Kategorialny Parser Składniowo-Semantyczny „ENIAM”
składnia leksykonu LCG

Wojciech Jaworski

Leksykon LCG składa sie z trzech sekcji:

`@PHRASE_NAMES` — zawiera listę symboli atomowych zdefiniowanych przez użytkownika

`@WEIGHTS` — zawiera listę definicji wag

`@LEXICON` — zawiera listę reguł.

Reguły zdefiniowane są zgodnie z następującą gramatyką:

```
            <rule> ::= <selectors>:<type>; |
                       <selectors>:QUANT[<quantifiers>]<type>;
       <selectors> ::= <selector> | <selector>,<selectors>
        <selector> ::= <gram>=<gram-values> |
                       <gram>!=<gram-values>
     <gram-values> ::= <gram-value> |
                       <gram-value>|<gram-values>
     <quantifiers> ::= <quantifier> | <quantifier>,<quantifiers>
      <quantifier> ::= <gram>=0 | <gram>=all_numbers |
                       <gram>=all_cases | <gram>=all_genders |
                       <gram>=all_persons | <gram>=<quant-gram-values>
<quant-gram-values> ::= <gram-value> |
                       <gram-value>&<quant-gram-values>
```

```
  <type>  ::=  <maybe> | <imp> | <impset>
   <imp>  ::=  <type><dir><type>
<impset>  ::=  <type>{<args>}
  <args>  ::=  <dir><type> | <dir><type>,<args>
   <dir>  ::=  / | | | \
 <maybe>  ::=  <plus> | ?<plus>
  <plus>  ::=  <tensor> | 1 | <tensor>+<plus> | 1+<plus>
<tensor>  ::=  <atom> | <atom>*<tensor>
  <atom>  ::=  <user> | <gram> | T
```

Symbol `<user>` można rozwinąć do dowolnego symbolu z listy zdefiniowanej w sekcji `@PHRASE_NAMES`.

Poniższa tabela definiuje dostępne selektory wraz z ich możliwymi wartościami:

<gram>	<gram-value>
lemma	bez ograniczeń
pos	subst depr ppron12 ppron3 siebie prep num intnum realnum intnum-interval realnum-interval symbol ordnum date date-interval hour-minute hour hour-minute-interval hour-interval year year-interval day day-interval day-month day-month-interval month-interval roman roman-interval roman-ordnum match-result url email obj-id adj adjc adjp adja adv ger pact ppas fin bedzie praet winien impt imps pred aglt inf pcon pant qub comp conj interj sinterj burk interp unk
cat	bez ograniczeń
number	sg pl
case	nom gen dat acc inst loc postp pred
gender	m1 m2 m3 f n1 n2 p1 p2 p3
person	pri sec ter
grad	pos com sup
praep	praep npraep praep-npraep
acm	congr rec
ctype	int rel sub coord
mode	abl adl locat perl dur temp mod

aspect	perf imperf
negation	neg aff
mood	indicative imperative conditional
tense	past pres fut
nsyn	proper pronoun common
nsem	count time mass measure

Poniższa tabela zawiera listę selektorów zdefiniowanych dla poszczególnych części mowy:

<pos>	<gram>
subst	lemma cat number case gender person nsyn nsem
depr	lemma cat number case gender person nsyn nsem
ppron12	lemma number case gender person
ppron3	lemma number case gender person praep
siebie	lemma number case gender person
prep	lemma cat case
compar	lemma cat case
num	lemma number case gender person acm
intnum	lemma number case gender person acm
realnum	lemma number case gender person acm
intnum-interval	lemma number case gender person acm
realnum-interval	lemma number case gender person acm
symbol	lemma number case gender person
ordnum	lemma number case gender grad
date	lemma nsyn nsem
date-interval	lemma nsyn nsem
hour-minute	lemma nsyn nsem
hour	lemma nsyn nsem
hour-minute-interval	lemma nsyn nsem
hour-interval	lemma nsyn nsem
year	lemma nsyn nsem
year-interval	lemma nsyn nsem
day	lemma nsyn nsem
day-interval	lemma nsyn nsem
day-month	lemma nsyn nsem
day-month-interval	lemma nsyn nsem
month-interval	lemma nsyn nsem
roman-ordnum	lemma number case gender grad
roman	lemma nsyn nsem
roman-interval	lemma nsyn nsem

match-result	lemma nsyn nsem
url	lemma nsyn nsem
email	lemma nsyn nsem
obj-id	lemma nsyn nsem
adj	lemma cat number case gender grad
adjc	lemma cat number case gender grad
adjp	lemma cat number case gender grad
apron	lemma number case gender grad
adja	lemma cat
adv	lemma cat grad mode
ger	lemma cat number case gender person aspect negation
pact	lemma cat number case gender aspect negation
ppas	lemma cat number case gender aspect negation
fin	lemma cat number gender person aspect negation mood tense
bedzie	lemma cat number gender person aspect negation mood tense
praet	lemma cat number gender person aspect negation mood tense
winien	lemma cat number gender person aspect negation mood tense
impt	lemma cat number gender person aspect negation mood tense
imps	lemma cat number gender person aspect negation mood tense
pred	lemma cat number gender person aspect negation mood tense
aglt	lemma number person aspect
inf	lemma cat aspect
pcon	lemma cat aspect
pant	lemma cat aspect
qub	lemma
comp	lemma
conj	lemma
interj	lemma
sinterj	lemma
burk	lemma
interp	lemma
unk	lemma number case gender person

Selektor postaci `<gram>=<gram-values>` oznacza, że aby reguła została użyta token musi mieć symbole należące do `<gram-values>` wśród wartości kategorii `<gram>`; reguła ta zostanie wykonana dla `<gram>` z usuniętymi pozostałymi wartościami. Selektor postaci `<gram>!=<gram-values>` oznacza, że aby reguła została użyta token musi mieć symbole nie należące do `<gram-values>` wśród wartości kategorii `<gram>`; reguła zostanie wykonana dla `<gram>` z usuniętymi wartościami z `<gram-values>`.

Nazwy kategorii gramatycznych (`<gram>`) użyte jako symbole atomowe (`<atom>`) pełnią rolę zmiennych. Zmienne te są automatycznie kwantyfikowane na podstawie wartości kategorii gramatycznych przetwarzanego leksemu. Np. reguła

```
pos=prep:  prepnp*cat*lemma*case/np*cat*T*case*T;
```

zostanie dla leksemu w mającego kategorię *case* równą acc bądź loc przetworzona do postaci:

$$\bigwedge_{cat=0} \bigwedge_{case=\mathrm{acc\&loc}} \mathrm{prepnp} \bullet cat \bullet \mathrm{w} \bullet case \,/\, \mathrm{np} \bullet cat \bullet \top \bullet case \bullet \top$$

W powyższym przykładzie została wprowadzona zmienna *case* mająca dwie możliwe wartości, zaś nie posiadająca ograniczeń kategoria *cat* została skwantyfikowana jako $cat = 0$ co oznacza, na zmienną *cat* można przypisać dowolną wartość.

Aby ograniczyć kwantyfikację należy dodać selektor wskazujący możliwe wartości danej kategorii gramatycznej. Np. reguła

```
lemma=w,pos=prep,case=loc:  location/np*MIASTO*T*case*T;
```

zostanie przetworzona do postaci:

$$\mathrm{location} \,/\, \mathrm{np} \bullet \mathrm{MIASTO} \bullet \top \bullet \mathrm{loc} \bullet \top$$

Selektor `case=loc` spowodował, że *case* ma tylko jedną możliwą wartość.

W niektórych sytuacjach (np. przy koordynacji) trzeba wprowadzić kwantyfikację po selektorach nie zdefiniowanych dla danej kategorii gramatycznej. Służy do tego konstrukcja `QUANT[<quantifiers>]`, gdzie `<quantifiers>` opisuje dodatkowe selektory i zbiory ich wartości. `0` oznacza dowolną wartość, `all_cases` to zbiór wszystkich wartości przypadka, podobnie `all_numbers`, `all_genders` i `all_persons`. Konkretne wartości można podać za pomocą listy rozdzielonej symbolem `&`.

Powyższą konstrukcję można również wykorzystać do tego by w dowolny sposób zmienić zakres kwantyfikatora dla selektora zdefiniowanego dla danej kategorii gramatycznej. Gdy taki selektor pojawi się na liście kwantyfikatorów, jego dopuszczalne wartości będą takie jak określone na tej liście, a nie takie jak wynikają z analizy morfologicznej i selektorów reguły.

Poniższa reguła definiuje koordynację fraz rzeczownikowych, w której uzgadniana jest kategoria *cat* oraz przypadek *case*, natomiast liczba *number* i rodzaj *gender* skoordynowanych fraz mogą być dowolne.

Reguła

```
lemma=i,pos=conj:
QUANT[cat=0,number=all_numbers,case=all_cases,gender=all_genders]
(np*cat*number*case*gender\np*cat*T*case*T)/np*cat*T*case*T;
```

zostanie przetworzona do postaci:

$$\underset{cat:=0}{\&}\ \underset{number:=\mathrm{sg}\&\mathrm{pl}}{\&}\ \underset{case:=\mathrm{nom}\&\mathrm{gen}\&\mathrm{dat}\&\mathrm{acc}\&\mathrm{inst}\&\mathrm{loc}\&\mathrm{voc}}{\&}\ \underset{gender:=\mathrm{m}_1\&\mathrm{m}_2\&\mathrm{m}_3\&\mathrm{f}\&\mathrm{n}_1\&\mathrm{n}_2\&\mathrm{p}_1\&\mathrm{p}_2\&\mathrm{p}_3}{\&}$$

$$(\mathrm{np} \bullet cat \bullet number \bullet case \bullet gender \setminus \mathrm{np} \bullet cat \bullet \top \bullet case \bullet \top) \,/\, \bullet cat \bullet \top \bullet case \bullet \top$$