Format wejścia w polsko-angielskim słowniku elektronicznym konstruowanym dla potrzeb tłumaczenia automatycznego

Format of the Entry into the Electronic Polish-English Dictionary designed for MT

Krzysztof Jassem

Uniwersytet im. Adama Mickiewicza, Poznań
Wydział Matematyki i Informatyki
jassem@math.amu.edu.pl.

STRESZCZENIE

W poniższej pracy omówiono format wejścia w słowniku polsko-angielskim konstruowanym dla potrzeb tłumaczenia automatycznego. Przedstawiona jest graficznie struktura logiczna hasła słownikowego oraz sposób zapisu słownika w pliku tekstowym w postaci dokumentu typu SGML.
W obecnym stanie badań przygotowano listy leksemów języka polskiego (patrz [1]), które mają być zawarte w słowniku. Planuje się zakończenie prac nad opisem haseł do końca 1998 roku.

ABSTRACT

The following paper presents a format of the entry into the Polish-English dictionary designed for MT. The logical structure of the entry is shown graphically. The representation of the dictionary in a text file as a SGML-type document is proposed.
In the current state of research, lists of Polish lexemes have been prepared in order to be included in the dictionary (see [1]). By the end of 1998 the work on describing the entries is scheduled to be concluded.

1. Wstęp

W pracy [2] opisano format wejścia w słowniku dla prototypowego systemu tłumaczenia automatycznego z języka polskiego na język angielski. Format ten był ściśle związany z algorytmem tłumaczenia, który dokonywał analizy leksykalnej w oparciu o informacje zawarte w słowniku. W słowniku zawartych było około 30000 form fleksyjnych wyrazów i fraz leksykalnych wygenerowanych z 2000 form kanonicznych wybranych w sposób niesystematyczny z różnych źródeł (słowników oraz tekstów informatycznych). Celem obecnie prowadzonych działań jest stworzenie polsko-angielskiego słownika elektronicznego o uniwersalnym charakterze. Uniwersalność słownika realizowana jest w dwóch płaszczyznach:

1) słownik ma umożliwiać tłumaczenie tekstów z różnych działów informatyki,

2) słownik może być wykorzystywany w różnych systemach tłumaczenia maszynowego.

Pierwszy z powyższych postulatów realizowany jest poprzez skonstruowanie słownika w oparciu o analizę korpusu tekstów informatycznych wybranych z szerokiego wachlarza źródeł. Analiza korpusu tekstów wydaje się niezbędna (alternatywą może wydawać się skorzystanie z istniejących słowników "papierowych"), gdyż w słowniku skonstruowanym dla potrzeb tłumaczenia automatycznego zawarte powinny być wszelkie wyrazy i frazy leksykalne występujące w tekstach informatycznych, także takie, które nie należą do leksykonu informatycznego. Wielkość zanalizowanego korpusu (patrz [1]) implikuje określoną wielkość słownika (około 10000 leksemów).

Spełnienie drugiego postulatu ma być możliwe dzięki następującym dwóm rozwiązaniom:

- format tekstowy słownika (format binarny przedstawiony jest w pracy [3]) zgodny jest z popularnym standardem dokumentów SGML (min. standard HTML, w którym tworzone są internetowe strony www, jest zgodny z notacją SGML),
- kod wartości kluczowego pola, nazwanego *Complementation* jest opisany przy pomocy gramatyki bezkontekstowej, co sprawia, że jest łatwy do parsowania.

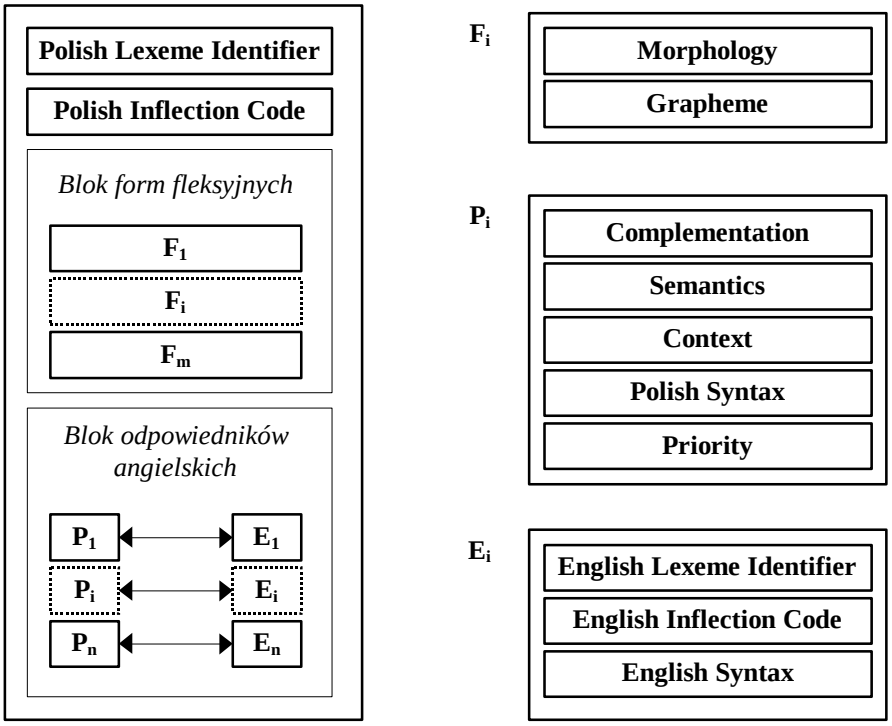
2. Format tekstowy słownika

Hasłem (wejściem) w słowniku jest leksem frazy słownikowej języka polskiego. Definicja frazy słownikowej podana jest w [3]. W szczególności fraza słownikowa może być pojedynczym wyrazem.

Hasła homograficzne (hasła o różnym znaczeniu reprezentowane przez taką samą formę kanoniczną) rozdzielane są tylko wtedy, gdy różnią się przynajmniej jedną formą fleksyjną.

2.1. Graficzna reprezentacja formatu hasła w słowniku

Na *rysunku 1* przedstawiona jest struktura logiczna hasła w słowniku.



Rysunek 1. Graficzna reprezentacja formatu hasła w słowniku.

Charakterystyka każdego z pól przedstawionych na *rysunku 1* podana jest w paragrafie 3.

2.2. Reprezentacja słownika jako dokumentu typu SGML

Poniżej podana jest specyfikacja słownika jako dokumentu typu SGML.

```
<!-- POLENG / Polish-to-English Machine Translation / DTD -->

<!ENTITY        % doctype    "POLENG">

<!ELEMENT POLENG            O  O   (Dictionary | D)>

<!ELEMENT (Dictionary | D)  -  O   ((Lexeme | L)*)>
<!ATTLIST  (Dictionary | D)  name                CDATA      #IMPLIED
                             version             CDATA      #IMPLIED
                             authors             CDATA      #IMPLIED
                             updated             CDATA      #IMPLIED>

<!ELEMENT (Lexeme | L)      -  O   ((Form | F)*, (Translation | T)*)>
<!ATTLIST  (Lexeme | L)      id                  ID         #REQUIRED
```

```
                          polishInflection  CDATA      #REQUIRED>

<!ELEMENT (Form | F)       -  O  (#PCDATA)>
<!ATTLIST  (Form | F)      morphology        CDATA      #REQUIRED>

<!ELEMENT (Translation | T) -  O  (#PCDATA)>
<!ATTLIST  (Translation | T) complementation CDATA      #IMPLIED
                          semantics         CDATA      #IMPLIED
                          context           CDATA      #IMPLIED
                          polishSyntax      CDATA      #IMPLIED
                          priority          CDATA      #IMPLIED
                          englishInflection CDATA      #IMPLIED
                          englishSyntax     CDATA      #IMPLIED>
```

Rysunek 2. Specyfikacja słownika jako dokumentu w formacie SGML

W dalszej części pracy, gdy mowa będzie o strukturze logicznej słownika (*rysunek 1*), używane będą pojęcia: ramka, pole, wartość pola. W przypadku, gdy słownik rozpatrywany będzie jako dokument w formacie SGML (*rysunek 2*), wykorzystywane będą pojęcia: element, atrybut, wartość elementu, wartość atrybutu.

Przykładowo ramce F_i struktury logicznej odpowiada element *Form* dokumentu SGML. Pole *Morphology* ramki F_i odpowiada atrybutowi *Morphology* elementu *Form*, (wartość pola *Morphology* ramki F_i odpowiada wartości atrybutu *Morphology* elementu *Form*), a wartość pola *Grapheme* ramki F_i reprezentowana jest przez wartość elementu *Form*.

3. Charakterystyka poszczególnych pól słownika.

3.1. Pole *Polish Lexeme Identifier*

Wartością tego pola jest forma kanoniczna leksemu, tj. forma mianownika liczby pojedynczej (mianownika liczby mnogiej dla *plurale tantum*) dla rzeczowników i fraz rzeczownikowych, bezokolicznik dla czasowników i fraz czasownikowych, mianownik liczby pojedynczej rodzaju męskiego dla przymiotników i fraz przymiotnikowych.

3.2. Pole *Polish Inflection Code*

Wartością tego pola jest kod "komputerowego paradygmatu odmiany". Na podstawie tego kodu możliwe jest automatyczne wygenerowanie form fleksyjnych leksemu z formy kanonicznej bez pomocy kompetencji ludzkiej. Wartość tego pola pozwala również określić, jaką część mowy reprezentuje leksem. Dla leksemów czasownikowych pierwszą literą kodu jest *C*, dla leksemów rzeczownikowych pierwszą literą kodu jest *R*, itd.

3.3. Blok form fleksyjnych

Blok form fleksyjnych może składać się z różnej liczby elementów. Dla niefleksyjnych części mowy blok składa się z jednego elementu. Dla rzeczowników (i fraz rzeczownikowych) blok ten składa się najczęściej z 14 elementów odpowiadających wszystkim przypadkom liczby pojedynczej i mnogiej. Dla przymiotników blok ten może składać się maksymalnie z 30 elementów, w tym również przysłówków odprzymiotnikowych (uwzględniono zjawisko synkretyzmu: te same formy przymiotnika odpowiadające różnym wartościom przypadka, rodzaju i liczby występują dokładnie raz). Dla czasowników blok form fleksyjnych może składać się maksymalnie z 58 elementów; z formy kanonicznej generowane są wszystkie formy imiesłowu biernego i czynnego oraz formy liczby pojedynczej odsłowników (rzeczowników odczasownikowych).

Każdy element bloku (ramka F_i) składa się z postaci grafemicznej danej formy fleksyjnej oraz zakodowanej informacji morfologicznej. Jeśli jednej formie leksemu odpowiadają dwa grafemy (np. dwa grafemy "rękami", "rękoma" odpowiadają formie narzędnika liczby mnogiej leksemu "ręka"), to każdy grafem jest reprezentowany przez osobny element.

3.4. Blok odpowiedników angielskich

Blok odpowiedników angielskich składa się z nieograniczonej liczby par ramek. Każda para ramek charakteryzuje jeden odpowiednik angielski wyrazu polskiego (frazy polskiej). Ramka P określa warunki wyboru odpowiednika opisywanego przez ramkę E. Ponadto w ramkach P i E zawarte są inne informacje, które mogą być istotne w analizie wyrażenia języka polskiego oraz syntezie wyrażenia języka angielskiego.

3.4.1. Pole *Complementation*

Pole to określa związki zdaniowe, które dane hasło tworzy z elementami podrzędnymi (modyfikatorami) w zdaniu polskim i angielskim. Przyjmuje się następujące założenie:

W polu *Complementation* umieszczona jest informacja tylko o takich modyfikatorach, które przetłumaczone kompozycyjnie dałyby błędny rezultat.

Takie podejście możliwe jest tylko dla słownika dwujęzycznego i algorytmu dokonującego tłumaczenia metodą transferu (a nie przy pomocy konstrukcji interlingwistycznych).

Przykład 1.

W zdaniu: "Tłumaczę teksty przy pomocy komputera" dopełnienie "przy pomocy komputera" może być tłumaczone kompozycyjnie - w oderwaniu od czasownika "tłumaczę". W polu *Complementation* przy opisie leksemu "tłumaczyć" nie wystąpi grupa przyimkowa "przy pomocy G" (*G - genitive* - oznacza grupę rzeczownikową w dopełniaczu). W zdaniu "Tłumaczę teksty na język angielski" dopełnienie "na język angielski" nie może być tłumaczone w

oderwaniu od czasownika, gdyż dałoby to efekt: “I translate texts on English” (“dyżurnym” tłumaczeniem przyimka “na” jest przyimek “on”) zamiast poprawnego zdania “I translate texts into English”.

Zgodnie z definicją podaną w [2] dopełnienia, które nie mogą być tłumaczone kompozycyjnie nazywane będą dalej *wymaganiami*.

Pole *Complementation* może być niepuste dla następujących części mowy (i odpowiadających im fraz leksykalnych):

- czasowniki,
- rzeczowniki (np. grupa przyimkowa “nad I” (I: instrumental) dla rzeczownika “praca”),
- przymiotniki (np. grupy przyimkowe “od G” dla przymiotników “zależny”, niezależny),
- przyimki (wartością pola C*omplementation* są przypadki frazy rzeczownikowej, z którą przyimek łączy się w grupę przyimkową, np. A | L (accusative, locative) dla przyimka “na”.

Istnienie pola *Complementation* ma następujące implikacje w algorytmie tłumaczenia:

- Możliwość poprawnej analizy wyrażenia języka polskiego, np. wyrażenie składające się z frazy czasownikowej i dopełnienia frazą rzeczownikową w bierniku powinno być analizowane jako poprawne tylko w przypadku, gdy informacja o takim typie dopełnienia zawarta jest w polu *Comlementation* czasownika konstytutywnego (wyrażenia typu “lubić coś” są analizowane jako poprawne w odróżnieniu od wyrażeń typu “rozmawiać coś”).
- Możliwość ustalenia w analizie wyrażenia wejściowego, czy dana fraza jest wymaganiem czasownika (modyfikuje grupę czasownikową), czy wymaganiem rzeczownika (modyfikuje grupę rzeczownikową).
- Możliwość wyboru najlepszego odpowiednika angielskiego, np. informacja o typie wymagania umożliwia poprawne przetłumaczenie frazy “powiedzieć coś komuś” na frazę “to tell smth to sone” (innym odpowiednikiem czasownika “powiedzieć” jest angielski czasownik “to say”, stosowany w innych konstrukcjach składniowych).
- Możliwość poprawnego przetłumaczenia modyfikatora na język angielski (patrz *przykład 1*).
- Możliwość ustalenia poprawnej kolejności modyfikatorów w zdaniu angielskim.

Opis własności wymagań podany w polu *Complementation* ma spełniać następujące 2 postulaty:

- wprowadzanie opisu przez leksykografa ma być czynnością możliwie mało czasochłonną,
- opis powinien być “dobrze” parsowalny przez komputer.

W celu zrealizowania pierwszego postulatu przyjęto, że kategorie gramatyczne będą kodowane możliwie krótkimi ciągami znaków. W celu zrealizowania drugiego postulatu przyjęto, że opis będzie zgodny z pewną gramatyką bezkontekstową.

W rezultacie zaproponowano sposób kodowania wartości pola *Complementation* zgodnie z poniższym schematem podanym w notacji Backhusa - Naura):

Język opisu wymagań w notacji Backusa-Naura.

Symbole nieterminalne podane są zwykłym drukiem, symbole metajęzyka notacji Backusa- Naura podane są czcionką pochyłą, symbole terminalne - czcionką pogrubioną. Za znakiem procentu znajduje się komentarz.

Complementation *::=* **‘’** % brak wymagań
Complementation *::=* Transfer

Transfer *::=* Source_Category → Target_Category
% każda kategoria wymagania skojarzona jest z odpowiadającą jej kategorią
% w języku angielskim
Transfer *::=* **[** Transfer **]**
% obligatoryjność wystąpienia wymagania w zdaniu polskim
Transfer *::=* **(**Transfer**)**
% priorytet (znak nawiasu “porządkuje” skomplikowane konstrukcje
% współzależności wymagań)
Transfer *::=* Transfer **|** Transfer *{***|** Transfer*}*
% dysjunkcja wymagań (możliwe jest wystąpienie dokładnie jednego
% z wymagań w zdaniu polskim)
Transfer *::=* **<** Transfer **,** Transfer *{***,** Transfer*}* **>**
%określony szyk wymagań w zdaniu polskim
Transfer *::=* **{** Transfer **,** Transfer *{***,** Transfer*}* **}**
%dowolny szyk wymagań w zdaniu polskim

Source_Category *::=* Noun_Category
Source_Category *::=* Source_Preposition Noun_Category
Source_Category *::=* **IN** | % bezokolicznik
GR | % odsłownik
AJ | % przymiotnik
AV | % przysłowek
TH | % zdanie podrzędne “że”
BY | % zdanie podrzędne “by”
JK | % zdanie podrzędne “jak”
OB | % inne zdanie podrzędne dopełnieniowe
LC | % okolicznik miejsca
% Przyjęto konwencję, że kody kategorii języka źródłowego są
% “w miarę możliwości spójne” z kodami kategorii języka docelowego,
% np. kategoria przymiotnika w języka polskim kodowana jest tak
% samo (*AJ*), jak kategoria przymiotnika w języku angielskim,
% a kategoria odsłownika (rzeczownika odczasownikowego) jest kodowana

% tak samo, jak kategoria odpowiadającej kategorii języka angielskiego (*GR*).
Noun_Category *::*= Noun_Case
Noun_Category *::*= Noun_Case **:** Semantic_Value

Noun_Case ::= **N** | **G** | **D** | **A** | **I** | **L**

Semantic_Value *::*= Semantic_Feature
Semantic_Value *::*= **-** Semantic_Feature

Semantic_Feature *::*= **Hum** | **Anim** | **Abstr**

Source_Preposition *::*= **od** | **do** | **z** |*…* % przyimki języka polskiego

Target_Category *::*= **NP** |
Target_Preposition **NP** |
GR | % gerund
TO | % forma "to" + infinitive
IN | % infinitive
AJ | % przymiotnik
AV | % przysłówek
OB | %objective relative clause
RC | %zdanie podrzędne bez spójnika
TH | %zdanie podrzędne "*that"*
LC %okolicznik miejsca
Target_Preposition *::*= **from** | **to** | **with** | ... % przyimki języka angielskiego

Przykłady kodowania pola *Complementation* podane zostaną w sekcji 4.

3.4.2. Pole *Semantics*

W polu tym opisane są podstawowe cechy semantyczne rzeczowników (fraz rzeczownikowych) oraz podmiotów czynności dla czasowników. W przypadku czasowników, pole *Semantics* jest niepuste tylko wtedy, gdy informacja o cechach semantycznych podmiotu może mieć znaczenie w algorytmie tłumaczenia (wydaje się, że nie jest to częsty przypadek, jednak przykładem potwierdzającym zasadność wprowadzenia takiego rozwiązania może być czasownik “tłumaczyć się”, który w przypadku osobowego podmiotu czynności znajduje swój angielski odpowiednik jako: “to excuse oneself”, - względnie “to explain oneself” - , natomiast w przypadku nieożywionego podmiotu czynności - np. w zdaniu - “Ta książka dobrze się tłumaczy” - znajduje swój angielski odpowiednik jako: “to translate”).

Wartościami pola *Semantics* są wyrażenia, które mogą być derywowane z symbolu *Semantic_value* w gramatyce podanej w punkcie 3.4.1.

3.4.3. Pole *Context*

Wartość pola kontekst ma postać *?Context_Value* dla haseł, w których kontekst jest “daną wejściową” lub *+Context_Value* dla haseł, w których kontekst

jest “daną wyjściową”. Na przykład dla rzeczownika “hasło” kontekst wystąpienia wyrazu w zdaniu decyduje o tym, czy odpowiednikiem angielskim jest wyraz “entry’, czy “password”. Blok odpowiedników angielskich wejścia “hasło” składa się z dwóch elementów: wartością pola *Context* pierwszego elementu jest *?Leksykografia* (odpowiednik: “entry”), a wartością pola *Context* drugiego elementu jest *?Bezpieczeństwo* (odpowiednik “password”). W słowniku będą znajdowały się również hasła, które nadają określony kontekst wypowiedzi (dla takich haseł kontekst jest “daną wyjściową”), np. hasło “słownik” ma w polu *Context* wartość *+Leksykografia*.

Pole *Context* ma w założeniu być niepuste i mieć wartość pierwszego znaku: ‘?’ tylko dla haseł polisemicznych z dziedziny informatyki, którym w zależności od kontekstu odpowiadają inne wyrazy (frazy) języka angielskiego. Nie zakłada się skonstruowania pełnej hierarchii semantycznej pojęć informatycznych. Na podstawie analizy list leksemów uzyskanych z korpusu planuje się stworzenie prostego podziału dziedziny informatyki na konteksty tak, aby możliwe było rozdzielenie znaczeń wyrazów polisemicznych z listy.

3.4.4. Pole *Polish Syntax*

Wartością tego pola jest typ frazy dla fraz leksykalnych. Ponadto w polu tym przechowywana jest informacja o zwrotności polskiego czasownika (takie niezbyt “estetyczne” rozwiązanie ma na celu niemnożenie liczby pól opisu).

3.4.5. Pole *Priority*

Pole *Priority* nie jest puste tylko dla tych haseł, w których dla jednego wyrazu (frazy) języka polskiego istnieje kilka odpowiedników języka angielskiego i nie jest możliwe wskazanie warunków wyboru któregoś z nich w żadnym z pól: Complemenetation, *Semnatics, PolihSyntax, Context.* Wartością pola jest liczba naturalna, wskazująca priorytet wyboru odpowiednika.

3.4.6. Pole *English Lexeme Identifier*

Wartością tego pola jest forma kanoniczna leksemu odpowiednika angielskiego.

3.4.7. Pole *English Inflection Code*

Wartością tego pola jest kod odmiany odpowiednika angielskiego. Przyjmuje się, że kod powinien być tak skonstruowany, aby algorytm tłumaczenia mógł na jego podstawie wygenerować formy fleksyjne bez konieczności konsultowania jakiegokolwiek pliku klasyfikacji (kod taki opisano w [Jassem, 1997] i nazwano “kodem konstruktywnym”).

3.4.8. Pole *English Syntax*

Pole to jest niepuste tylko dla czasowników (i fraz czasownikowych) języka angielskiego.

W polu tym zawarte są informacje składniowe niezbędne do poprawnej syntezy wyrażenia języka angielskiego takie jak: nietworzenie (być może tylko w

pewnych warunkach określonych przez pole *Complementation*) przez czasownik form ciągłych, tendencja do łączenia się z czasownikami posiłkowymi, czy zwrotność czasownika.

4. Przykłady wejść w słowniku

W tej sekcji podanych zostanie kilka przykładów wejść w słowniku, które mają zobrazować przydatność przyjętych konwencji w automatycznym (algorytmicznym) rozwiązaniu niektórych problemów tłumaczenia z języka polskiego na język angielski. Z lingwistycznego punktu widzenia dokonano pewnych uproszczeń.

Przykład 2.

```
<L id ="tłumaczyć" polishInflection="C58N">
  <F morphology = "OP1">tłumaczę</F>
  <F morphology = "OP2">tłumaczysz</F>
  ...
  <T complementation="" semantics="Hum" context=""
    polishSyntax="refl" priority="1" englishInflection="V1"
    englishSyntax="refl">excuse</T>
  <T complementation="" semantics="Hum" context=""
    polishSyntax="refl" priority="2" englishInflection="V1"
    englishSyntax="refl">explain</T>
  <T complementation="{z G → from NP, na A → into NP}"
    semantics="-Hum" context="" polishSyntax="refl"
    priority="" englishInflection="V1"
    englishSyntax="">translate</T>
  <T complementation={"A:-Anim → NP, z G → from NP, na A → NP}"
    semantics="Hum" context="?Przekład" polishSyntax=""
    priority="" englishInflection="V1"
    englishSyntax="">translate</T>
  <T complementation={"A:Abstr → NP, D: Hum → to NP}"
    semantics="Hum" context="" polishSyntax="" priority=""
    englishInflection="V1"
    englishSyntax="">explain</T>
</L>
```

Powyższe wejście opisuje czasownik tłumaczyć (się). Przyjmuje się założenie, że formy zwrotne i niezwrotne czasownika nie są rozdzielane. W algorytmie tłumaczenia, podczas analizy leksykalnej, która najczęściej poprzedza analizę syntaktyczną, nie można z reguły określić, czy czasownik wystąpił w zdaniu w formie zwrotnej czy niezwrotnej, dlatego w słowniku łączy się obie formy.

Wyjaśnione zostanie znaczenie wartości atrybutów elementów zawartych między znacznikami <T , T>. Pierwsze trzy elementy odpowiadają czasownikowi zwrotnemu “tłumaczyć się” (wartość atrybutu *polishSyntax* wynosi “refl”). W pierwszym i drugim elemencie wartość atrybutu *Semantics* wynosi “Hum”, co

oznacza, że element opisuje czynność wykonywaną przez człowieka. Wartością pierwszego elementu *T* (odpowiednikiem angielskim) jest czasownik zwrotny (wartość atrybutu *EnglishSyntax* wynosi "refl") "excuse", który tworzy swoje formy fleksyjne w sposób regularny (wartość atrybutu *EnglishInflectionCode* wynosi "V1"). Wartości atrybutów drugiego elementu (o wartości "explain") są takie same jak dla pierwszego elementu z wyjątkiem atrybutu *Priority*. Trzeci element opisuje czasownik "tłumaczyć się" z podmiotem nieosobowym, któremu odpowiada angielski czasownik "translate" Wartość atrybutu *Complementation* wskazuje, że czasownik może być modyfikowany grupami przyimkowymi (np. w zdaniu: To się dobrze tłumaczy z języka polskiego na angielski), występującymi w dowolnym szyku, przy czym grupy przyimkowe zostaną odpowiednio przetransferowane na język angielski.

Czwarty i piąty element opisują czynność tłumaczenia dokonywaną przez człowieka i dotyczą składni odpowiednio: "tłumaczyć coś z czegoś na coś" oraz "tłumaczyć coś komuś". Żadne z wymagań w opisie nie jest ujęte w nawiasy kwadratowe (nawiasy kwadratowe opisują obligatoryjność wystąpienia wymagania w analizowanym zdaniu), co oznacza, że w wyrażeniu języka polskiego dopuszczalne jest wystąpienie dowolnego podzbioru wymagań opisu. W szczególności fraza "tłumaczyć coś" będzie spełniała warunki nałożone przez atrybuty *Complementation* obu elementów. Przy analizie konstrukcji "tłumaczyć coś" wybór odpowiednika będzie możliwy na podstawie analizy kontekstu (wartość atrybutu *Context* czwartego elementu wynosi "?Przekład").

Przykład 3.

```
 <L id ="praca" polishInflection="R414">
   <F morphology = "ŻMP">praca</F>
   <F morphology = "ŻDP">pracy</F>
   ...
   <T complementation="nad I:Abstr → on NP" semantics="Abstr"
     context="?Nauka" polishSyntax="" priority=""
     englishInflection="N00" englishSyntax="">research</T>
   <T complementation="" semantics="-Anim"
     context="?Artykuł naukowy"     polishSyntax=""
     priority="" englishInflection="N1"
     englishSyntax="">paper</T>
   <T complementation="" semantics="Abstr" context="?posada"
     polishSyntax="" priority="1" englishInflection="N1"
     englishSyntax="">job</T>
   <T complementation="" semantics="Abstr" context="?posada"
     polishSyntax="" priority="2" englishInflection="N00"
     englishSyntax="">occupation</T>
   <T complementation="" semantics="Abstr" context=""
     polishSyntax="" priority="" englishInflection="N1"
     englishSyntax="">work</T>
</L>
```

Powyższy przykład opisuje wejście o polskim identyfikatorze “praca”. Pierwszy element wskazuje, że wyraz “praca” powinien być przetłumaczony na wyraz “research”, jeśli występuje w kontekście nauki. Możliwa jest modyfikacja frazy rzeczownikowej dopełnieniem przyimkowym “nad czymś” - wtedy angielskim odpowiednikiem dopełnienia jest fraza “on smth”. Wartość atrybutu *englishInflection* wynosi “N00”, co oznacza, że rzeczownik ten nie występuje w j. angielskim w liczbie mnogiej.

W kontekście “artykułu naukowego” odpowiednikiem angielskim jest wyraz “paper”, którego cechy semantyczne oznaczone są jako *-Anim,* (obiekt nieożywiony, nie abstrakcyjny).

Istnieją dwa odpowiedniki wyrazu praca w kontekście “posady”: “job” oraz “occupation”. Odpowiednik “job” wybrano jako preferowany (wartość atrybutu *Priority* wynosi “1”).

W pozostałych przypadkach (w innych kontekstach niż wskazane w poprzednich elementach) odpowiednikiem angielskim jest wyraz “work”.

Przykład 4.

```
<L id ="aktualny" polishInflection="P46">
  <F morphology = "MoMPR">aktualny</F>
  <F morphology = "MoMPR">aktualnego</F>
  ...
  <T complementation="" semantics="" context=""
    polishSyntax="" priority="1" englishInflection="A0"
    englishSyntax="">current</T>
  <T complementation="" semantics="" context=""
    polishSyntax="" priority="2" englishInflection="A0"
    englishSyntax="">topical</T>
  <T complementation="" semantics="" context=""
    polishSyntax="" priority="3" englishInflection="A0"
    englishSyntax="">up-to-date</T>
</L>
```

W tym przykładzie opisano polski przymiotnik “aktualny”, któremu odpowiada kilka przymiotników języka angielskiego. Nie potrafiono określić warunków wyboru angielskiego odpowiednika, dlatego niezbędne było określenie wartości atrybutów *Priority*.

Przykład 5.

```
<L id ="liczba całkowita" polishInflection="R414;b-bi P38">
  <F morphology = "ŻMP">liczba całkowita</F>
  <F morphology = "ŻDP">liczby całkowitej</F>
  ...
  <T complementation="" semantics="Abstr" context=""
    polishSyntax="attr_phr" priority=""
    englishInflection="N1" englishSyntax="">integer</T>
</L>
```

Powyżej przedstawiono opis frazy atrybutywnej “liczba całkowita”, której angielskim odpowiednikiem jest rzeczownik “integer”.

5. Podsumowanie

Powyższy format wejścia w elektronicznym słowniku dwujęzycznym nie ma ambicji rozstrzygnięcia wszystkich, czy nawet większości, lingwistycznych problemów przekładu z języka polskiego na język angielski. Celem pracy było sformalizowanie niektórych własności wyrazów i fraz języka polskiego w taki sposób, aby informacje opisane zgodnie z przyjętym formalizmem mogły być automatycznie zinterpretowane i wykorzystane w algorytmie komputerowego tłumaczenia. W najbliższym czasie algorytm tłumaczenia automatycznego opisany w [2] zostanie tak zmodyfikowany, aby mógł wykorzystać każdy typ informacji leksykalnej, ujęty w przedstawionym formalizmie.

W pracach nad utworzeniem słownika na bazie list leksemów wygenerowanych z korpusu tekstów niezbędna będzie współpraca lingwistów. Współpraca ta może zaowocować uchwyceniem pewnych innych cech lingwistycznych, które dadzą się zinterpretować w sposób algorytmiczny. Format haseł w słowniku - zgodny ze standardem SGML - oraz metodologia oprogramowania do obsługi słownika opisana w [4] i [3], umożliwia uzupełnienie opisu haseł o nowe pola (atrybuty).

BIBLIOGRAFIA

[1] Graliński F., (1998) *Realizacja pół-automatycznej ekstrakcji leksemów występujących w korpusie polskich tekstów informatycznych*, w : “Speech and Language Technology. Vol. 2”, Poznań.

[2] Jassem K., (1997) *POLENG - a Machine Translation System Based on an Electronic Dictionary*, w: “Speech and Language Technology. Vol. 1”, Poznań.

[3] Lison M., (1998) *Model i realizacja struktury danych leksykalnych słownika elektronicznego skonstruowanego dla potrzeb automatycznego tłumaczenia*, w: “Speech and Language Technology. Vol. 2”, Poznań.

[4] Rutkowski B., (1998) *Edytor polsko-angielskiego słownika w formacie SGML*, w: “Speech and Language Technology Vol. 2”, Poznań 1998.