Szkoła Podstawowa nr 4, Liceum Ogólnokształcące im. Juliusza Słowackiego, Państwowe Technikum Korespondencyjne,*

[6] W nazwach występuje także termin *zakład* w znaczeniu jednostki organizacyjnej instytucji, np. *Zakład Anatomii Prawidłowej ŚAM.*

Technikum Hotelarskie, a także *Szkoła Oficerska Centralnego Ośrodka Szkolenia Służby Więziennej, Szkoła Policealna dla Dorosłych*), większość uczelni ma w nazwie *politechnika, uniwersytet* (*Politechnika Wrocławska, Uniwersytet Gdański, Uniwersytet Kardynała Wyszyńskiego*), itp., itd. (*Ministerstwo Rolnictwa, Wielkopolski Bank Kredytowy*). Dotyczy to także niektórych nazw geograficznych, np. ulic, dzielnic, i osiedli. W szczególności, w przykładach zdań prezentowanych m.in. na rys. 4 występuja nazwy *ulica Gdyńska, ulica Piaskowa, ulica Okrężna, osiedle 40-lecia*. Dlatego w sytuacji, gdy dana nazwa nie występuje w żadnym z rozważanych źródeł danych bezpośrednio, warto zastosować heurystykę szacowania jej interpretacji na podstawie jej centrum.

Heurystyka wykorzystuje fakt, że w *Składnicy* przechowywana jest informacja, który z terminali będących potomkami wierzchołka reprezentującego daną frazę stanowi jego centrum składniowe (w wypadku fraz rzeczownikowych centrum składniowe i semantyczne są sobie równe, por. Przepiórkowski, 2006; Hajnicz, 2011, s. 96). W tym celu podczas pierwotnego przeglądania drzewa każdemu wierzchołkowi przypisywany jest lemat jego centrum. Lemat ten jest następnie wyszukiwany w *Słowosieci* i tabeli nazw.

Bardzo dużą grupę wieloczłonowych jednostek nazewniczych stanowi para złożona z imienia i nazwiska danej osoby (*Barbara Blida, Paweł Piskorski, Kazimierz Pytko*). W większości wypadków centrum takich fraz stanowi imię. Jednak właściwą interpretacją nazwy nie jest *kobieta/mężczyzna określany imieniem*, lecz *człowiek ze względu na miano* (por. punkt 4.2.2). Dlatego w tym przypadku heurystyka centrum jest zastępowana bardziej specyficzną heurystyką „imienia i nazwiska".

Po zastosowaniu heurystyki formy żeńskiej nazwiska bądź heurystyki centrum może rzecz jasna być zastosowana heurystyka zmiany kaszty. Zauważmy jednak, że lemat centrum nie musi być (i zazwyczaj nie jest) zgodny pod względem kaszty z lematem nazwy własnej (i formą ortograficzną w tekście): jeśli jest to wyraz pospolity, to jego lemat pisany jest małą literą. Może to powodować drobną konfuzję, gdy dla nazwy *Uniwersytet Gdański* mamy `<plwn_units part="head" case_agreement="Full" polysemy="true">`, a następnie jego potomne elementy `unit` posiadają lemat *uniwersytet*.

Wszystkie omówione powyżej heurystyki dotyczą nazw będących frazami rzeczownikowymi. Nazwy mające postać fraz przyimkowych są bardzo nieliczne; niektóre ulice (*Przy Bażantarni, Przy Rondzie, Za Łąkami*), budynki (*dom Bez Kantów, dom Pod Sedesami*) czy organizacje (*kabaret Pod Egidą, Piwnica pod Baranami, restauracja U Aktorów*). Nazwy takie zazw. pojawiają się w tekście wraz z determinującym ich interpretację wyrazem pospolitym *dom, ulica, kabaret, restauracja*, i wtedy ma zastosowanie heurystyka centrum[7]. W przeciwnym razie zastosowanie może mieć co najwyżej heurystyka zmiany kaszty: takie nazwy nie bywają kompozycyjne.

[7] Heurystyka ta rzecz jasna czasem zawiedzie, np. dla *Piwnicy pod Baranami*.

Jak już wspominałam, w NKJP przymiotniki pochodne od nazw własnych są znakowane na poziomie nazw własnych, co jest zaznaczana przez dodanie typu derywacyjnego (`derivType`) o wartości `relAdj`. Z drugiej strony, w *Składnicy* nazwom takim odpowiadają frazy przymiotnikowe. Informację tę można więc wydobyć i dopasować.

Rzecz jasna nazwy rzeczownikowe interpretowane są za pomocą jednostek i synsetów rzeczownikowych; dotyczy to także nazw będących frazami przyimkowymi. Jest to spójne z ogólnymi zasadami znakowania semantycznego *Składnicy* jako całości[8]. Ta sama zasada musi dotyczyć nazw przymiotnikowych.

W *Słowosieci* występuje pewna liczba przymiotników pochodnych od nazw własnych; są one przetwarzane i reprezentowane w analogiczny sposób jak nazwy rzeczownikowe, tyle że element `plwn_units` zawiera atrybut `adjective="true"` (por. rys. 9). Natomiast nazwy takie nie są jak dotąd reprezentowane w tabeli nazw. Dlatego przygotowałam dla nich heurystyki ogólne:

— przypisanie jednego ze sztucznych synsetów przymiotnikowych (por. punkt 4.2.3):
  — pochodny od nazwy dzielnicy (`subtype="district"`),
  — pochodny od nazwy miejscowości (`subtype="settlement"`),
  — pochodny od nazwy obszaru (`subtype="bloc/country/region"`),
  — pochodny od nazwy geograficznej dla wszystkich pozostałych nazw posiadających typ `placeName` bądź `geogName`;
— przypisanie listy synsetów przymiotnikowych:
  — dla nazw typu `orgName`: partyjny, urzędowy, firmowy,
  — dla nazw typu `persName`: własny, osobniczy, swój.

Na rys. 9 widnieje przykład nazwy przymiotnikowej znajdującej się w *Słowosieci*, na rys. 10 — przykład nazwy mającej atrybut `subtype="settlement"` (identyfikator synsetu 200212 oznacza *pochodny od nazwy miejscowości*), zaś na rys. 11 — przykład nazwy posiadającej atrybut `type="orgName"`. Zwróćmy uwagę na fakt, że atrybut `case_aggreement` ma wartość `irrel`: procedura dopasowywania lematu zawiodła, a użyta heurystyka ogólna jest niezależna od lematu.

Istnieją jednak dwie sytuacje, w których nazwy o charakterze rzeczownikowym przypisywane są frazom morfoskładniowo przymiotnikowym:

— przymiotnikowe nazwy ulic (*Gdyńska, Okrężna, Piaskowa*), jezior (*Białe, Maszewskie*) itp.
— określenia czasu (*19.27, 12.50, 1299-ty, 756, dziesiąty, piąta*).

Jeśli nazwa zawiera centrum będące wyrazem pospolitym (*u zbiegu ulic Gdyńskiej i Piaskowej, na początku 1995 roku*), jest to nazwa rzeczownikowa, do

[8] Także w wypadku fraz przyimkowo-nominalnych; fraza przyimkowa dziedziczy sens rzeczownika będącego centrum dopełnienia rzeczownikowego przyimka (centrum semantycznym frazy przyimkowej jest rzeczownik).

```
<named name_id="named_36.40-s_n3">
    <nambase>hiszpański</nambase>
    <nkjp_type type="relAdj" subtype="country"/>
    <plwn_units part="whole" case_agreement="Full" polysemy="false"
                adjective="true">
        <unit luid="n3-sv1" status="auto" chosen="true">
            <lubase>hiszpański</lubase>
            <lusense>1</lusense>
            <luident>2134</luident>
            <synset>8567</synset>
        </unit>
    </plwn_units>
</named>
```

Rysunek 9. Przykład nazwy przymiotnikowej uwzględnionej w *Słowosieci*

```
<named name_id="named_131.36-s_nd">
    <nambase>sztokholmski</nambase>
    <nkjp_type type="relAdj" subtype="settlement"/>
    <plwn_types part="whole" case_agreement="irrel" polysemy="false"
                adjective="true">
        <type typid="nd-tv1" status="auto" chosen="true">
            <nbase>sztokholmski</nbase>
            <plwntype>200212</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 10. Przykład nazwy przymiotnikowej z dopasowanym sztucznym synsetem

```
<named name_id="named_184.67-s_n1">
    <nambase>LPR-owski</nambase>
    <nkjp_type type="relAdj" subtype="orgName"/>
    <plwn_types part="whole" case_agreement="irrel" polysemy="true"
                adjective="true">
        <type typid="n1-tv2" status="suggested">
            <nbase>LPR-owski</nbase>
            <plwntype>8344</plwntype>
        </type>
        <type typid="n1-tv3" status="suggested">
            <nbase>LPR-owski</nbase>
            <plwntype>9411</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 11. Nazwa przymiotnikowa, dla której zastosowano ogólną heurystykę

której stosuje się heurystykę centrum. Jednak brak takiego centrum (*dziś od 19.27, o piątej po południu,około dziesiątego, w 1299-tym, mieszkanie na Obożnej*) nie zmienia faktu, że nazwa taka interpretowana jest jako *godzina, dzień, rok, ulica* itp., czyli posiada typ rzeczownikowy.

Dlatego nazwa taka przypisywana jest frazie rzeczownikowej (`fno`) nadrzędnej nad frazą przymiotnikową (`fpt`), o ile taka fraza o tych samych granicach w zdaniu występuje. Przykład takiej nazwy (ze zdania *Już w 756 było biskupstwo.*) można zobaczyć na rys. 12.

Przyjęte zostało założenie, ze nazwy typu `time` oraz `date` posiadają wyłącznie interpretację rzeczownikową. Jako że nazwy tego typu tak naprawdę nie posiadają lematu, atrybut `part` ma wartość `irrel`.

Dla nazw rzeczownikowych, dla których nie udało się znaleźć lematu za pomocą heurystyk omówionych powyżej, a którym udało się przyporządkować wierzchołek w drzewie, tworzona była pusta nazwa zawierająca element `<plwn_types part="whole" case_agreement="irrel" polysemy="irrel">`, posiadająca jednego potomka `type` z atrybutami `status="added"` i `chosen="true"`.

5.2.3. Dyzambiguacja

Spora liczba nazw własnych jest niejednoznaczna, np. *Mars* to planeta, bóg wojny oraz baton czekoladowy, *Alberta* to imię żeńskie, jezioro i prowincja w Kanadzie, *Brzezina* ma w tabeli nazw 6 interpretacji, niezależnie od pospolitej nazwy lasu; z kolei lematowi *szkoła* uzyskiwanemu na zasadzie heurystyki

```
<named name_id="named_82.35-s_n1">
    <nambase></nambase>
    <nkjp_type type="date"/>
    <plwn_types part="irrel" case_agreement="irrel" polysemy="true">
        <type typid="n1-tv1" status="suggested">
            <nbase>0756</nbase>
            <plwntype>5084</plwntype>
        </type>
        <type typid="n1-tv2" status="suggested">
            <nbase>0756</nbase>
            <plwntype>21452</plwntype>
        </type>
        <type typid="n1-tv3" status="suggested">
            <nbase>0756</nbase>
            <plwntype>48395</plwntype>
        </type>
        <type typid="n1-tv4" status="suggested" chosen="true">
            <nbase>0756</nbase>
            <plwntype>65377</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 12. Nazwa semantycznie rzeczownikowa, morfoskładniowo przymiotnikowa

centrum przypisane jest w *Słowosieci* 1.6 i 1.8 aż 10 jednostek! Dlatego podjęta została próba dyzambiguacji niejednoznacznych nazw własnych, w których wykorzystywana jest informacja o typie i podtypie nazwy przypisanej jej w NKJP. W tym celu każdemu typowi i podtypowi rozważanemu w NKJP (por. punkt 2.1) przypisane została lista identyfikatorów synsetów, jakie zostały przypisane nazwom tego typu w tabeli nazw. Liczność tych list jest bardzo zróżnicowana, por. tabela 1.

Procedura dyzambiguacji oparta była na heurystyce najczęstszego sensu (por. Gale *et al.*, 1992; Agirre i Edmonds, 2006). Jednak heurystyka ta stosowana była jedynie do podzbioru interpretacji zgodnych z typem i (w wypadku jego występowania) podtypem nazwy. W wypadku nazw występujących w *Słowosieci* (element `plwn_units`) zgodność badana była dla bezpośredniego typu nazwy i wszystkich jego hiperonimów. Jedynie w wypadku, gdy żaden z sensów wyrazów nie był zgodny z taką listą, rozważany był cały zbiór interpretacji.

Zastosowanie właściwej heurystyki polegało na wyborze najwyższego numeru znaczenia (`lusense`) dla danego lematu, w wypadku tabeli nazw wybierany był synset umieszczony jako pierwszy (spośród zgodnych) na liście interpreta-

geogName	orgName	persName		placeName		time
23	55		89		1	8
		addName	1	bloc	1	
		forename	2	country	3	
		surname	1	district	2	
				region	11	
				settlement	5	

Tabela 1. Liczba interpretacji przypisana poszczególnym typom nazw NKJP

cji nazwy. Wybrana interpretacja opatrywana była atrybutami `status="auto"` oraz `chosen="true"`.

W wypadku nazw przymiotnikowych nie jest dokonywana dyzambiguacja, bo nie ma do niej podstaw. Zamiast tego wszystkie elementy `type` opatrywane są atrybutem `status="suggested"`. Jeden z elementów może zostać następnie opatrzony atrybutem `chosen="true"` podczas ręcznej korekty znakowania, ale nie jest to obowiązkowe.

5.3. Ręczna korekta znakowania nazw własnych

Wersja *Składnicy frazowej* w sposób automatyczny opatrzona interpretacją semantyczna nazw własnych zgodnie z procedurą omówioną powyżej została następnie w całości poddana ręcznej korekcie. Korekta ta miała za zadanie

1. poprawę heurystycznie dobranych słowosieciowych typów nazw,
2. określenie słowosieciowych typów nazw, których nie udało się odnaleźć w *Słowosieci* ani w tabeli nazw,
3. znalezienie adekwatnych wierzchołków dla nazw, dla których nie udało się dopasować wierzchołka w sposób automatyczny,
4. dodanie nazw zignorowanych w znakowaniu NKJP.

Punkty 3–4 były realizowane wyłącznie wtedy, gdy ich zaniechanie powodowało pozostawienie w zdaniu jakichś niezinterpretowanych wyrazów (rzeczowników bądź przymiotników), które nie mogły być interpretowane jako wyrazy pospolite. Tak więc nie były dodawane interpretacje dla nazw

— będących częścią składową zinterpretowanej nazwy,
— których wszelkie części składowe zostały zinterpretowane.

Na przykład najmniejszą frazą obejmującą nazwę *Wendy Melvoin* w zdaniu *Jednak główną rolę w pracy nad brzmieniem płyty odegrały gitarzystka Wendy Melvoin i grająca na instrumentach klawiszowych Lisa Coleman.* jest *gitarzystka Wendy Melvoin.* Centrum tej frazy stanowi wyraz *gitarzystka*, który poprawnie interpretuje całą nazwę. Z drugiej strony w NKJP występują tez nazwy

Wendy oraz *Melvoin*, które zostały poprawnie przekonwertowane do *Składnicy*. Dodawanie tej nazwy do *Składnicy* jest więc zbędne. Ponadto znakowanie takie nie było realizowane w wypadku gdy nadfraza obejmowała zdanie względne ze względu na niedopuszczalny poziom nadmiarowości takiej nadfrazy[9].

Punkt 1 dotyczy błędnego wyboru interpretacji nazwy i realizowany jest na dwa sposoby. Jeśli na liście elementów `unit` bądź `type` znajduje się poprawna interpretacja, element ten uzupełniany jest o atrybuty `status="modif"` oraz `chosen="true"` (por. rys. 7). W przeciwnym razie dodawany jest nowy element `type` (element `nbase` dziedziczy nazwę po `nambase`), przy czym atrybut `status` uzyskuje wartość `added`. Jest to ważne rozróżnienie z dwóch powodów. Przede wszystkim wskazuje brak danej interpretacji w tabeli nazw. Informacja ta wykorzystywana jest także podczas oceny procesu konwersji: jeśli poprawnej interpretacji nie było na liście, automatyczne dokonanie poprawnego wyboru nie było możliwe (por. rys. 13). W wypadku gdy element `named` nie zawierał potomka `plwn_types` (a wyłącznie `plwn_units`), element ten dodawany jest w całości (por. rys. 14). W obu wypadkach błędnie wybrana interpretacja ma zmieniany atrybut `chosen` na `false`. Z jednym wyjątkiem: gdy znaczenie wybrane automatycznie i to poprawne są trudne do rozróżnienia, atrybut ten uzyskuje wartość `close` (por. rys. 14). W wyjątkowych wypadkach dotyczy to także elementu `plwn_types` (np. *Maria* jako imię męskie: *Jan Maria Rokita*), a nie tylko zbliżonych znaczeń w *Słowosieci* dla elementu `plwn_units`. Jest to informacja ważna z wielu względów: dla przyszłego rozwoju *Słowosieci*, w wypadku wykorzystywania przyszłej *Składnicy Semantycznej* do ewaluacji algorytmów ujednoznaczniania znaczenie wyrazów (ang. *Word Sense Disambiguation*, WSD, Agirre i Edmonds, 2006). Szczególny przypadek stanowi sytuacja, gdy poprawna interpretacja znajduje się w jednym z elementów `plwn_units, plwn_types`. Wówczas niepoprawna interpretacja uzyskuje `chosen="false"` bez konieczności znajdowania interpretacji poprawnej (por. rys. 22).

Realizacja punktu 2 ogranicza się do dobrania właściwej interpretacji (element `plwntype`, por. rys. 15). Specyficzny przypadek stanowi sytuacja, gdy fiasko dopasowania lematu wynikało z faktu, że w zdaniu wystąpił jedynie fragment nazwy (np. *Brytania* zamiast *Wielkiej Brytanii*). Inny specyficzny przypadek stanowią inicjały imion i nazwisk, które oczywiście nie są nigdzie reprezentowane. Przeciwnie niż w NKJP, przyjęłam konwencję, że inicjały są zawsze interpretowane jako człowiek ze względu na miano, bez względu na to, czy występują na pozycji imienia czy nazwiska (*A. Mickiewicz, Adam M.*).

W *Słowosieci* znajduje się pewna liczba jednostek wieloczłonowych (*szkoła podstawowa, szkoła ponadpodstawowa, szkoła średnia, szkoła wyższa; kopalnia soli, kopalnia węgla brunatnego, kopalnia węgla kamiennego*); ich udział rośnie

[9] Zgodnie z powyższymi zasadami, mogło to wymusić znakowanie podfraz nazwy.

```
<named name_id="named_4.62-s_n2">
    <nambase>Wik</nambase>
    <nkjp_type type="geogName"/>
    <plwn_types part="whole" case_agreement="Full" polysemy="false">
        <type typid="n2-tv1" status="auto" chosen="false">
            <nbase>Wik</nbase>
            <plwntype>200001</plwntype>
        </type>
        <type typid="n2-tv2" status="added" chosen="true">
            <nbase>Wik</nbase>
            <plwntype>4267</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 13. Przykład nazwy własnej z dodaną właściwą interpretacją

wraz z rozwojem zasobu (w *Słowosieci* 1.6 stanowią już ok. 20% jednostek rzeczownikowych). Heurystyka centrum dla nazw takich jak *Szkoła Podstawowa nr 4, Wyższa Szkoła Bankowa, Kopalnia Soli w Wieliczce* znajdzie jednostki odpowiednio o lematach *szkoła, kopalnia* gdy tymczasem bardziej adekwatne byłyby lematy *szkoła podstawowa, szkoła wyższa, kopalnia soli*. Dlatego w ramach ręcznej korekty poprawiana była także interpretacja takich nazw[10]. Korekta taka dotyczy jedynie elementu `plwn_units`, któremu przypisywany jest atrybut `part="sub"`. Wpisywane są wszystkie jednostki o danym lemacie (w *Słowosieci 1.6* istnieją 3 jednostki o lemacie *szkoła podstawowa*), jedna opatrywana jest atrybutem `chosen="true"`. Rzecz jasna wartość tego atrybutu dla interpretacji wybranej automatycznie zmieniana jest na `close`, o ile tylko heurystyka wyboru interpretacji była poprawna. Przykład takiej nazwy znajduje się na rys. 16. Czasem samo zastosowanie heurystyki centrum mogło być nieadekwatne (nazwa *Dom Dziecka na Oruni* byłaby błędnie interpetowana jako *dom* i poprawnie jako *dom dziecka*, który nie jest w żadnym znaczeniu hiponimem *domu*).

W wypadku, gdy pewna nazwa nie występuje w *Słowosieci*, lecz znajduje się tam jej synonim, element `plwn_units` wprowadzany jest z atrybutem `part="synonym"`. Przykładem takiej sytuacji jest termin *Wysoka Izba* (rys. 17). Termin ten nie jest egzemplarzem *sejmu* czy też *senatu* (jak *Warszawa* jest egzemplarzem *miasta*), więc zastosowanie tu elementu `plwn_types` byłoby w tym wypadku niepoprawne[11].

[10] Pomysł ten pojawił się już w trakcie korekty, dlatego nie został zrealizowany w pełni konsekwentnie.

[11] *Wysoka Izba* nie została uznana przez leksykografów NKJP za nazwę własną. Ponie-

```
<named name_id="named_4.19-s_n1">
    <nambase>Rada Europy</nambase>
    <nkjp_type type="orgName"/>
    <plwn_units part="head" case_agreement="Full" polysemy="true">
        <unit luid="n1-sv1" status="auto" chosen="close">
            <lubase>rada</lubase>
            <lusense>1</lusense>
            <luident>7249</luident>
            <synset>7801</synset>
        </unit>
        <unit luid="n1-sv2">
            <lubase>rada</lubase>
            <lusense>2</lusense>
            <luident>21080</luident>
            <synset>2845</synset>
        </unit>
        <unit luid="n1-sv3">
            <lubase>rada</lubase>
            <lusense>3</lusense>
            <luident>7250</luident>
            <synset>2844</synset>
        </unit>
    </plwn_units>
    <plwn_types part="whole" case_agreement="irrel" polysemy="irrel">
        <type typid="n1-tv1" status="added" chosen="true">
            <nbase>Rada Europy</nbase>
            <plwntype>200013</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 14. Nazwa własna ze zbliżoną interpretacją wybraną automatycznie

```
<named name_id="named_48.31-s_n2">
    <nambase>LPR</nambase>
    <nkjp_type type="orgName"/>
    <plwn_types part="whole" case_agreement="irrel" polysemy="irrel">
        <type typid="n2-tv1" status="added" chosen="true">
            <nbase>LPR</nbase>
            <plwntype>7692</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 15. Przykład nazwy własnej spoza *Słowosieci* i tabeli nazw

Realizacja punktów 3 i 4 wymaga stworzenia nowego elementu `named`; atrybut `case_aggreement` ma wówczas zawsze wartość `irrel`. Realizacja punktu 3 wymaga znalezienia w drzewie rozbioru zdania nadfrazy bądź podfrazy nazwy. Dotyczy on wyłącznie nazw wieloczłonowych, w drzewie zawsze istnieją frazy obejmujące dowolną nazwę jednoczłonową. W wypadku podfrazy prefiks identyfikatora nazwy (atrybut `name_id`) zmieniany jest na `subname` (por. rys. 18), zaś w wypadku nadfrazy — na `supername` (por. rys. 19): jest to istotne, gdyż jest to jedynie przybliżona reprezentacja danej nazwy. Lemat nazwy kopiowany jest z NKJP. Dalsza procedura sprowadza się do znalezienia właściwej nazwy w *Słowosieci* (element `plwn_units`) lub tablicy nazw (element `plwn_types`) i wybrania poprawnej interpretacji spośród dostępnych, a w wypadku fiaska ręczne dobranie dodatkowej interpretacji. W tym drugim wypadku atrybut `part` ma wartość `sub, super`, odpowiednio, zaś element `nbase` przyjmuje wartość lematu całej frazy. Szczególny przypadek stanowią przymiotniki przyprzymiotnikowe (*białorusko*), które w *Składnicy* nie posiadają odrębnego nieterminala, i są interpretowane w wierzchołku przymiotnika złożonego (*białorusko-polski*, por. rys. 20).

Punkt 4 realizowany jest w sposób analogiczny, tyle że dla nazwy pominiętej w NKJP powinna istnieć w *Składnicy* adekwatna fraza. Prefiks identyfikatora nazwy zmieniany jest na `ignored`, element `nkjp_type` posiada atrybut `type="undefined"`. Numerację takich identyfikatorów stanowią kolejne liczby ujemne[12]. Przykład takiej nazwy znajduje się na rys. 21.

W kilku przypadkach poza nazwami zostały ręcznie oznakowane idiomy, gdyż ich interpretacja semantyczna jest niekompozycyjna. Prefiks nazwy ma

waż nie jest to oczywisty błąd, a termin ten jest ewidentnie niekompozycyjny, identyfikator nazwy posiada prefiks `lexicalized`.

[12] Jako że identyfikatory mają inny prefiks, możnaby im nadawać kolejne liczby dodatnie. Jednak wówczas identyfikatory `luid` i `typid` nie byłyby unikalne w ramach zdania.

```
<named name_id="named_81.12-s_n1">
    <nambase>Klub PSL</nambase>
    <nkjp_type type="orgName"/>
    <plwn_units part="head" case_agreement="Full" polysemy="true">
        <unit luid="n1-sv1" status="auto" chosen="close">
            <lubase>klub</lubase>
            <lusense>1</lusense>
            <luident>108270</luident>
            <synset>78523</synset>
        </unit>
        <unit luid="n1-sv2">
            <lubase>klub</lubase>
            <lusense>2</lusense>
            <luident>2580</luident>
            <synset>1256</synset>
        </unit>
        <unit luid="n1-sv3">
            <lubase>klub</lubase>
            <lusense>3</lusense>
            <luident>2579</luident>
            <synset>4855</synset>
        </unit>
        <unit luid="n1-sv4" status="added" chosen="true">
            <lubase>klub parlamentarny</lubase>
            <lusense>1</lusense>
            <luident>108271</luident>
            <synset>78527</synset>
        </unit>
    </plwn_units>
</named>
```

Rysunek 16. Przykład nazwy własnej interpretowanej przez jednostkę wieloczłonową

```
<named name_id="lexicalized_23.13-s_n-1">
   <nambase>Wysoka Izba</nambase>
   <nkjp_type type="undefined"/>
   <plwn_units part="synonym" case_agreement="irrel" polysemy="true">
       <unit luid="n-1-sv1">
           <lubase>sejm</lubase>
           <lusense>1</lusense>
           <luident>21211</luident>
           <synset>7813</synset>
       </unit>
       <unit luid="n-1-sv2">
           <lubase>sejm</lubase>
           <lusense>2</lusense>
           <luident>7855</luident>
           <synset>7814</synset>
       </unit>
       <unit luid="n-1-sv3">
           <lubase>sejm</lubase>
           <lusense>3</lusense>
           <luident>17981</luident>
           <synset>8865</synset>
       </unit>
       <unit luid="n-1-sv4" status="added" chosen="true">
           <lubase>sejm</lubase>
           <lusense>4</lusense>
           <luident>7856</luident>
           <synset>3038</synset>
       </unit>
   </plwn_units>
</named>
```

Rysunek 17. Przykład interpretacji frazy za pomocą jej synonimu

```
 <named name_id="subname_13.51-s_n1">
     <nambase>Alpy Julijskie</nambase>
     <nkjp_type type="geogName"/>
     <plwn_types part="sub" case_agreement="irrel" polysemy="irrel">
         <type typid="n1-tv1" status="added" chosen="true">
             <nbase>Alpy</nbase>
             <plwntype>8363</plwntype>
         </type>
     </plwn_types>
 </named>
```

Rysunek 18. Przykład nazwy interpretowanej w wierzchołku podfrazy

```
<named name_id="supername_1.62-s_n1">
    <nambase>ul. Wilcza</nambase>
    <nkjp_type type="geogName"/>
    <plwn_types part="super" case_agreement="irrel" polysemy="irrel">
        <type typid="n1-tv1" status="added" chosen="true">
            <nbase>ul. Wilcza w Warszawie</nbase>
            <plwntype>1167</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 19. Przykład nazwy interpretowanej w wierzchołku nadfrazy

```
<named name_id="supername_1.67-s_n1">
    <nambase>białoruski</nambase>
    <nkjp_type type="relAdj" subtype="country"/>
    <plwn_types part="super" case_agreement="irrel" polysemy="irrel"
                adjective="true">
        <type typid="n1-tv1" status="added" chosen="true">
            <nbase>białorusko-polski</nbase>
            <plwntype>2262</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 20. Przykład nazwy będącej przymiotnikiem przyprzymiotnikowym

```
<named name_id="ignored_30.62-s_n-2">
    <nambase>WIG20</nambase>
    <nkjp_type type="undefined"/>
    <plwn_types part="whole" case_agreement="irrel" polysemy="irrel">
        <type typid="n-2-tv1" status="added" chosen="true">
            <nbase>WIG20</nbase>
            <plwntype>200016</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 21. Przykład nazwy własnej pominiętej w znakowaniu NKJP

wówczas postać `lexicalized`, a odpowiednie elementy `type` są opatrzone atrybutem `usage` o wartości `metaph` lub `idiom`. Na przykład w zdaniu (3) nazwa *Pokój 240* została użyta metaforycznie, w znaczeniu grupy ludzi zajmujących ten pokój.

(3) *Pokój 240 naparł silnie na dziennikarzy i prawem fizyki wywalił ich z przepełnionego namiotu.*
(4) *Gramy Beethovena, dziś nawet idą tradycyjnie koncerty fortepianowe.*

Podobnie rzecz się ma w wypadku metonimii. W zdaniu (4) nie chodzi o granie kompozytora, tylko jego utworów. Oczywiście nazwa *Beethoven* jako kompozytor także jest reprezentowana. Aby nie umieszczać dwóch elementów `named` w jednym wierzchołku, interpretacja metonimiczna była umieszczana w wierzchołku o poziom wyżej (kategorii `fno` zamiast `formarzecz`).

W analogiczny sposób traktowane były neologizmy (*fenomulat*, prefiks nazwy `neologism`), wyrażenia obce (*tupacamaros*, prefiks nazwy `foreign`) oraz wyrażenia gwarowe (*mućka*, prefiks nazwy `dialect`). Jako że zazwyczaj taki idiom, a neologizm, czy wyrażenie obce lub gwarowe nie jest nazwą własną, atrybut `part` ma wartość `irrel`.

5.4. Ocena automatycznej konwersji nazw własnych

Ręczna korekta automatycznego znakowania *Składnicy frazowej* nazwami własnymi stanowi podstawę do oceny jakości procesu automatycznego.

Spośród 4473 nazw z milionowego NKJP znajdujących się w zdaniach posiadających drzewo rozbioru, a więc takich, które potencjalnie mogły zostać przekonwertowane, w *Składnicy* znalazły się 4294 nazwy, czyli aż 96%. Fiasko konwersji dla pozostałych nazw spowodowane było różnicami pomiędzy

geogName	orgName	persName		placeName		date	time
268	683		1915		1233	166	29
		addName	81	bloc	9		
		forename	759	country	553		
		surname	721	district	19		
				region	38		
				settlement	587		

Tabela 2. Liczba przekonwertowanych nazw poszczególnych typów NKJP

zakresem nazwy a granicami odpowiadającej jej frazy. Dla 15 takich nazw dopasowana została w *Składnicy* podfraza[13], a dla 28 — nadfraza.

Ponadto oznakowane zostały 122 nazwy pominięte w znakowaniu NKJP, 7 zwrotów obcojęzycznych, 5 leksykalizacji, 4 neologizmy, 3 przypadki metonimii i jedno wyrażenie gwarowe.

Dalsza analiza dotyczy jedynie nazw, dla których automatyczna konwersja została zrealizowana. W tabeli 2 przedstawiony został podział przekonwertowanych nazw według ich typów i podtypów. Jak widać, największą ich liczbę stanowią nazwy osobowe (ok. 45%), jeśli jednak potraktować nazwy miejsc i nazwy geograficzne łącznie, okaże się, że jest ich niewiele mniej (ok. 35%).

W tabeli 3 przedstawione zostały dane liczbowe dotyczące decyzji co do wyboru jednostki (*Słowosieć*) lub typu (tabela nazw) względem wyboru podjętego automatycznie. Kolumna `chosen` (od nazwy atrybutu) wskazuje typ podjętej decyzji: true oznacza akceptację decyzji podjętej automatycznie, modif, added oznaczają, że atrybut `chosen="true"` przypisany został elementowi o adekwatnej wartości atrybutu `status`, zaś cl&mod, cl&add oznaczają to samo pod warunkiem, że element wybrany automatycznie miał skorygowany atrybut na `chosen="close"` (a nie `false`). Wiersz false opisuje sytuację, w której żaden element nie został wybrany i oznacza, że dodatkowy element `named` został dodany do innego wierzchołka (dotyczy to przymiotnikowych nazw ulic; odpowiednia heurystyka została wdrożona dopiero w poprawionej wersji programu). Na koniec, wiersz c&f&a ma zastosowanie jedynie dla lematu rozważonego w obu źródłach i oznacza, że w jednym wypadku element wybrany automatycznie uzyskał atrybut `chosen="close"`, a w drugim `false`.

Kolumny `unit, type` oznaczają odpowiednio, że nazwa reprezentowana jest przez `plwn_units` bądź `plwn_types` (w wypadku podkolumn kolumny OBA zaznaczane jest, w którym elemencie pojawił się atrybut `chosen="true"`). Specyficzny charakter ma wiersz ADDED, oznaczający, że dodany został cały element `plwn_types` (w jednym wypadku `plwn_units`), a nie tylko element `unit`

[13] W wypadku nazwy wieloczłonowej jeden nazwie mogło zostać przyporządkowane kilka podfraz.

chosen	unit	type	OBA				RAZEM
			unit	type	oba	razem	
true	1305 (314)	1599 (22)	39	70	38	147	3051
modif	50 (2)	3 (0)	1	0	2	3	56
added	10 (1)	36 (0)	0	9	0	9	55
cl&mod	39 (0)	3 (0)	0	0	0	0	42
cl&add	5 (0)	81 (3)	0	0	0	0	86
false	5 (5)	0 (0)	0	0	0	0	5
c&f&a	0 (0)	0 (0)	0	12	0	13	12
razem	1414 (322)	1722 (25)	40	91	40	171	3307
ADDED	6 (0)	765 (151)	1	215	—	216	987
RAZEM	1420 (322)	2487 (176)	41	306	40	387	4294

Tabela 3. Rodzaj decyzji podjętej w zależności od źródła informacji o nazwie

czy type. Liczby w podkolumnie type kolumny OBA oznaczają w tym wypadku dodanie elementu plwn_types przy istniejącym elemencie plwn_units. Liczby w nawiasach oznaczają liczbę przymiotników w wypadku kolumny unit (i wiersza ADDED) oraz liczby pochodnych nazw żeńskich dla pozostałych wierszy kolumny type.

W tabeli 4 widnieją takie same dane, ale wyłącznie dla nazw niejednoznacznych w (przynajmniej jednym) ostatecznie wybranym źródle danych[14]. 46,2% wszystkich nazw posiadało interpretację w *Słowosieci*, 48,9% znajdowało się w tabeli nazw, zaś 4,9% uwzględnione zostało w obu zasobach (odpowiednio 78,8%, 16,6% i 4,6% dla nazw wieloznacznych). Udało się automatycznie zinterpretować 82% nazw, przy czym 19,2% z nich było niejednoznacznych. Sumarycznie, automatyczne heurystyki zadziałały poprawnie w 86,6% sytuacji (80,1% dla jednostek *Słowosieci* oraz 92,8% dla typów z tabeli nazw), jednak gdy weźmiemy pod uwagę wyłącznie nazwy niejednoznaczne (w danym źródle interpretacji), uzyskamy już jedynie 65,4%, 61,2% i 86,6%, odpowiednio[15]. Wyniki te stają się odrobinę lepsze po uwzględnieniu znakowania uznanego za zbliżone.

Poniżej uzyskane wyniki przeanalizowane zostaną pod kątem użytych heurystyk. Pod uwagę zostaną wzięte wyłącznie wyniki uzyskane automatycznie (zignorowany będzie wiersz ADDED). Analiza zostanie przeprowadzona dla każdego źródła oddzielnie, tak jak aplikowane były heurystyki, oraz w wypadku

[14] W wypadku wiersza ADDED oznacza to status"suggested".

[15] W powyższych wyliczeniach z wiersza ADDED została uwzględniona tylko kolumna OBA.

chosen	unit	type	OBA				RAZEM
			unit	type	oba	razem	
true	326 (32)	97 (1)	1	10	8	19	442
modif	50 (2)	3 (0)	1	0	2	3	56
added	5 (0)	9 (0)	0	4	0	4	18
cl&mod	39 (0)	2 (0)	0	0	0	0	41
cl&add	5 (0)	1 (0)	0	0	0	0	6
false	2 (2)	0 (0)	0	0	0	0	2
c&f&a	0 (0)	0 (0)	0	5	0	5	5
razem	427 (37)	112 (1)	2	19	10	31	570
ADDED	2 (0)	33 (14)	0	106	—	106	141
RAZEM	429 (37)	145 (15)	2	125	10	137	711

Tabela 4. Rodzaj decyzji podjętej w zależności od źródła informacji o nazwie wieloznacznej

Słowosieci także dla przymiotników. Przeprowadzanie odrębnej analizy dla 25 form żeńskich wydaje się zbędne.

Zacznijmy od nazw reprezentowanych w *Słowosieci*. W tabeli 5 zaprezentowane są statystyki dla wszystkich nazw, a w tabeli 6 — nazwy wieloznaczne. Kolumny wskazują, jaka heurystyka doboru kaszty została zastosowana. Każda kolumna zawiera trzy podkolumny wskazujące, że nazwa była analizowana w całości (W), zastosowana została heurystyka centrum (H) oraz wynik sumaryczny (R). Wiersze oznaczane są tak jak poprzednio. Nowość stanowią dwa ostatnie wiersze podające wyniki procentowe: dla interpretacji wybranych automatycznie (wiersz true) oraz dla z uwzględnieniem interpretacji zbliżonych do prawidłowych (suma wierszy true, cl&mod, cl&add i close). Procenty liczone są zawsze względem wszystkich adekwatnych interpretacji (wartość w prawym dolnym rogu tabeli).

Spośród 475 segmentów przymiotnikowych pochodnych od nazw własnych przekonwertowanych z milionowego NKJP do *Słowosieci*, lematy 323 (68%) znalazły się w *Słowosieci*, z czego aż 313 (96,9%) zostało poprawnie zinterpretowane bez użycia jakichkolwiek heurystyk (32 spośród 37 niejednoznacznych — 86,5%), 5 było błędnymi nazwami ulic, dla 2 zastosowano heurystykę `FirstLow` (raz z sukcesem), a 3 wymagały modyfikacji. Można więc powiedzieć, że nazwy przymiotnikowe, o ile tylko w *Słowosieci* wystąpiły, nie powodowały większych kontrowersji.

Przejdźmy teraz do tabeli nazw, w której występują wyłącznie nazwy rzeczownikowe. W tabeli 7 widnieją statystyki dla wszystkich nazw, a w tabeli 8 — dla nazw wieloznacznych.

chosen	Full			FirstLow			FirstUp			AllLow			RAZEM		
	W	H	R	W	H	R	W	H	R	W	H	R	W	H	R
true	922	356	1278	88	13	101	0	0	0	4	0	4	1014	369	1383
modif	10	33	43	7	2	9	0	0	0	0	0	0	17	35	52
added	1	6	7	2	0	2	0	1	1	0	0	0	3	7	10
cl&mod	6	18	24	12	3	15	0	0	0	0	0	0	18	21	39
cl&add	0	2	2	3	0	3	0	0	0	0	0	0	3	2	5
close	3	20	23	4	2	6	0	0	0	0	0	0	7	22	29
false	17	85	102	138	32	170	0	3	3	9	1	10	164	121	285
razem	959	520	1479	254	52	306	0	4	4	13	1	15	1226	577	1803
procent	51,1	19,7	70,9	4,9	0,7	5,6	0,0	0,0	0,0	0,2	0,0	0,2	56,2	20,5	76,7
proc4cl	51,6	22,0	73,6	5,9	1,0	6,9	0,0	0,0	0,0	0,2	0,1	0,2	57,8	23,0	80,8

Tabela 5. Statystyki dla wszystkich nazw ze *Słowosieci*

chosen	Full			FirstLow			FirstUp			AllLow			RAZEM		
	W	H	R	W	H	R	W	H	R	W	H	R	W	H	R
true	37	215	252	68	7	75	0	0	0	1	0	1	106	222	328
modif	10	33	43	7	2	9	0	0	0	0	0	0	17	35	52
added	1	2	3	2	0	2	0	0	0	0	0	0	3	2	5
cl&mod	6	18	24	12	3	15	0	0	0	0	0	0	18	21	39
cl&add	0	2	2	3	0	3	0	0	0	0	0	0	3	2	5
close	1	17	18	1	2	3	0	0	0	0	0	0	2	19	21
false	1	42	43	66	15	81	0	0	0	2	0	2	69	57	126
razem	56	329	385	159	29	188	0	0	0	3	0	3	218	358	576
procent	6,4	37,3	43,8	11,8	1,2	13,0	0,0	0,0	0,0	0,2	0,0	0,2	18,4	38,5	56,9
proc4cl	7,6	43,7	51,4	14,6	2,1	16,7	0,0	0,0	0,0	0,2	0,0	0,2	22,4	45,8	68,2

Tabela 6. Statystyki dla wieloznacznych nazw ze *Słowosieci*

chosen	Full			FirstLow			FirstUp			AllLow			AllUp			RAZEM		
	W	H	R	W	H	R	W	H	R	W	H	R	W	H	R	W	H	R
true	1669	38	1707	0	0	0	0	0	0	0	0	0	1	0	1	1670	38	1708
modif	4	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4
added	29	14	43	0	0	0	0	2	2	2	0	2	0	0	0	31	16	47
cl&mod	3	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	3
cl&add	85	5	90	0	0	0	0	0	0	0	0	0	0	0	0	85	5	90
close	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
false	12	1	13	25	0	25	0	1	1	1	0	1	0	0	0	38	2	40
razem	1803	58	1861	25	0	25	0	3	3	3	0	3	1	0	1	1832	61	1893
procent	88,2	2,0	90,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,1	88,2	2,0	90,2
proc4cl	92,9	2,3	95,1	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,1	0,0	0,1	92,9	2,3	95,2

Tabela 7. Statystyki dla wszystkich nazw z tabeli nazw

chosen	Full			FirstLow			FirstUp			AllUp			RAZEM		
	W	H	R	W	H	R	W	H	R	W	H	R	W	H	R
true	108	7	115	0	0	0	0	0	0	1	0	1	109	7	116
modif	4	0	4	0	0	0	0	0	0	0	0	0	4	0	4
added	8	5	13	0	0	0	0	1	1	0	0	0	8	6	14
cl&mod	2	0	2	0	0	0	0	0	0	0	0	0	2	0	2
cl&add	1	4	5	0	0	0	0	0	0	0	0	0	1	4	5
razem	123	16	139	0	0	0	0	1	1	1	0	1	124	17	141
procent	76,6	5,0	81,6	0,0	0,0	0,0	0,0	0,0	0,0	0,7	0,0	0,7	77,3	5,0	82,3
proc4cl	78,7	7,8	86,5	0,0	0,0	0,0	0,0	0,1	0,1	0,7	0,0	0,7	79,4	7,8	87,2

Tabela 8. Statystyki dla nazw wieloznacznych z tabeli nazw

Jak widać, bez zastosowania żadnej heurystyki zinterpretowano 53,2% nazw posiadających interpretacje słowosieciowe (i jedynie 9,7% spośród wieloznacznych) oraz aż 95,2% nazw występujących w tabeli nazw (87,2% wieloznacznych). Widać więc wyraźnie, że heurystyki dopasowywania lematu mają wyraźnie większe zastosowanie podczas poszukiwania interpretacji w *Słowosieci*. Najbardziej skuteczna jest heurystyka centrum oraz zamiana pierwszej litery lematu (ew. wszystkich pierwszych liter wyrazów składających się na nazwę wieloczłonową), w wypadku nazw z tabeli jakiekolwiek znaczenie ma tylko pierwsza z nich. Aż o 12 punktów procentowych (do 90,2%) podnoszą heurystyki skuteczność interpretacji nazw w *Słowosieci* (o 50 — od 6,4% do 56,9% w wypadku nazw wieloznacznych). Zgodnie z przewidywaniami, dotyczy to głównie nazw interpretowanych przez centrum będące wyrazem pospolitym, typu *szkoła* czy *ministerstwo*; wyrazy pospolite są w większości bardziej wieloznaczne od nazw.

Nieliczne przypadki zastosowania heurystyki centrum dla nazw z tabeli dotyczą przede wszystkim nazw geograficznych, takich jak miejscowości (*Kamieniec Podolski, Bukowina Tatrzańska*), rzeki (*Drwęca Warmińska*) czy pasma górskie (*Beskid Żywiecki, Tatry Wysokie*), lecz także nazw instytucji (*PKO BP, KC PZPR, ZOZ w Piasecznie* o centrach *PKO, KC, ZOZ* występujących w tabeli nazw).

5.5. Źródła błędów konwersji

W założeniu *Słowosieć* oraz tabela nazw powinny być rozłączne. Istnieją jednak przypadki, w których nie do końca jest to prawdą. Dotyczy to umieszczonych w tabeli nazw:

— homonimicznych z wyrazami pospolitymi, dotyczy to imion (*Róża*), nazwisk (*Dymek, Pieróg, Sikora, Słomka, Wilk*), ale i miast (*Łódź*)[16];
— homonimicznych nazw, z których jedna występowała w *Słowosieci*, a inna w tabeli: w *Słowosieci* są miasta *Jarosław, Kalisz, Paryż, Warszawa*, zaś w tabeli znajdują się imię i nazwiska oraz marka samochodu *warszawa*, odpowiednio (por. rys. 22);
— w *Słowosieci* znajduje się podzbiór interpretacji z tabeli, np. *Podhale* jako region pojawia się w obu źródłach, w tabeli dodatkowo interpretowane jest jako klub piłkarski; *Mars, Merkury, Wenus* w *Słowosieci* zostały uznane wyłącznie za planety, w tabeli interpretowane są także jako bóstwa (a *Mars* dodatkowo jako baton);

Ciekawy przypadek stanowi ulica *Nowy Świat*. W *Słowosieci* jest to określenie *Ameryki*, w tabeli jest nazwisko *Świat* uzyskiwane przez heurystykę centrum; żadna z tych interpretacji nie jest adekwatna.

Spora liczba nazw homonimicznych z wyrazami pospolitymi w ogóle w tabeli się nie pojawiła (nazwiska *Chmura, Gorczyca, Kruczek, Pióro, Srebro*, imiona *Malwa* itp.), gdyż ich obecność w korpusie w ogóle nie została wykryta.

Do pewnego stopnia niezależny rozwój obu źródeł spowodował, że niektóre nazwy (np. *Chopin, Jezus, Wielkopolska*) są poprawnie interpretowane w obu. Prowadzi to do redundancji bez wystąpienia błędu (por. rys. 8).

W *Słowosieci* występuje wiele jednostek o lematach jednoliterowych, takich jak witaminy (*A, B, C, D, E, K*), symbole pierwiastków (*węgiel C, wodór H, potas K, sód N, tlen O*), jednostki fizyczne (*jednostki temperatury w skali Celsjusza C, Fahrenheita F czy Kelvina K, jednostka siły newton N*). Jako że inicjały składają się z dwóch segmentów: litery i kropki, heurystyka centrum prowadzi do dość zaskakującej i nie da się ukryć niepoprawnej interpretacji (por. rys. 23).

Heurystyka zmiany kaszty stosowana była do pierwszego dopasowania. Miało to na celu zmniejszenie wieloznaczności, ale mogło też być potencjalnym źródłem błędów. Na przykład w *Słowosieci Kościół* oznacza instytucję, zaś *kościół* — świątynię. W większości przypadków da to pozytywny wynik, ale na przykład nazwa *Kościół Mariacki* zostanie zinterpretowana błędnie.

We frazach typu *św. Piotr*, *święty Piotr* centrum stanowi imię *Piotr*, przez co heurystyka „imię i nazwisko" uzna to za człowieka ze względu na miano, a nie świętego.

Heurystyka centrum zawodzi natomiast wówczas, gdy właściwą interpretację nazwy stanowi jednostka wieloczłonowa — takie jednostki zawsze dobierane były ręcznie.

[16] Ponieważ miasto *Łódź* jest reprezentowane bezpośrednio w *Słowosieci*, kwestia ta w praktyce nie wystąpiła.

```
<named name_id="named_128.16-s_n1">
    <nambase>Warszawa</nambase>
    <nkjp_type type="placeName" subtype="settlement"/>
    <plwn_units part="whole" case_agreement="Full" polysemy="false">
        <unit luid="n1-sv1" status="auto" chosen="true">
            <lubase>Warszawa</lubase>
            <lusense>1</lusense>
            <luident>58505</luident>
            <synset>39237</synset>
        </unit>
    </plwn_units>
    <plwn_types part="whole" case_agreement="FirstLower"
                polysemy="false">
        <type typid="n1-tv1" status="auto" chosen="false">
            <nbase>warszawa</nbase>
            <plwntype>7462</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 22. Przykład nazwy własnej, której odmienne interpretacje występują w obu zbiorach danych

```
<named name_id="named_18.44-s_n3">
    <nambase>N.</nambase>
    <nkjp_type type="persName" subtype="surname"/>
    <plwn_units part="head" case_agreement="Full" polysemy="false">
        <unit luid="n3-sv1" status="auto" chosen="false">
            <lubase>N</lubase>
            <lusense>1</lusense>
            <luident>55236</luident>
            <synset>5246</synset>
        </unit>
    </plwn_units>
    <plwn_types part="whole" case_agreement="irrel" polysemy="irrel">
        <type typid="n3-tv1" status="added" chosen="true">
            <nbase>N.</nbase>
            <plwntype>200004</plwntype>
        </type>
    </plwn_types>
</named>
```

Rysunek 23. Przykład błędnej interpretacji inicjału za pomocą heurystyki centrum

W wypadku wieloznaczności stosowana była heurystyka dopasowania interpretacji słowosieciowej do typów i podtypów NKJP. Heurystyka ta działała gorzej dla synsetów niż dla nazw z listy ze względu na niedostateczne powiązanie hiperonimów wykorzystywanej wersji Słowosieci.

Na koniec warto zaznaczyć, że niepoprawną interpretację mogą powodować także błędy w znakowaniu NKJP. Błędna może być identyfikacja nazwy. Na przykład w zdaniu *Sam Mendel to prosty Żyd.* wyraz *Sam* został uznany za imię, choć to przymiotnik (*Mendel* za nazwisko). Czasem błędna jest jedynie kategoryzacja. Na przykład w zdaniu *Rybnik błyskawicznie złożył trzy projekty i pieniądze te otrzymał.* nazwa *Rybnik* została uznana za orgName, a nie settlement.

6. Znakowanie semantyczne wyrazów pospolitych

Podstawowym zadaniem nie jest reinterpretacja nazw własnych w oparciu o *Słowosieć*, lecz przypisanie jednostek *Słowosieci* wszystkim wyrazom pospolitym (rzeczownikom, czasownikom i przymiotnikom) występującym w zdaniach posiadających drzewa rozbioru w *Składnicy frazowej*. Znakowanie nazw własnych dla *Składnicy frazowej 0.5* zostało już przeprowadzone. Natomiast właściwe znakowanie wyrazów pospolitych dopiero ma się rozpocząć. W niniejszym rozdziale omówione zostaną pokrótce zasady, którym znakowanie to ma podlegać.

W przeciwieństwie do nazw własnych, standardowej interpretacji semantycznej podlegać będą wyłącznie pojedyncze segmenty, a znakowanie to przyporządkowywane będzie wierzchołkom terminalnym (potomkom nieterminali kategorii formarzecz, zaimos, psubst dla rzeczowników, formaczas dla czasowników oraz formaprzym, padj dla przymiotników)[17].

Zadanie realizowane będzie za pośrednictwem dedykowanego programu *Semantykon*.

6.1. Składnia xml-owa

W wyniku znakowania semantycznego *Składnicy frazowej* pliki XML zawierające upakowane lasy drzew rozbioru są rozbudowane na dwóch poziomach. Po pierwsze, dodawane są informacje dotyczące całego zdania. Z drugiej strony, dodawana jest informacja dotycząca znakowania poszczególnych nazw własnych (element `named`, patrz punkt 5.1) oraz wyrazów pospolitych (element `plwn_interpretation`, patrz poniżej).

[17] Zaimki osobowe zaimos, rzeczownikowe psubst i przymiotnikowe padj będą zazwyczaj interpretowane za pomocą anafory, por. punkt 6.2.1.

6.1.1. Informacje ogólne

Podczas znakowania nazw własnych do elementu `forest` został dodany atrybut `plwn_version` o wartości `plwn-1-6` oznaczający, że identyfikatory jednostek leksykalnych i synsetów pochodziły ze *Słowosieci* 1.6. Znakowanie wyrazów pospolitych będzie już realizowane za pomocą *Słowosieci* 1.8 (znakowanie nazw własnych zostanie wpierw zaktualizowane, por. rozdz. 8). Ze względu na dalszy rozwój *Słowosieci* atrybut ten zostanie zastąpiony trzema:

— `named_plwn_ver` wskazujący wersję *Słowosieci* za pomocą której realizowane było znakowanie nazw własnych,
— `sense_plwn_ver` wskazujący wersję *Słowosieci* za pomocą której realizowane było znakowanie wyrazów pospolitych,
— `updated_plwn_ver` wskazujący wersję *Słowosieci*, do której została dokonana aktualizacja.

Ponadto zostanie dodany specjalny element `sem_answers`, będący odpowiednikiem elementu `answer-data` (por. punkt 3.1). Atrybut `type` będzie posiadać następujące wartości:

— FULL oznaczający bezproblematyczną realizację zdania znakowania semantycznego,
— SYNTAX oznaczający błędnie dobrane drzewo rozbioru,
— MISSING_LEMMA oznaczający brak lematu któregoś z segmentów w *Słowosieci*,
— MISSING_SENSE oznaczający brak w *Słowosieci* jednostki leksykalnej poprawnie identyfikującej któryś z segmentów,
— IGNORED_NAME oznaczający nieoznakowaną nazwę własną,
— ANAPHORA oznaczający wystąpienie anafory,
— ELIPSIS oznaczający wystąpienie elipsy,
— PRED_NOUN oznaczający pojawienie się w zdaniu rzeczownika predykatywnego,
— PRED_ADJ oznaczający pojawienie się w zdaniu przymiotnika predykatywnego.

Jako że wymienione zjawiska nie wykluczają się wzajemnie, odpowiedź może stanowić ich listę. Jedyny wyjątek stanowi odpowiedź FULL, która oznacza, że żadne z powyższych zjawisk nie wystąpiło. Nawet błędny rozbiór zdania (SYNTAX) nie stanowi przesłanki do zaprzestania znakowania semantycznego, gdyż znakowanie takie może zostać w przyszłości przeniesione na skorygowane drzewo rozbioru (por. rozdz. 7).

6.1.2. Znakowanie segmentu

Wierzchołki terminalne reprezentujące segmenty interpretowane za pomocą *Słowosieci* zostaną rozbudowane o potomka o nazwie `plwn_interpretation`

posiadającego atrybuty sem_id złożone z prefiksu sem i kolejnej liczby dziesiątkowej (począwszy od 1) rozdzielonych znakiem podkreślenia. Jako że lemat interpretacji będzie z natury rzeczy tożsamy z lematem segmentu, nie ma powodu powtórnie go wypisywać.

Dla większości wyrazów, *Słowosieć* będzie zawierała jednostki leksykalne stanowiące poprawną interpretację danego segmentu w kontekście. Do ich reprezentacji służyć będzie, podobnie jak dla nazw, element plwn_units. Dopuszczalna jest jednak sytuacja, w której żadna jednostka potomna tego elementu nie jest opatrzona atrybutem chosen. Poza elementem plwn_units, w wyjątkowych okolicznościach mogą pojawić się następujące typy elementów:
— other_units,
— derived_units,
— anaphora.
Elementy te będą posiadać listę argumentów zmodyfikowaną w stosunku do elementu plwn_units dla nazw własnych:
— atrybut part zostanie pominięty,
— atrybut relat będzie występował jedynie w elemencie other_units przyjmując wartości synonym, hypernym, multiunit,
— atrybut case_agreement będzie miał jedynie wartości true, false,
— atrybut polisemy będzie miał wartości true, false.
Element anaphora nie posiada atrybutu case_agreement, zaś other_units posiada go opcjonalnie, wyłącznie dla relat="multiunit", gdyż lemat synonimu, hiperonimu czy poprzednika anaforycznego nie ma nic wspólnego z lematem segmentu. Podobnie jak w wypadku nazw, wymienione elementy będą posiadały listę potomków unit z takim samym zestawem potomków oraz atrybutami:
— atrybut luid złożony z prefiksu tożsamego z wartością sem_id przodka oraz (po dywizie) liter sv po których następuje numer sensu (czyli wartość elementu lusense),
— opcjonalny atrybut chosen o wartościach true, close,
— opcjonalny atrybut update.
Dla zachowania unikalności identyfikatorów, prefiks atrybutu sem_id (i konsekwentnie luid) zostanie zmieniony na syn, hyp, mul dla elementu other_units w zależności od wartości atrybutu relat, der dla elementu derived_units oraz ana dla elementu anaphora[18].

6.2. Zasady znakowania semantycznego segmentów w *Składnicy Semantycznej*

Zdanie jest oznakowane poprawnie, jeżeli każdy wierzchołek terminalny reprezentujący dowolny segment rzeczownikowy, czasownikowy i przymiotnikowy

[18] Zmiana prefiksu wartości atrybutu sem_id ma miejsce tylko wówczas, gdy element plwn_units nie występuje, gdyż w *Słowosieci* brak jednostek o adekwatnym lemacie.

```
<plwn_interpretation sem_id="sem_5">
    <plwn_units case_agreement="true" polysemy="true">
        <unit luid="sem_5-sv1" chosen="true">
            <lubase>zbieg</lubase>
            <lusense>1</lusense>
            <luident>143884</luident>
            <synset>103314</synset>
        </unit>
        <unit luid="sem_5-sv1">
            <lubase>zbieg</lubase>
            <lusense>2</lusense>
            <luident>28750</luident>
            <synset>13274</synset>
        </unit>
    </plwn_units>
</plwn_interpretation>
```

Rysunek 24. Przykład znakowania niejednoznaczengo wyrazu pospolitego

posiada element potomny `plwn_interpretation`, w którym dokładnie jeden element `unit` jest opatrzony atrybutem `chosen="true"`. Wyjątek stanowią nazwy własne: są one interpretowane semantycznie jako całość, więc składające się nań segmenty być znakowane nie muszą. Jednak dla nazw mających charakter kompozycyjny (*Ministerstwo Nauki i Szkolnictwa Wyższego*) znakowanie segmentów składowych jest zalecane.

Element `plwn_units` jest obowiązkowy dla wszystkich segmentów, dla których w *Słowosieci* występuje przynajmniej jedna jednostka o lemacie tożsamym z lematem segmentu. Wartość atrybutu `polisemy` zależy od liczby jednostek (długości listy potomków) i wynosi `false` w wypadku listy jednoelementowej. Na liście jednostek (element `unit`) muszą wystąpić wszystkie jednostki o lemacie tożsamym z lematem segmentu:
— z zachowaniem zgodności kaszty (atrybut `case_agreement="true"`),
— z dowolną kasztą (atrybut `case_agreement="false"`).
Co najwyżej jeden z tych elementów może posiadać atrybut `chosen="true"`, jednak dowolna ich ilość może posiadać atrybut `chosen="close"` (oznaczający jednostkę bliską znaczeniowo wybranej).

Przykładowa interpretacja segmentu *zbieg* ze zdania (1) przedstawiona jest na rys. 24.

Potomkiem `plwn_interpretation` może być także element `other_units`, który pojawia się w wypadku braku jednostki (o takim lemacie jak znakowany segment) stanowiącej poprawną interpretację wyrazu w kontekście. Atrybut `relat` ma wówczas wartość:

```
<plwn_interpretation sem_id="sem_6">
    <plwn_units case_agreement="true" polysemy="false">
        <unit luid="sem_6-sv1">
            <lubase>koncert</lubase>
            <lusense>1</lusense>
            <luident>2725</luident>
            <synset>1325</synset>
        </unit>
    </plwn_units>
    <other_units relat="hypernym" polysemy="false">
        <unit luid="hyp_6-sv1" chosen="true">
            <lubase>utwór muzyczny</lubase>
            <lusense>1</lusense>
            <luident>17030</luident>
            <synset>6994</synset>
        </unit>
    </other_units>
</plwn_interpretation>
```

Rysunek 25. Przykład znakowania wyrazu pospolitego, dla którego brak poprawnej interpretacji w *Składnicy*

— `synonym`, jeśli jednostka o danym lemacie jest synonimem segmentu w kontekście użycia,
— `hypernym`, jeśli jednostka o danym lemacie jest hiperonimem segmentu w kontekście użycia.

Synonim lub hiperonim może być jednostka jedno- lub wieloczłonową. Przykład znakowania wyrazu *koncert* ze zdania (4) widnieje na rys. 25. Inne przykłady to znakowanie segmentu *CB-radio* przez *radio* czy segmentu *ciemnowiśniowy* przez *wiśniowy*.

Jak już pisałam, w *Słowosieci* znajduje się wiele jednostek wieloczłonowych. Jednostki te bardzo często pojawiają się w tekście w postaci ciągu segmentów. W takich sytuacjach segment stanowiący centrum danej frazy może zostać oznakowany za pomocą jednostki wieloczłonowej, ktorej lemat obejmuje lemat tego segmentu. Wówczas element `other_units` występuje z atrybutem `relat="multiunit"`. Przykładem takiej sytuacji jest fraza *prawem fizyki* ze zdania (3) (por. rys. 26). W tym wypadku atrybut `case_aggreement` dotyczy wyłącznie lematu znakowanego segmentu (*prawo*), a nie jednostki wieloczłonowej.

Zdarzają się także sytuacje, w których jednostki wieloczłonowe mają zastosowanie, chociaż „uzupełniające" składowe ich lematu w zdaniu nie występują. Taka sytuacja ma miejsce w zdaniach (1) i (2) dla wyrazu *sygnalizacja*, dla której najwłaściwszą interpretacją jest jednostka *sygnalizacja świetlna*. Zazwyczaj

w takich sytuacjach istnieje jednostka o wyjściowym lemacie jednoczłonowym (*prawo, sygnalizacja*) będąca hiperonimem (rzadziej synonimem) wybranej jednostki wieloczłonowej. Jest ona opatrywana atrybutem `chosen="close"`. Jednak nie zawsze tak jest. W zdaniu *Lubimy zaglądać do takich dużych centrów* segment *centrów* powinien być interpretowany jako *centrum handlowe*, które nie jest hiponimem żadnej jednostki leksykalnej o lemacie *centrum*.

Brakujące jednostki mogą dotyczyć wyrazów derywowanych od lematów uwzględnionych w *Słowosieci*. Są to

— zdrobnienia (*rączyna*),
— zgrubienia *łapsko, nochal*,
— pokrewieństwo: żona (*wójtowa, imperatorowa*), córka (*wójtówna*), syn (*kasztelanic*)
— przedrostek (*antysemicki, kryptokomunista, pseudoeuropejczyk, quasi-mocarstwowy*).

Dlatego element `other_units` posiada opcjonalne atrybuty `deriv_type` i `deriv_source`. Wartością pierwszego z tych atrybutów będzie odpowiednio demin, augment, cognat, prefix. Wystąpienie tych atrybutów wymusza pojawienie się dodatkowo elementu `derived_units` stanowiącego interpretacje wyrazu pierwotnego.

Przykładem wystąpienia derywatywu jest segment *imperatorowej* ze zdania *Przez burtę statku Imperatorowej zaczęły się wychylać różne postaci [...]*; jego interpretacja widnieje na rys. 27.

6.2.1. Segmenty specjalnej troski

Istnieje kilka rodzajów segmentów, które wymagają specjalnego traktowania. Wynika to z faktu, że ich interpretacji semantycznej nie da się standardowo i bezpośrednio ustalić przypisując im odpowiednie jednostki *Słowosieci*.

Pierwszą taką grupę stanowią gerundia i imiesłowy. Ich podstawową cechę stanowi fakt, że morfologicznie są one czasownikami (i lematyzują się do czasownika), zaś składniowo rzeczownikami (gerundia (ger — terminale potomne względem nieterminala formarzecz) oraz przymiotnikami (imiesłowy przymiotnikowe pact, ppas — terminale potomne względem nieterminala formaprzym). Wbrew nazwie, imiesłowy przysłówkowe (pcon, pant) są potomne względem nieterminala formaczas i nie zachowują się jak właściwe przysłówki. Ich znakowanie jak typowych czasowników jest więc w pełni uzasadnione.

Jako że gerundia i imiesłowy przymiotnikowe stanowią pewien specyficzny rodzaj derywatów, do ich interpretacji wykorzystamy element `derived_units`. Ponieważ sytuacja jest poniekąd przeciwna do „typowej" derywacji: dysponujemy lematem wyrazu źródłowego, a nie pochodnego, atrybuty `deriv_source` i `deriv_dest` zostaną użyte w nieco innej konfiguracji. Zawsze jednak właściwa interpretacja dotyczy derywatu.

```
<plwn_interpretation sem_id="sem_4">
    <plwn_units case_agreement="true" polysemy="true">
        <unit luid="sem_4-sv1">
            <lubase>prawo</lubase>
            <lusense>1</lusense>
            <luident>6460</luident>
            <synset>7310</synset>
        </unit>
        <unit luid="sem_4-sv2" chosen="close">
            <lubase>prawo</lubase>
            <lusense>2</lusense>
            <luident>50813</luident>
            <synset>52811</synset>
        </unit>
        <unit luid="sem_4-sv3">
            <lubase>prawo</lubase>
            <lusense>3</lusense>
            <luident>6461</luident>
            <synset>1311</synset>
        </unit>
        <unit luid="sem_4-sv4">
            <lubase>prawo</lubase>
            <lusense>4</lusense>
            <luident>20905</luident>
            <synset>8188</synset>
        </unit>
    </plwn_units>
    <other_units relat="multiunit" case_agreement="true"
                 polysemy="false">
        <unit luid="mul_4-sv1" chosen="true">
            <lubase>prawo fizyki</lubase>
            <lusense>1</lusense>
            <luident>6460</luident>
            <synset>7310</synset>
        </unit>
    </other_units>
</plwn_interpretation>
```

Rysunek 26. Przykład znakowania wyrazu pospolitego jednostką wieloczłonową

```
<plwn_interpretation sem_id="sem_6">
    <other_units relat="hypernym" polysemy="false"
                 deriv_type="cognat" deriv_source="imperator">
        <unit luid="hyp_6-sv1" chosen="true">
            <lubase>żona</lubase>
            <lusense>1</lusense>
            <luident>11751</luident>
            <synset>6303</synset>
        </unit>
    </other_units>
    <derived_units case_agreement="true" polysemy="false"
                   deriv_type="cognat" deriv_dest="imperatorowa">
        <unit luid="der_6-sv1" chosen="close">
            <lubase>imperator</lubase>
            <lusense>1</lusense>
            <luident>45586</luident>
            <synset>29133</synset>
        </unit>
        <unit luid="der_6-sv2">
            <lubase>imperator</lubase>
            <lusense>2</lusense>
            <luident>24112</luident>
            <synset>5985</synset>
        </unit>
    </derived_units>
</plwn_interpretation>
```

Rysunek 27. Przykład znakowania derywatywu, dla którego brak poprawnej interpretacji w *Składnicy*

```
    <plwn_interpretation sem_id="sem_2">
        <plwn_units case_agreement="true" polysemy="false"
                    deriv_type="ger" deriv_dest="funkcjonowanie">
            <unit luid="sem_2-sv1" chosen="match">
                <lubase>funkcjonować</lubase>
                <lusense>1</lusense>
                <luident>1824</luident>
                <synset>54227</synset>
            </unit>
        </plwn_units>
        <derived_units case_agreement="true" polysemy="false"
                       deriv_type="ger" deriv_source="funkcjonować">
            <unit luid="der_2-sv1" chosen="true">
                <lubase>funkcjonowanie</lubase>
                <lusense>1</lusense>
                <luident>126208</luident>
                <synset>91200</synset>
            </unit>
        </derived_units>
    </plwn_interpretation>
</node>
```

Rysunek 28. Znakowanie gerundium mającego poprawną interpretację w *Słowosieci*

W *Słowosieci* znajduje się bogaty zestaw gerundiów i znacznie skromniejszy zestaw imiesłowów przymiotnikowych (na dodatek słabo powiązanych relacjami). Przykładowa interpretacja segmentu *funkcjonowania* ze zdania *Rok funkcjonowania wyższych szczebli samorządów nie upoważnia jeszcze do ocen* (występujacego w *Słowosieci*) znajduje się na rys. 28. Natomiast przykładowa interpretacja segmentu *Prowadzący* ze zdania *Prowadzący suzuki grand vitara 48-letni Zbigniew S. [...] stracił panowanie nad kierownicą* nie posiadającego w *Słowosieci* interpretacji przymiotnikowej widnieje na rys. 29.

Drugi przypadek wymagający szczególnej uwagi stanowią zaimki. Poza segmentami kategorii gramatycznych ppron12 (o lematach *ja, ty, my, wy*) i ppron3 (o lemacie *on*), będące bezpośrednimi potomkami nieterminala kategorii zaimos), lecz także rzeczowniki (*ktoś, ktokolwiek, nikt, kto, któż, coś, cokolwiek, nic, co*). Pierwsze pięć określa obiekty osobowe i jako takie interpretowane jest jako osoba[19]. Natomiast interpretacja pozostałych jest w zasadzie dowolna i powinna być determinowana przez kontekst. W wypadku zaimków powinno być możliwe znalezienie poprzednika anaforycznego. Do interpretacji zaimków

[19] *Słowosieć* 1.8 zawiera jednostki o lematach *ktoś, ktokolwiek* będące hiponimami osoby, jednak już *nikt* występuje jedynie w znaczeniu *człowiek mało ważny*.

```
    <plwn_interpretation sem_id="sem_1">
        <plwn_units case_agreement="true" polysemy="true"
                    deriv_type="pact" deriv_dest="prowadzący">
            <unit luid="sem_1-sv1">
                <lubase>prowadzić</lubase>
                <lusense>1</lusense>
                <luident>93035</luident>
                <synset>66370</synset>
            </unit>
            ..........................
            <unit luid="sem_1-sv7" chosen="match">
                <lubase>prowadzić</lubase>
                <lusense>7</lusense>
                <luident>6597</luident>
                <synset>2630</synset>
            </unit>
            ..........................
            <unit luid="sem_1-sv12">
                <lubase>prowadzić</lubase>
                <lusense>12</lusense>
                <luident>6595</luident>
                <synset>66367</synset>
            </unit>
        </plwn_units>
        <other_units relat="synonym" polysemy="false"
                     deriv_type="pact" deriv_source="irrel">
            <unit luid="syn_1-sv1" chosen="true">
                <lubase>kierujący</lubase>
                <lusense>1</lusense>
                <luident>2527</luident>
                <synset>9586</synset>
            </unit>
        </derived_units>
    </plwn_interpretation>
</node>
```

Rysunek 29. Znakowanie imiesłowu przymiotnikowego bez interpretacji w *Słowosieci*

służy element `anaphora`. Posiada on dodatkowe atrybuty `ref_sent` i `ref_node` wskazujące wierzchołek (zazwyczaj segment reprezentujący wyraz pospolity, lecz może to być także nazwa własna) będący poprzednikiem. Pierwszy z nich posiada trzy specjalne wartości: `self` oznaczającą, że poprzednik znajduje się w tym samym zdaniu co zaimek, `context` oznaczający, że brak jest jednoznacznego poprzednika i przyjęta interpretacja została przyjęta ogólnie z kontekstu oraz `none` obsługujący sytuację, gdy dostępny kontekst nie pozwala na ustalenie interpretacji zaimka. Specjalną wartością drugiego jest `none`, przyjmowana w wypadku ogólnego kontekstu oraz w sytuacji, gdy zdanie zawierające poprzednik nie posiada poprawnego drzewa rozbioru, a w konsekwencji nie jest interpretowane semantycznie.

Jeśli poprzednik anaforyczny znajduje się w zdaniu posiadającym drzewo rozbioru (w najprostszym przypadku w tym samym zdaniu, co zaimek), to jest on zinterpretowany semantycznie, i zaimek dziedziczy jego interpretację semantyczną. Taka sytuacja ma miejsce w zdaniach (5) i (6) (w przykładach koreferencja zaznaczona jest za pomocą indeksów). Interpretacja nie jest kopiowania, tylko jest umieszczane odwołanie do wierzchołka, z którego pochodzi. Rys. 30 zawiera interpretację segmentu *jego* z przykładu (5) posiadającego poprzednik w tym samym zdaniu, zaś rys. 31 zawiera interpretację segmentu *im* ze zdania **(b)** przykładu (6), posiadającego poprzednik w zdaniu **(a)**. Natomiast interpretacja zaimka *ją* z przykładu (9), którego poprzednik jest nazwą własną, widnieje na rys. 32. W obu wypadkach wierzchołek referencyjny jest interpretowany semantycznie.

(5) *Jeszcze po południu zawieszony prezes*$_1$ *[...] nie wiedział, czy członkowie rady podejmą jakieś decyzje w jego*$_1$ *sprawie.*

(6) **(a)** *Demonstranci*$_1$ *mieli flagi Liberalno-Demokratycznej Partii Rosji[...].* **(b)** *Nie udało się im*$_1$ *rozwinąć transparentów.*

(7) **(a)** *Ksiądz*$_1$ *Stanisław jest w S. od roku 1992.* **(b)** *Po awanturze z rodziną K. wygrzebano przeciwko niemu*$_1$ *stare sprawy.*

Jeśli bezpośrednim poprzednikiem jest inny zaimek (lub elipsa, patrz punkt 6.3), nawiązanie realizowane jest do pierwszego odwołania do danego obiektu w kontekście (zazwyczaj jest nim interpretowalny wyraz pospolity; por. przykład (8), czasem nazwa własna).

Jeśli poprzednik anaforyczny znajduje się w zdaniu nie posiadającym drzewa rozbioru w *Składnicy*, interpretacja semantyczna musi być reprezentowana bezpośrednio. Jako że pierwsze zdanie z (7) takiego drzewa nie posiada (`<base-answer type="NO_TREE">`), dotyczy to interpretacji poprzednika anaforycznego segmentu *niemu* ze zdania drugiego (rys. 33). Pewien problem stanowią zaimki, których poprzednikiem anaforycznym jest zaimek (lub elipsa!)

```
<node nid="20" from="4" to="5" subtrees="1" chosen="true">
    ......................
    <plwn_interpretation sem_id="sem_3">
        <plwn_units case_agreement="true" polysemy="false">
            <unit luid="sem_3-sv1" chosen="true">
                <lubase>prezes</lubase>
                <lusense>1</lusense>
                <luident>36989</luident>
                <synset>23067</synset>
            </unit>
        </plwn_units>
    </plwn_interpretation>
 </node>
.........................
<node nid="105" from="18" to="19" subtrees="1" chosen="true">
    ......................
    <plwn_interpretation sem_id="sem_13">
        <anaphora ref_sent="self" ref_node="20"/>
    </plwn_interpretation>
 </node>
```

Rysunek 30. Przykład znakowania anaforycznego zaimka posiadającego poprzednik w tym samym zdaniu

```
<node nid="13" from="3" to="4" subtrees="1" chosen="true">
    ......................
    <plwn_interpretation sem_id="ana_2">
        <anaphora ref_sent="morph_162.48-s" ref_node="5"/>
    </plwn_interpretation>
 </node>
```

Rysunek 31. Przykład znakowania anaforycznego zaimka posiadającego poprzednik w innym zdaniu akapitu

```
<node nid="7" from="1" to="2" subtrees="1" chosen="true">
   <named name_id="named_30.38-s_n1">
       <nambase>Agatka</nambase>
       <nkjp_type type="persName" subtype="forename"/>
       <plwn_types part="whole" case_agreement="Full" polysemy="false">
           <type typid="n1-tv1" status="auto" chosen="true">
               <nbase>Agatka</nbase>
               <plwntype>200002</plwntype>
           </type>
       </plwn_types>
   </named>
</node>
.....................................
<node nid="27" from="6" to="7" subtrees="1" chosen="true">
   <plwn_interpretation sem_id="ana_4">
       <anaphora ref_sent="self" ref_node="7"/>
   </plwn_interpretation>
</node>
```

Rysunek 32. Przykład znakowania anaforycznego zaimka, którego poprzednik jest nazwą własną

```
<node nid="57" from="8" to="9" subtrees="1" chosen="true">
    ..................
    <plwn_interpretation sem_id="ana_5">
        <anaphora ref_sent="morph_6.10-s" ref_node="none">
            <unit luid="sem_3-sv1" chosen="true">
                <lubase>ksiądz</lubase>
                <lusense>1</lusense>
                <luident>2972</luident>
                <synset>5943</synset>
            </unit>
        </anaphora>
    </plwn_interpretation>
</node>
```

Rysunek 33. Przykład znakowania anaforycznego zaimka posiadającego poprzednik w zdaniu bez drzewa rozbioru

w zdaniu, które nie posiada drzewa rozbioru: brak jest wierzchołka, do którego możnaby się odwołać, oraz lematu, który możnaby zinterpretować. W takiej sytuacji dodawana jest sztuczna jednostka o lemacie (element `base`) `on` (lub pustym w wypadku elipsy) oraz o jednostce i synsecie o identyfikatorze `0`. W ten sposób zaznaczamy, że interpretacja tych segmentów, jakakolwiek by nie była, musi być identyczna.

6.3. Elipsy

Reprezentacja wystąpień zjawiska elipsy i ich semantycznej interpretacji jest kwestią kontrowersyjną. Zważywszy jednak na fakt, że bank drzew będzie podstawą do tworzenia semantycznego słownika walencyjnego, jest ona wysoce uzasadniona.

Na poziomie semantycznym, interpretacja elipsy bardzo przypomina anaforę: trzeba znaleźć segment, mogący stanowić wypełnienie luki w strukturze semantycznej wypowiedzenia. Jest tylko jeden problem: elipsy nie występują na powierzchni, i nie są w żaden sposób w *Składnicy frazowej* reprezentowane (nie ma odpowiadających im wierzchołków). Dlatego muszą być reprezentowane jako oddzielne elementy drzewa rozbioru.

6.4. Składnia xml-owa

Do reprezentowania elipsy służy element `elipsis` będący bezpośrednim potomkiem elementu `forest`[20]. Element ten posiada trzy podstawowe atrybuty: `elipsis_id` będący identyfikatorem elipsy, `category` wskazujący kategorię gramatyczną elipsy oraz `parent` będący identyfikatorem wierzchołka, którego potencjalnym potomkiem (wymaganym) byłoby dane wystąpienie elipsy, gdyby zostało zrealizowane na powierzchni. Ponadto uwzględniany jest atrybut `case` dla opisu elipsy frazy rzeczownikowej lub przymiotnikowej. Dla reprezentacji elipsy „częściowej" wprowadzono opcjonalny atrybut `sibling`. Identyfikator elipsy składa się z prefiksu `eli` oraz kolejnej liczby naturalnej. Jedynym potomkiem elementu `elipsis` jest element `plwn_interpretation` wraz z potomkiem `anaphora`, analogiczne jak dla zaimków (por. punkt 6.2.1).

6.5. Zasady znakowania elips

Zaznaczane są wyłącznie elipsy występujące w całym zdaniu, a ich adekwatnym rodzicem jest zawsze wierzchołek kategorii `zdanie` znajdujący się najniżej w drzewie. Imiesłowy przysłówkowe oraz bezokoliczniki, będące potomkami

[20] Elementy te będą umieszczane na końcu, za wszystkimi elementami `node`.

```
<elipsis elipsis_id="eli_1" category="fno" case="mian" parent="1">
    <plwn_interpretation sem_id="ana_1">
        <anaphora ref_sent="morph_5.14-s" ref_node="16"/>
    </plwn_interpretation>
</elipsis>
```

Rysunek 34. Przykład reprezentacji elipsy posiadającej poprzednik w innym zdaniu

```
<elipsis elipsis_id="eli_1" category="fno" case="mian" parent="83">
    <plwn_interpretation sem_id="ana_1">
        <anaphora ref_sent="self" ref_node="7"/>
    </plwn_interpretation>
</elipsis>
```

Rysunek 35. Reprezentacja elipsy posiadającej poprzednik będący nazwą własną

wierzchołka fwe, systemowo nie posiadają podmiotu, więc jego brak nie może być uznany za elipsę. Mamy tu do czynienia z regularnym zjawiskiem kontroli, które zostanie rozwiązane na poziomie słownikowym, więc zrezygnowałam z oznaczania ich w korpusie.

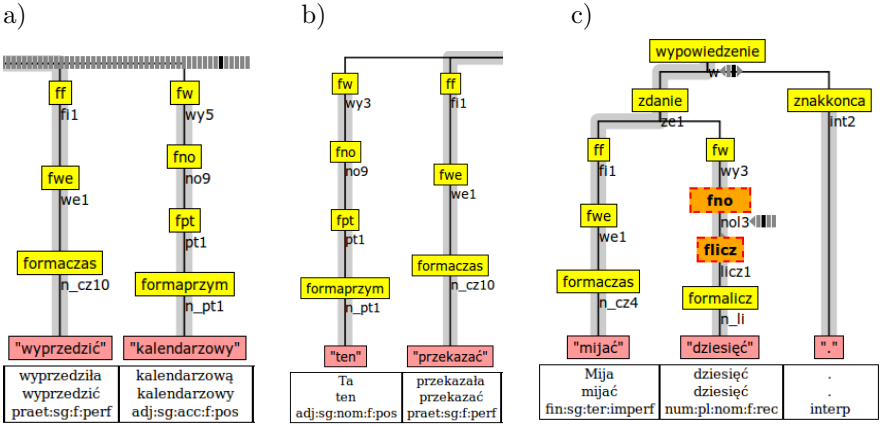
Zasady semantycznej interpretacji elips są analogiczne do interpretacji zaimków: należy powiązać ją z segmentem stanowiącym jej poprzednik w akapicie. W przykładach (8)–(10) wystąpienia elips oznaczone są przez $[\mathsf{el}]_i$. Interpretacja elipsy podmiotu w zdaniu **(b)** przykładu (8) widnieje na rys. 34, zaś interpretacja elipsy podmiotu zdania podrzędnego z przykładu (9) można obejrzeć na rys. 35 (por. też dotyczący tego samego zdania rys. 32).

(8) **(a)** *Paul Broca, lekarz*$_1$ *paryski, zachował mózgi*$_2$ *swoich pierwszych pacjentów dla potomności.* **(b)** *Po utrwaleniu ich*$_2$ *w formalinie* $[el]_1$ *przekazał je*$_2$ *do muzeum w Paryżu.* **(c)** *W 2007 roku neurolodzy*$_3$ *amerykańscy wyciągnęli je*$_2$ *z muzeum i* $[el]_3$ *poddali badaniu rezonansem magnetycznym.* **(d)** $[el]_3$ *Potwierdzili ścisłość obserwacji Broki.*

(9) *Agatka*$_1$ *mówiła, że to ją*$_1$ *strasznie bolało, ale* $[el]_1$ *nie mogła nawet pisnąć...*

(10) **(a)** *Niemcy wyrzucają z mieszkania Konstantego Prusa*$_1$*.* **(b)** *Gdzie się* $[el]_1$ *podzieje? Jak* $[el]_1$ *uratuje resztki księgozbioru?*

W przeciwieństwie do anafory, reprezentacja elipsy musi zawierać informację charakteryzującą frazę, która została pominięta. Służą temu atrybuty `category` oraz, w wypadku frazy rzeczownikowej lub przymiotnikowej, `case`. Podobnie jak w wypadku anafory, segment determinujący kontekstowo znaczenie pominiętego składnika zdania może znajdować się w tym samym ((9), (11), (13)) bądź wcześniejszym ((8), (10), (12), (14)) zdaniu.



Rysunek 36. Przykładowe poddrzewa rozbioru fraz nominalnych realizowanych jako a), b) frazy przymiotnikowe c) frazy liczebnikowe

Przyjrzyjmy się jeszcze poniższym zdaniom. W przeciwieństwie do poprzednich przykładów, w których pewne pozycje składniowe nie były wypełnione, tym razem rzeczownikowe pozycje są wypełnione przez frazy przymiotnikowe ((11) i (12)) bądź liczebnikowe ((13) i (14)), co pokazują fragmenty drzew z rys. 36.

(11) *W tym roku prawdziwa zima*$_1$ *wyprzedziła kalendarzową*$_1$.

(12) **(a)** *Komitet Wierzycieli MAAD złożył doniesienie karne [...] w poznańskiej Prokuraturze*$_1$ *Wojewódzkiej.* **(b)** *Ta*$_1$ *przekazała sprawę do PW w Lesznie.*

(13) *Jedna osoba*$_1$ *zmarła, a pięć*$_1$ *ciężko zatruło się alkoholem niewiadomego pochodzenia.*

(14) **(a)** *Mija pięć minut*$_1$. **(b)** *Nie przychodzi.* **(c)** *Mija dziesięć*$_1$.

W takim wypadku nie mamy do czynienia z całkowitą elipsą, a jedynie z „częściową”. Jednak z punktu widzenia znakowania semantycznego, a zwłaszcza tworzenia semantycznego słownika walencyjnego, sytuacja nie jest zadowalająca, gdyż pozycja nie jest poprawnie zinterpretowana semantycznie: pozycja rzeczownikowa nie może mieć interpretacji przymiotnikowej, a liczebniki w ogóle nie są semantycznie interpretowane. Dlatego zdecydowałam się na reprezentowanie takich elips. Posiadają one dodatkowy atrybut `sibling` wskazujący frazę stanowiącą faktyczną realizację wymagania. Jej bezpośredni przodek jest najlepszym kandydatem na przodka frazy eliptycznej, i n i e jest to w tym wypadku fraza kategorii zdanie (w omawianych przypadkach jest to wierzchołek

```
<elipsis elipsis_id="eli_1" category="fno" sibling="67" parent="66">
    <plwn_interpretation sem_id="ana_1">
        <anaphora ref_sent="self" ref_node="50"/>
    </plwn_interpretation>
</elipsis>
```

Rysunek 37. Zapis elipsy frazy rzeczownikowej realizowanej jako przymiotnikowa

```
<elipsis elipsis_id="eli_1" category="fno" sibling="39" parent="38">
    <plwn_interpretation sem_id="ana_1">
        <anaphora ref_sent="morph_7.38-s" ref_node="14"/>
    </plwn_interpretation>
</elipsis>
```

Rysunek 38. Zapis elipsy frazy rzeczownikowej realizowanej jako liczebnikowa

nieterminalny kategorii fno). Przykład reprezentacji elipsy frazy nominalnej ze zdania 11 realizowanej jako fraza przymiotnikowa (*kalendarzową*) znajduje się na rys. 37. Rodzeństwem jest w tym wypadku fraza przymiotnikowa fpt.

Przykładowa interpretacja elipsy frazy nominalnej realizowanej jako liczebnikowa w zdaniu **(c)** przykładu 14 została zaprezentowana na rys. 38. Zauważmy, że w przykładach tych ominięty został argument `case`. W wypadku frazy realizowanej jako przymiotnikowa, rzeczownik uzgadnia przypadek z przymiotnikiem. Natomiast w wypadku frazy realizowanej jako liczebnikowa to liczebnik jest centrum składniowym frazy (centrum semantycznym jest w obu wypadkach rzeczownik), i to on wskazuje, w jakim przypadku występuje fraza jako całość. W obu wypadkach przypadek frazy nominalnej jest ustalony i uwzględniony w potomku typu rekcja nieterminala kategorii formaczas potomnego względem wierzchołka wskazywanego przez atrybut `parent`.

7. Przenoszenie znakowania semantycznego na nowe wersje *Składnicy*

Jak pisałam we wstępie, *Składnica* jest zasobem intensywnie rozwijanym. Co ważne, nie jest to rozwój przyrostowy, gdyż rozwijana jest także gramatyka GFJP (Świdziński, 1992; Świdziński i Woliński, 2009) będąca podstawą konstrukcji drzew rozbioru w *Składnicy*. Tak więc w kolejnych wersjach *Składnicy* nie tylko pojawią się nowe akapity, dotychczas nie rozważane, ale także:

1. identyczne poprawne drzewa rozbioru w zmienionym upakowanym lesie,
2. zmienione poprawne drzewa rozbioru,

3. poprawne drzewa rozbioru zdań, które wcześniej takowych nie posiadały,
4. brak poprawnego drzewa rozbioru zdań, które wcześniej takowe posiadały.

Zmiany te w wyraźny sposób wpływają na strukturę upakowanego lasu. Nie tylko zmiana drzewa wybranego jako poprawne (punkt 2), ale już sam zmiana składu lasu (punkt 3) powoduje, że nie zostaje zachowana numeracja wierzchołków (atrybut `nid` elementu `node`). Powstała więc konieczność opracowania modułu przenoszącego znakowanie z wcześniejszej wersji *Składnicy* na nowszą w taki sposób, by ograniczyć konieczność ingerencji ludzkiej do minimum.

Podstawowy mechanizm tego przenoszenia wykorzystuje fakt, że atrybuty `from` i `to` wyznaczające granice frazy reprezentowanej przez wierzchołek zostają zachowane. Co więcej, w każdym drzewie istnieje dokładnie jeden wierzchołek terminalny reprezentujący dany segment. Jako że znakowanie semantyczne wyrazów pospolitych (por. rozdz. 6) przypisywane jest terminalom, przeniesienie elementu `plwn_interpretation` nie powinno stanowić problemu.

Szczególnej uwagi wymagają jednak dwa specyficzne przypadki: anafora i elipsa. W obu wypadkach, o ile tylko poprzednik anaforyczny znajduje się w innym zdaniu, istnieje możliwość, że zdanie to straciło lub zyskało rozbiór. Jeśli zdanie straciło rozbiór, interpretacja poprzednika musi zostać przeniesiona z aktualizowanej wersji *Składnicy*. Jeśli zyskało, można podjąć próbę automatycznego odnalezienia w nim poprzednika anaforycznego (po jego oznakowaniu) — zarówno lemat, jak i sama interpretacja muszą się zgadzać. Jeśli zgodność taka istnieje dla dokładnie jednego segmentu, można automatycznie zastąpić element `plwn_units` odpowiednim odnośnikiem (atrybut `ref_node`), w przeciwnym razie niezbędna jest ingerencja anotatora. W wypadku elipsy istotną kwestią jest też przeniesienie elementu, pod który elipsa jest podczepiana. Jeśli nie uda się odnaleźć wierzchołka kategorii `zdanie` o takich samych granicach frazy, znów niezbędny będzie udział anotatora.

Nieco bardziej skomplikowana sytuacja występuje w wypadku przenoszenia znakowania nazw własnych (por. rozdz. 5). Wynika to z faktu, że element `named` jest przypisywany do wierzchołków nieterminalnych. Ogólna zasada przepisywania nazw zakłada więc zgodność nie tylko granic frazy reprezentowanej przez wierzchołek, ale i jej kategorii. Większości nazw jednoczłonowych jest przypisywana do bezpośredniego przodka terminala, więc w takim wypadku (mało prawdopodobna) niezgodność kategorii nie musi być sztywnym wymaganiem. Istniały jednak sytuacje, w których element `named` był przypisywany do wierzchołka wyżej położonego w drzewie (patrz str. 42), i tu, tak jak w wypadku nazw wieloczłonowych, potrzebna jest pełna zgodność.

Zmiana poprawnego drzewa może powodować nie tylko zniknięcie wierzchołka adekwatnego dla danej nazwy, ale w wypadku, gdy takowego brakowało (prefiks `subname`, `supername` atrybutu `name_id`[21]), także jego pojawienie się.

[21] Niektóre nazwy zostały całkowicie zignorowane.

Jako że domyślne granice takich nazw (wyliczane w procesie konwersji, por. punkt 5.2.1) nie są w obecnej wersji *Składnicy semantycznej* przechowywane, sprawdzenie, czy w zmienionym drzewie pojawiły się wierzchołki adekwatne dla takich nazw wymagałoby konfrontacji z podkorpusem milionowym NKJP. Dlatego taka kontrola musi zostać przeprowadzona ręcznie.

8. Aktualizacja znakowania semantycznego względem kolejnych wersji *Słowosieci*

Znacznie poważniejszą i trudniejszą kwestię stanowią zmiany wprowadzane do *Słowosieci*. Można je zaklasyfikować w następujący sposób:

1. dodanie nowego lematu,
2. dodanie nowej jednostki leksykalnej dla rozważonego lematu,
3. zmiana identyfikatora jednostki,
4. zmiana identyfikatora synsetu,
5. przesunięcie jednostki do innego synsetu,
6. usunięcie jednostki leksykalnej,
7. usunięcie synsetu,
8. zmiana relacji łączącej jednostki bądź synsety.

Aktualizację znakowania semantycznego można podzielić na dwie fazy: identyfikację zmian w *Słowosieci* oraz wprowadzanie ich do *Składnicy*.

8.1. Identyfikacja zmian wprowadzonych w *Słowosieci*

Twórcy *Słowosieci* na bieżąco monitorują zmiany wprowadzane do zasobu. Jednak do utworzenie kolejnej oficjalnej wersji *Słowosieci* towarzyszy tylko jeden plik kodujący zmianę numeracji znaczeń pomiędzy kolejnymi wersjami *Słowosieci* (por. rys 39). Nowa (po lewej stronie) i stara (po prawej) identyfikacja jednostki oddzielone są średnikiem, każda z nich składa się z lematu, kategorii gramatycznej (1 — czasownik, 2 — rzeczownik, 4 — przymiotnik) oraz numeru znaczenia oddzielonych kropkami. Zmianie ulega co najwyżej numer znaczenia, z pliku tego można więc bez trudu utworzyć 3 trzykolumnowe tabele ⟨lemat, stary numer sensu, nowy numer sensu⟩ dla każdej kategorii gramatycznej.

Posługując się adekwatnymi tabelami zmian, można wykryć w *Słowosieci* jednostki, dla których zmieniły się identyfikatory. W tym celu należy porównać pliki `jednostki.txt` obu wersji dla wszystkich kategorii gramatycznych; umożliwia to jednocześnie wykrycie usuniętych i dodanych jednostek (i całych lematów). W ten sposób tworzone są pliki `added-<pos>-entities.txt` i `deleted-<pos>-entities.txt` zawierające trzykolumnowe tabele ⟨lemat, znaczenie, identyfikator⟩ jednostek dodanych bądź usuniętych, `modified-<pos>-`

```
aparycja.2.1;aparycja.2.1
aparycja.2.2;aparycja.2.2
apteka.2.1;apteka.2.1
arabski.4.1;arabski.4.1
aranżować.1.2;aranżować.1.1
aranżować.1.1;aranżować.1.2
arbiter.2.2;arbiter.2.1
arbiter.2.1;arbiter.2.2
archaiczny.4.2;archaiczny.4.1
archaiczny.4.1;archaiczny.4.2
atrakcyjny.4.1;atrakcyjny.4.1
atrybut.2.1;atrybut.2.1
atrybut.2.3;atrybut.2.2
atrybut.2.2;atrybut.2.3
```

Rysunek 39. Fragment pliku kodującego zmianę numeracji znaczeń pomiędzy *Słowosiecią* 1.4 a 1.5

`entities.txt` zawierający czterokolumnową tabelę ⟨lemat, znaczenie, stary identyfikator, nowy identyfikator⟩[22]. Porównanie dwóch wersji plików `synsety.txt` (przy użyciu tabeli zmian identyfikatorów jednostek) umożliwia stworzenie pliku `different-<pos>-synsets.txt` zawierającego czterokolumnową tabelę ⟨lemat, znaczenie, stary synset, nowy synset⟩ zmian przynależności jednostek do synsetów. `<pos>` koduje `verbs, nouns` bądź `adjects` w zależności od kategorii gramatycznej.

Następnie wykrywane są synsety, które różnią się identyfikatorami przy tym samym składzie; tworzą one dwukolumnową tabelę przechowywaną na pliku `modified-<pos>-synsets.txt`. Pozostałe synsety, z których usunięto wszystkie jednostki wypisywane tworzą tabelę jednokolumnową na pliku `deleted-<pos>-synsets.txt`, zaś nowopowstałe — takąż tabelę na pliku `added-<pos>-synsets.txt`. Informacja ta jest potrzebna ze względu na fakt, że część nazw interpretowana jest za pomocą typu słowosieciowego reprezentowanego przez identyfikator synsetu. Jako że informacja dotycząca zmian ma być nie tylko wykorzystywana w procesie automatycznej aktualizacji *Składnicy semantycznej*, lecz także przeglądana przez człowieka, istnieje możliwość utworzenia wersji plików zawierających wyłącznie identyfikatory synsetów oraz uzupełnionych o składające się nań jednostki. W tym drugim wypadku jednostki usunięte bądź dodane oznaczane są odpowiednio przez '*' lub '+'.

Tabele takie tworzone są dla wszystkich jednostek (synsetów) bądź wyłącznie dla tych, które zostały wykorzystane w *Składnicy*.

Usunięcie jednostki leksykalnej spowodowane jest zazwyczaj zbytnim rozdrobnieniem znaczeń; dlatego jest jej przypisywana lista jednostek (zazwyczaj jednoelementowa), z którymi usuwana jednostka została „zunifikowana". Lista taka tworzona jest automatycznie na podstawie *Składnicy semantycznej* poprzez powiązanie jednostek interpretujących konkretne segmenty z wystąpienia-

[22] Tabele zawierają zawsze stary numer znaczenia. Nowy uzyskiwany jest poprzez wspomnianą tabelę.

```
Czarny          1    2,38796,11372          pałac          2    1
fundusz         2    2,357359,226986        spółdzielnia   2    1
klub sportowy   1    1,357363,226991        towarzystwo    1    2
ministerstwo    1    3
```

Rysunek 40. Fragment pliku kodującego zastępowanie jednostek usuniętych bądź brakujących pomiędzy *Słowosiecią* 1.6 a 1.8

mi oznakowanymi `chosen="true"` i `chosen="false"`. Uporządkowanie listy determinowane jest w pierwszej kolejności frekwencją współwystępowania, a następnie rangą znaczenia. W rezultacie tabela `deleted-<pos>-entities.txt` uzupełniana jest o listę numerów znaczeń (zakładamy, że lemat nie ulega zmianie). Uwzględniona została także sytuacja, w której usunięta jednostka została zastąpiona jednostką nowododaną[23] (z pliku `added-<pos>-entities.txt`). Wówczas w miejsce numeru znaczenia wstawiana jest trójka ⟨znaczenie, identyfikator, synset⟩. W powstałej tabeli `suggested-<pos>-entities.txt` uwzględniane są (na podobnej zasadzie) jednostki, których brakowało we wcześniejszej wersji *Słowosieci* (były zastąpione atrapami o numerze jednostki 0 i sztucznie wysokim numerze synsetu). Fragment takiej tabeli znajduje się na rys. 40. Numer znaczenia jednostki o lemacie *Czarny* uległ zmianie; jednostki o lematach *fundusz* i *klub sportowy* zostały dodane.

Jako że w *Słowosieci semantycznej* zostały dotychczas oznakowane wyłącznie nazwy własne, mechanizm wykorzystania synonimów bądź hiperonimów jednostek nie został jeszcze zaimplementowany. W przyszłości planowane jest uzupełnienie jednostek dodanych z pliku `added-<pos>-entities.txt` o listę ich synonimów lub, jeśli jednostka tworzy synset jednoelementowy, o listę jej bezpośrednich hiperonimów. Ma to umożliwić automatyczne zastępowanie synonimów/hiperonimów adekwatnymi jednostkami.

8.2. Przeniesienie zmian wykrytych w *Słowosieci* do tabeli nazw

Do obsługi elementów `plwn_types` interpretujących nazwy własne niezbędna jest także lista zmian identyfikatorów synsetów. Poza tabelą nazw, istnieje lista wszystkich synsetów, które zostały w niej wykorzystane.

W *Słowosieci* są one traktowane jako zbiory jednostek. Usunięcie pojedynczej jednostki z synsetu zmienia jego tożsamość. Dlatego zmiany dotyczące synsetów można podzielić na dwie grupy:

1. zmiana identyfikatora synsetu bez zmiany jego składu (ew. z dodanymi całkiem nowymi jednostkami),

[23] Zdarzyły się takie przypadki. Ich celem była prawdopodobnie zmiana kolejności znaczeń.

```
20
ADD:    przedsiębiorstwo:1,
DEL:    zakład:1,
SAME:   biznes:2,  business:1,  firma:1,  interes:3,

597
DEL:    medyk:1,
SAME:   doktor:1,  lek.:1,  lekarz:1,

740
ADD:    dzielnica miasta:1+,
DEL:    dystrykt:1,
SAME:   dzielnica:1,  rejon:2,

871
ADD:    jednostka:7+,
SAME:   formacja:1,  oddział:1,  poczet:2,  związek taktyczny:1,

11427
DEL:    istota fantastyczna:2*,
```

Rysunek 41. Fragment pliku kodującego zmiany składu synsetów zachodzące pomiędzy *Słowosiecią* 1.6 a 1.8

2. zmiana składu synsetu:
   — synsety usunięte (brak jednostek),
   — synsety zmodyfikowane (dodane i usunięte jednostki, lecz przynależność przynajmniej jednej jednostki pozostała bez zmian),
   — synsety dodane (nowy identyfikator lub istniejący z zupełną zmianą składu).

W praktyce, każdemu synsetowi można przypisać trzy listy, z których przynajmniej jedna musi być niepusta: dodanych, usuniętych i zgodnych jednostek. Identyfikuje to sytuacje wymienione powyżej. Synsety z pierwszej grupy mogą być przetwarzane automatycznie (przy wykorzystaniu tabeli z pliku `modified-<pos>-synsets.txt`). Pozostałe wymagają niestety ręcznego przejrzenia. W tym celu tworzona jest dodatkowa tabela synsetów (plik `synset-<pos>-changes.txt`), których egzemplarze w tabeli nazw muszą być przejrzane ręcznie. Fragment takiej tabeli znajduje się na rys. 41.

Zmiany do tabeli nazw są wprowadzane ręcznie. Jednak w celu aplikacji tych zmian w *Składnicy* tworzone są dwie tabele: `suggested-<pos>-synsets.txt` zawierająca pary ⟨stary synset, nowy synset⟩ oraz tabela `suggested-<pos>-names.txt` zawierająca trójki ⟨lemat nazwy, stary synset, nowy synset⟩ dla tych

nazw z tabeli, których słowosieciowy typ semantyczny uległ w sposób indywidualny. Tabele te nie muszą być rozłączne ze względu na stary identyfikator synsetu; indywidualne sugestie z drugiej z tabel aplikowane są jako pierwsze. Przykładem takiej sytuacji jest nazwa *Paszczaczek*, której przypisany został synset *paszczak* w miejsce *istota fantastyczna*, której synset nota bene także uległ zmianie.

8.3. Wprowadzenie zmian wykrytych w *Słowosieci* do *Składnicy*

Relacje łączące jednostki i synsety nie są w *Składnicy semantycznej* przechowywane. Jednak wybór jednostki adekwatnie interpretującej dany segment jest dokonywany na podstawie tych relacji, i ich zmiana może ten wybór zaburzyć. W celu zminimalizowania niezbędnej ingerencji ludzkiej podczas zmiany wersji *Słowosieci* przyjęte zostało założenie, że numer znaczenia stanowi stabilną interpretację lematu[24] i że wszelkie modyfikacje relacji zachowują tę interpretację na przyjętym poziomie granulacji. Dotyczy to także synonimii, czyli przesunięcia jednostki do innego synsetu oraz zmiany identyfikatora synsetu[25]. Dlatego wykrywanie zmiany relacji pomiędzy dwiema jednostkami bądź synsetami nie zostało zaimplementowane. Zmiana synsetu zaimplementowana została, ponieważ są one w *Składnicy* przechowywane.

Zmiana identyfikatora jednostki jest operacją czysto techniczną. Tak więc zmiany 3–5 wprowadzane są w sposób w pełni automatyczny: dla każdej nazwy (element `named`) czy segmentu (element `plwn_interpretation`) podmieniany jest identyfikator jednostki (element `luident`) lub synsetu (element `synset`) jednostki (element `unit`) o ustalonym lemacie (element `lubase`) i znaczeniu (element `lusense`). Powoduje to dodanie atrybutu `update` o wartości odpowiednio `unit`, `synset`.

Wystąpienia usuniętej jednostki (element `luident`) nie są ze *Składnicy* wyrzucane, tylko mają dodawany atrybut `update="deleted"`. Jeśli jednostka miała wartość atrybutu `chosen="true"`, jest on zmieniany na `old`. Następnie lista „zbliżonych" jednostek z tabeli `suggested-<pos>-entities.txt` jest unifikowana (przecinana) z jednostkami posiadającymi atrybut `chosen="close"`. Pierwszy element ze zunifikowanej listy ma zmieniany atrybut `chosen` na `true` oraz dodawany atrybut `update="close"`. Jeśli jest to jednostka, która pojawiła się dopiero w nowej wersji, dodawany jest nowy element `unit` (z atrybutem `update="added"`). Jeśli lista jest pusta, korekta musi być dokonana ręcznie (atrybut `update` uzupełniany jest o wartość `manual`).

[24] Sytuacja, w której została dokonana zmiana numeracji znaczeń miedzy wersjami *Słowosieci* jest obsługiwana automatycznie, patrz poniżej.

[25] Czyli technicznie rzecz biorąc przesunięcie wszystkich składających się nań jednostek do innego, pustego synsetu bez zmiany jego składu.

W wypadku wyrazów pospolitych jednostki dodane wykorzystywane będą także do korekty interpretacji tych segmentów (i nazw), których interpretacja się nie powiodła, i zostały oznakowane za pomocą synonimów bądź hiperonimów poprzez dodanie elementu `other_units` o wartości atrybutu odpowiednio `synonym` lub `hypernym`. Jako że anotatorzy będą mieli sporą dowolność w dobieraniu tychże, za poprawną interpretację uważa się nową jednostkę (o adekwatnym lemacie), jeśli

— lista synonimów zawiera:
  a. jednostkę wybraną jako synonim,
  b. hiponim jednostki wybranej jako hiperonim,
  c. jednostkę będącą kohiponimem jednostki wybranej jako synonim,
  d. jednostkę będącą kohiponimem jednostki wybranej jako hiperonim.

— lista bezpośrednich hiperonimów zawiera:
  a. jednostkę wybraną jako hiperonim,
  b. hiponim jednostki wybranej jako hiperonim,
  c. jednostkę wybraną jako synonim,
  d. jednostkę będącą kohiponimem jednostki wybranej jako synonim,
  e. jednostkę będącą kohiponimem jednostki wybranej jako hiperonim.

Powyższe heurystyki stosuje się w kolejności aż do znalezienia pasującej jednostki. Wówczas interpretacja uzupełniana jest o

— element `plwn_units` w wypadku, gdy nie brakowało jednostek o danym lemacie zawierający wszystkie jednostki o tym lemacie,
— elementy `unit` dla jednostek spełniających heurystykę.

Dodawane jednostki opatrywane są atrybutem `update="added"`. Jednostka spełnia heurystykę o najwyższym numerze znaczenia opatrywana jest atrybutem `chosen="true"`, pozostałe mają wartość tego atrybutu ustawioną na close. Zastąpione synonimy i hiperonimy także nie są ze *Składnicy* usuwane, tylko uzupełniane o atrybut `update="supplemented"`.

W wypadku fiaska niezbędna jest ingerencja anotatora[26]. Podobnie jak w podstawowym trybie pracy, może on wybrać jedną jednostkę właściwą i dowolną liczbę bliskich znaczeniowo. Są one opatrywane atrybutem `update="added,manual"`, zaś wartość atrybutu `chosen` jest ustalana na true bądź close.

Podobna procedura stosowana jest dla nazw własnych posiadających element `plwn_units`. Dla jednostek usuniętych, ze zmienionymi identyfikatorami oraz przesuniętych do innych synsetów ta analogia jest pełna. Natomiast w wypadku aktualizacji jednostki będącej atrapą dodawany jest atrybut `update="supplemented"`. Jako że *Słowosieć* rozbudowywana jest także o nazwy własne, należy przyjrzeć się także i im. Automatyczna korekta dokonywana jest tylko przy pełnej zgodności lematów i przy braku jednostki (element `unit`) opatrzonej atrybutem `chosen="true"`. Z drugiej strony nazwa interpretowana powinna

[26] Jest rzecz jasna możliwe, że adekwatna jednostka nie została dodana.

być za pośrednictwem tabeli nazw (element `plwn_types`), a dodawana jednostka powinna być egzemplarzem bądź hiponimem jej (skorygowanego, patrz poniżej) typu semantycznego (element `type` opatrzony atrybutem `chosen="true"`). Jeśli nazwa (element `named`) nie zawierała elementu `plwn_units`, dodawane są wszystkie jednostki o tym lemacie, podobnie jak w wypadku wyrazów pospolitych.

Wszystkie elementy `type` (wybrane bądź nie) podlegają korekcie zgodnej z tabelą korekty tabeli nazw (por. punkt. 8.2). Korekta ta nie obejmuje jednak nazw s p o z a tabeli (por. punkt 5.2.2 str. 5.2.2). Nazwy interpretowane przez synsety, dla których ustalono automatyczną zmianę identyfikatora, także mogą być skorygowane automatycznie. Pozostałe takie nazwy muszą być przejrzane ręcznie.

Także w tym wypadku element `type` uzupełniany jest o atrybut `update`. Nadawana jest mu wartość modified przy automatycznej zmianie identyfikatora synsetu, supplemented przy dodaniu odpowiednika dla stworzonej przeze mnie atrapy, lemma przy zmianie właściwej dla pojedynczego lematu oraz general przy zmianie właściwej dla wszystkich lematów nazw danego typu. W dwóch ostatnich wypadkach atrybut może posiadać dodatkowy składnik deleted, jeśli stary synset został usunięty. Rzecz jasna dla poprawek dokonywanych ręcznie wartością tego atrybutu jest manual.

8.3.1. Analiza wyników

Omówione powyżej rozwiązania zostały zastosowane do wprowadzenia różnic pomiędzy *Słowosiecią* 1.6 a 1.8 do *Składnicy nazw*, czyli *Składnicy Frazowej* 0.5 zawierającej oznakowane nazwy własne (por. rozdz. 5). W *Składnicy nazw* znajduje się w sumie 4479 nazw, zwrotów obcojęzycznych i innych leksykalizacji (w tym 498 nazw przymiotnikowych). 1807 (322 przymiotnikowe) z nich ma przypisane jednostki, zaś 2874 (176) — typ. Oczywiście liczba wystąpień jednostek i typów jest większa ze względu na niejednoznaczność nazw.

Znakowanie nazw własnych jest na tyle regularne, że w *Składnicy nazw* wykorzystane zostało 957 jednostek rzeczownikowych (382 spośród nich zostało wybrane) oraz 83 jednostki przymiotnikowe (64 wybrane). Ponadto wykorzystane zostało 311 typów nazw rzeczownikowych (290 z nich zostało wybrane) oraz 25 typów przymiotnikowych (22 wybrane). Jest to rzecz jasna znikoma część zasobów *Słowosieci*. Dlatego tak istotna była opcja „przycięcia" zmian do jednostek i synsetów wykorzystanych w *Składnicy*. W tablicy 9 widnieje liczba zmian dotyczących jednostek i synsetów rzeczownikowych i przymiotnikowych, odpowiednio. Kolumna wszystkie oznacza wszystkie zmiany, jakich dokonano w *Słowosieci*, kolumna Składn. oznacza ograniczenie się do jednostek i synsetów, które zostały użyte w *Składnicy*, zaś kolumna wybr. oznacza jednostki i synsety,

		jednostki			synsety		
		wszystkie	Składn.	wybr.	wszystkie	Składn.	wybr.
rzeczowniki	added	10736	55	24	8416	0	0
	deleted	473	13	5	636	2	2
	changed	775	19	7	10407	25	23
	identifier	58	3	1	19	0	0
przymiotniki	added	8988	16	12	8063	0	0
	deleted	51	0	0	98	0	0
	changed	399	1	0	8593	3	3
	identifier	3	0	0	5	0	0

Tabela 9. Liczba zmienionych jednostek i synsetów pomiędzy wersjami *Słowosieci*

które zostały uznane w *Składnicy* za poprawne. Z kolei wiersz added oznacza jednostki bądź synsety dodane (w wypadku synsetów ograniczanie dotyczyło wystąpień lematów), deleted — usunięte, changed oznacza zmianę przynależności jednostki do synsetu lub zmianę składu synsetu, w końcu identifier oznacza zmianę identyfikatora jednostki bądź synsetu.

Jak widać, ograniczenie się do zmian dotyczących jednostek i synsetów, które wystąpiły w *Składnicy*, oznacza tak istotne zmniejszenie liczby zmian, że mogły one bez trudu być przejrzane ręcznie. W ten sposób opracowany został ręcznie zbiór sugerowanych zmian dla jednostek usuniętych oraz uzupełnionych atrap. Sugestie takie zaproponowano dla 7 jednostek i 11 synsetów (7 z nich to uzupełnione atrapy) rzeczownikowych oraz dla jednego synsetu przymiotnikowego będącego atrapą.

Niewielka skala zmian w *Słowosieci* przekłada się na niewielką liczbę modyfikacji, jakich trzeba było dokonać w *Składnicy nazw*. Na tym etapie udało się uniknąć ingerencji ręcznej. Dane na temat liczby automatycznie dokonanych modyfikacji znajdują się w tabeli tab:mod-treebank. Jako nazwy zmian posłużyły wartości atrybutu `update`[27]. Zauważmy, że zmianie identyfikatora jednostki zawsze towarzyszyła zmiana identyfikatora synsetu.

[27] Z tą różnicą, że dla elementu `plwn_types` atrybut ma wartość general, nie synset.

	rzeczowniki		przymiotniki	
	jednostki	synsety	jednostki	synsety
deleted	32	2	0	0
close	13	0	0	0
unit,synset	3	—	0	–
synset	74	25	2	0
supplemented	5	58	2	1

Tabela 10. Liczba modyfikacji jednostek i synsetów wprowadzonych do *Składnicy*

9. Podsumowanie

W raporcie opisane zostały zasady znakowania leksykalno-semantycznego banku drzew *Składnica* jednostkami leksykalnymi pochodzącymi ze *Słowosieci* oraz aktualizacji tego znakowania do zmian zachodzących w obu tych zasobach. Sam proces znakowania nie został jeszcze rozpoczęty; na razie jedynie zrealizowano znakowanie nazw własnych.

Proces przenoszenia nazw własnych z korpusu NKJP do *Składnicy* zakończył się sukcesem, a analiza błędów wskazuje, w jaki sposób należy poprawić wykorzystywane w nich heurystyki. Po pierwsze, warto szerzej wykorzystać znakowanie NKJP, przede wszystkim dla nazw rzeczownikowych, których nie udało się zinterpretować (tak jak to jest robione dla nazw przymiotnikowych), dla inicjałów oraz nazw osobowych lematyzowanych do wyrazów pospolitych (zmiana kaszty na dolną). Po drugie, dla wyrazów posiadających zarówno interpretację słowosieciową, jak i w tabeli nazw, warto wybierać interpretację uzyskaną za pomocą heurystyki o wyższym priorytecie.

Proces aktualizacji *Składnicy* do nowszych wersji *Słowosieci* można będzie ocenić dopiero w wypadku kompletnego znakowania leksykalno-semantycznego; na razie zakres wymaganych modyfikacji był zbyt ograniczony.

Bibliografia

E. Agirre, P. Edmonds (red.) (2006) *Word Sense Disambiguation. Algorithms and Applications*, t. 33 serii *Text, Speech and Language Technology*, Springer-Verlag, Dordrecht, Holandia.

E. Agirre, D. Martinez (2001) *Learning class-to-class selectional preferences*, w: *Proceedings of the Conference on Natural Language Learning*, s. 15–22, Tuluza, Francja.

S. Bergsma, D. Lin, R. Goebel (2008) *Discriminative learning of selectional preference from unlabeled text*, w: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, s. 59–68, Association for Computational Linguistics, Stroudsburg, PA.

C. Brockmann, M. Lapata (2003) *Evaluating and Combining Approaches to Selectional Preference Acquisition*, w: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, s. 27–34, Budapeszt, Węgry.

M. Derwojedowa, S. Szpakowicz, M. Zawisławska, M. Piasecki (2008) *Lexical units as the centrepiece of a wordnet*, w: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń (red.), *Proceedings of the Intelligent Information Systems XVI (IIS'08)*, Challenging Problems in Science: Computer Science, Akademicka Oficyna Wydawnicza Exit, Zakopane.

K. Erk (2007) *A Simple, Similarity-based Model for Selectional Preferences*, w: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, s. 216–223, Praga, Czechy.

C. Fellbaum (red.) (1998) *WordNet — An Electronic Lexical Database*, MIT Press, Cambridge, MA.

W. Gale, K. Church, D. Yarowsky (1992) *Estimating upper and lower bounds on the performance of word-sense disambiguation programs*, w: *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, s. 249–256, Newark, DL.

K. Głowińska, A. Przepiórkowski (2010) *The Design of Syntactic Annotation Levels in the National Corpus of Polish*, w: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, ELRA, Valetta, Malta.

E. Hajnicz (2011) *Automatyczne tworzenie semantycznego słownika walencyjnego*, Problemy Współczesnej Nauki. Teoria i Zastosowania: Inżynieria Lingwistyczna, Akademicka Oficyna Wydawnicza Exit, Warszawa.

M. Lapata (1999) *Acquiring lexical generalizations from corpora: a case study for diathesis alternations*, w: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, s. 397–404, College Park, MA.

D. McCarthy (2001) *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*, Rozprawa doktorska, University of Sussex.

F. C. N. Pereira, D. H. D. Warren (1980) *Definite clause grammars for language analysis*, Artificial Intelligence, t. 13, nr 3, s. 231–278.

M. Piasecki, S. Szpakowicz, B. Broda (2009) *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

A. Przepiórkowski (2006) *What to acquire from corpora in automatic valence acquisition*, w: V. Koseska-Toszewa, R. Roszko (red.), *Semantyka a konfrontacja językowa*, t. 3, Slawistyczny Ośrodek Wydwniczy & Fundacja Slawistyczna, Warszawa.

— (2009) *A comparison of two morphosyntactic tagsets of Polish*, w: V. Koseska-Toszewa, L. Dimitrova, R. Roszko (red.), *Proceedings of the 4th MONDILEX Open Workshop on Representing Semantics in Digital Lexicography*, s. 138–144.

A. Przepiórkowski, M. Bańko, R. L. Górski, B. Lewandowska-Tomaszczyk (red.) (2012) *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa.

A. Przepiórkowski, P. Bański (2009) *XML Text Interchange Format in the National Corpus of Polish*, w: S. Goźdź-Roszkowski (red.), *Practical Applications in Language Corpora (PALC'09)*, s. 55–65, Peter Lang, Frankfurt am Main.

P. Resnik (1993) *Selection and Information: A Class-Based Approach to Lexical Relationships*, Rozprawa doktorska, University of Pennsylvania, Filadelfia, PA.

F. Ribas (1994) *An Experiment on Learning Appropriate Selectional Restrictions from Parsed Corpus*, w: *Proceedings of the 15th International Conference on Computational Linguistics (COLING-1994)*, s. 769–774, Kioto, Japonia.

M. Świdziński (1992) *Gramatyka formalna języka polskiego*, Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

M. Świdziński, M. Woliński (2009) *A New Formal Definition of Polish Nominal Phrases*, w: M. Marciniak, A. Mykowiecka (red.), *Aspects of Natural Language Processing*, t. 5070 serii *Lecture Notes in Computer Science*, s. 143–162, Springer-Verlag.

Z. Vetulani, J. Walkowska, T. Obrębski, J. Marciniak, P. Konieczka, P. Rzepecki (2009) *An Algorithm for Building Lexical Semantic Network and Its Application to PolNet — Polish WordNet project*, w: Z. Vetulani, H. Uszkoreit (red.), *Human Language Technology. Chalenges of the Information Society. 3rd Language & Technology Conference*, t. 5603 serii *Lecture Notes in Artificial Intelligence*, s. 369–381, Springer-Verlag. Revised Selected Papers.

M. Woliński (2004) *Komputerowa weryfikacja gramatyki Świdzińskiego*, Rozprawa doktorska, Instytut Podstaw Informatyki, Polska Akademia Nauk, Warszawa.

— (2006) *Morfeusz — a Practical Tool for the Morphological Analysis of Polish*, w: M. A. Kłopotek, S. T. Wierzchoń, K. Trojanowski (red.), *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'06*, Advances in Soft Computing, s. 503–512, Springer-Verlag, Ustroń.

— (2011) *A Preliminary Version of Składnica — a Treebank of Polish*, w: Z. Vetulani (red.), *Proceedings of the 5th Language & Technology Conference*, s. 299–303, Poznań.

A. Wróblewska, M. Woliński (2011) *Preliminary Experiments in Polish Dependency Parsing*, w: P. Bouvry, M. A. Kłopotek, F. Leprevost, M. Marciniak, A. Mykowiecka, H. Rybiński (red.), *International Joint Conference on Security and Intelligent Information Systems*, t. 7053 serii *Lecture Notes in Computer Science*, s. 279–292, Springer-Verlag, Warszawa.

Pracę zgłosił Adam Przepiórkowski

Adres autorki: Elżbieta Hajnicz
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa
Polska
e-mail: Elzbieta.Hajnicz@ipipan.waw.pl

Symbol klasyfikacji rzeczowej: CR: I.2.7

Na prawach rękopisu
Printed as manuscript

Nakład 100 egzemplarzy. Papier kserograficzny klasy III. Oddano do druku w grudniu 2012. Wydawnictwo IPI PAN. ISSN 0138-0648.