Pierwsze próby ekstrakcji ram walencyjnych z Korpusu IPI z użyciem programu Świgra

Marcin Woliński

INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. J. K. Ordona 21, 01-237 Warszawa

27 listopada 2006, IPI PAN

Gramatyka Świdzińskiego (GFJP)

Marek Świdziński

Gramatyka formalna języka polskiego
Wydawnictwa Uniwersytetu Warszawskiego, 1992

- realizacja komputerowa nie była celem Autora
- składnia powierzchniowa (bez semantyki)
- największa i najbardziej szczegółowa gramatyka formalna polszczyzny (460 reguł)

Formalizm gramatyczny

Gramatyki metamorficzne — prologowy formalizm gramatyczny zaproponowany przez Colmerauera:

Alain Colmerauer

Metamorphosis grammars
W: L. Bolc (ed.), Natural Language Communication with Computers. *Lecture Notes in Computer Science 63*. Springer-Verlag 1978, pp. 133–189.

Obecnie bardziej znany w wariancie **Definite Clause Grammars (DCG)**.

Jedna z reguł GFJP opisujących zdanie elementarne

$$\begin{aligned}
&\mathbf{ze}(Wf, A, C, T, Rl, O, Wa, Wb, Wc, Neg, I, Z, Ow) \longrightarrow \qquad \text{(e5)}\\
&\quad \mathbf{fw}(Wa, K, A, C, Rl, O, Neg, I, \text{np}),\\
&\quad \langle\!\langle \mathbf{ff}(Wf, A, C, T, Rl, O, Wa, Wb, Wc, K, Neg, \text{ni}, Z, Ow),\\
&\quad \mathbf{fw}(Wb, K, A, C, Rl, O, Neg, \text{ni}, \text{np}),\\
&\quad \mathbf{fw}(Wc, K, A, C, Rl, O, Neg, \text{ni}, \text{np})\rangle\!\rangle,\\
&\quad \{\ \text{różne}(Z, [\text{byxx}, \text{choćby}, \text{co}, \text{czyżby}, \text{gdyby}, \text{jakby}, \text{jaki},\\
&\qquad\qquad \text{jakoby}, \text{kto}, \text{który}, \text{p}, \text{px}, \text{pxx}, \text{pz}, \text{żeby}])\ \}.
\end{aligned}$$

Jedna z reguł GFJP opisujących zdanie elementarne

ze(*Wf*, *A*, *C*, *T*, *Rl*, *O*, *Wym*, *Neg*, *I*, z(*SwZ*, *NZ*), *Ow*,
@@@@0) $\longrightarrow$ (e5)
fw(*W0*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(*SwZ3*, *Z3*)),
{ zrowne(*Z3*, [np], *SwZ3*) },
wymagania([*W0*], *Wym*, *ResztaWym*,
ff(*Wf*, *A*, *C*, *T*, *Rl*, *O*, *Wym*, *K*, *Neg*, ni, z(*SwZ*, *Z*), *Ow*),
[*W1*/**fw**(*W1*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, ni, **z**(*SwZ1*, *Z1*)),
W2/**fw**(*W2*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, ni, **z**(*SwZ2*, *Z2*))]),
{ resztawym(*ResztaWym*),
zrozne(*Z*, [byxx, choćby, co, czyżby, gdyby, jakby, jaki,
jakoby, kto, który, p, px, pxx, pz, żeby], *NZ*),
zrowne(*Z1*, [np], *SwZ1*),
zrowne(*Z2*, [np], *SwZ2*) }.

Projekt walencyjny w IPI PAN

Automatyczna ekstrakcja wiedzy lingwistycznej z dużego korpusu języka polskiego.

grant „KBN” 3 T11C 003 28
kierownik projektu: Adam Przepiórkowski
czas trwania: 2005–2007

Projekt 3 T11C 003 28

- **Cel lingwistyczno-komputerowy:** Opracowanie i zaimplementowanie algorytmów automatycznego wydobywania powierzchniowych informacji walencyjnych z korpusów anotowanych morfosyntaktycznie.

Projekt 3 T11C 003 28

- **Cel lingwistyczno-komputerowy:** Opracowanie i zaimplementowanie algorytmów automatycznego wydobywania powierzchniowych informacji walencyjnych z korpusów anotowanych morfosyntaktycznie.
- **Cele uboczne**
  - **praktyczny:** Stworzenie słownika walencyjnego języka polskiego na podstawie Korpusu IPI PAN.
  - **teoretyczny:** Zbadanie teoretycznych i praktycznych aspektów dychotomii ‘argument vs okolicznik’.

„Walencja powierzchniowa”

Przykłady ram walencyjnych:

- wspierać:
  - ⟨NP[NOM], NP[ACC]⟩,
  - ⟨NP[NOM], NP[ACC], PP[W-LOC]⟩,
- pomagać:
  - ⟨NP[NOM], NP[DAT]⟩,
  - ⟨NP[NOM], NP[DAT], VP[INF]⟩,
  - ⟨NP[NOM], NP[DAT], PP[W-LOC]⟩,

Argumenty a okoliczniki

- **Argumenty:** idiosynkratyczne, tj. ich postać zależy od danego predykatu, a więc powinny być podane w słowniku walencyjnym.
- **Okoliczniki:** modyfikują predykaty według ogólnych reguł gramatycznych, a wiec nie muszą być podane w słowniku walencyjnym.

Argumenty a okoliczniki

A. Przepiórkowski analizuje kryteria odróżniania argumentów i okoliczników i stwierdza, że są one niespójne i niejednoznaczne.

Argumenty a okoliczniki

A. Przepiórkowski analizuje kryteria odróżniania argumentów i okoliczników i stwierdza, że są one niespójne i niejednoznaczne.

Decyzja: Brak rozróżnienia w słowniku walencyjnym.

Metoda

1. Podział tekstu na zdania elementarne
2. Identyfikacja czasowników i maksymalnych fraz w każdym zdaniu
3. Zastosowanie metod wnioskowania statystycznego do takiego zbioru obserwacji
4. Zapisanie w słowniku pozytywnie zweryfikowanych kombinacji czasowników i argumentów

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 5196 zdań z Korpusu IPI PAN:

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 5196 zdań z Korpusu IPI PAN:

- 51,7% zdań ma niepełny opis morfologiczny,

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 5196 zdań z Korpusu IPI PAN:

- 51,7% zdań ma niepełny opis morfologiczny,
- 3,7% zdań zawiera liczebniki,

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 5196 zdań z Korpusu IPI PAN:

- 51,7% zdań ma niepełny opis morfologiczny,
- 3,7% zdań zawiera liczebniki,
- 10,1% „zdań” nie zawiera żadnej finitywnej formy czasownika,

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 5196 zdań z Korpusu IPI PAN:

- 51,7% zdań ma niepełny opis morfologiczny,
- 3,7% zdań zawiera liczebniki,
- 10,1% „zdań” nie zawiera żadnej finitywnej formy czasownika,
- pozostałe 34,5% można poddać analizie składniowej,
- a w wyniku ...

Pierwszy eksperyment

Analiza niezmienioną *Świgrą* 2039 zdań z Korpusu IPI PAN:

	akceptowane	odrzucone
zdania	590	1449
	29,0%	71,0%
drzewa	300	
czas (s)	0,44	0,2
kroki wyw.	603983	319901

Sumaryczna liczba drzew analizy: 12 705 836 679
Sumaryczny czas: 1h 28m 13s

Wyniki Macieja Ogrodniczuka

Niezmieniona *Świgra* daje pokrycie 31% zdań na Korpusie Słownika Frekwencyjnego.

Wyniki Macieja Ogrodniczuka

Niezmieniona *Świgra* daje pokrycie 31% zdań na Korpusie Słownika Frekwencyjnego.

Po rozszerzeniu gramatyki o frazy liczebnikowe i koordynację wewnątrz fraz analizator akceptuje 84% zdań.

Co trzeba zmienić w Świgrze

Zmiany w gramatyce konieczne, dla ekstrakcji ram walencyjnych:

- wyłączenie wbudowanego słownika walencyjnego :-),
- usunięcie występującego w GFJP ograniczenia do trzech fraz wymaganych (wliczając podmiot),
- eliminacja rozróżnienia między frazą wymaganą, a luźną (na poziomie zdania elementarnego),
- odwzorowanie typów fraz na zaproponowane przez Adama.

Reguła e5 dopuszczająca dowolnie wiele fraz

ze(*Wf*, *A*, *C*, *T*, *Rl*, *O*, *Wym*, *Neg*, *I*, z(*SwZ*, *NZ*), *Ow*,
@@@@0) $\longrightarrow$ (e5)
fw(*W0*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(*SwZ3*, *Z3*)),
{ zrowne(*Z3*, [np], *SwZ3*) },
wymagania([*W0*], *Wym*, *ReszteWym*,
ff(*Wf*, *A*, *C*, *T*, *Rl*, *O*, *Wym*, *K*, *Neg*, ni, z(*SwZ*, *Z*), *Ow*),
[*W1*/**fw**(*W1*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, ni, **z**(*SwZ1*,
Z1))/**zrowne**(*Z1*, [np], *SwZ1*)]),
{ resztawym(*ReszteWym*),
zrozne(*Z*, [byxx, choćby, co, czyżby, gdyby, jakby, jaki,
jakoby, kto, który, p, px, pxx, pz, żeby], *NZ*) }.

Definicja frazy wymaganej w GFJP

fw(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, *Z*) $\longrightarrow$ (wy1)
  fw1(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, *Z*).

fw(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(_, [p])) $\longrightarrow$ (wy2)
  fw1(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(*SwZ1*, *Z1*)),
  fl(*A*, *C*, *Rl*, *O*, *Neg*, ni, z(*SwZ2*, *Z2*)),
  { zplubnp(*Z1*, *Z2*, *SwZ1*, *SwZ2*) }.

fw(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(*SwZ*, *NZ*)) $\longrightarrow$ (wy3)
  fw1(*Tfw*, *K*, *A*, *C*, *Rl*, *O*, *Neg*, *I*, z(*SwZ*, *Z*)),
  { zrowne(*Z*, [pz], *NZ*) },
  fl(*A*, *C*, *Rl*, *O*, *Neg*, ni, z(*SwZ1*, *Z1*)),
  { zrowne(*Z1*, [np, p], *SwZ1*) }.

Problem

Frazy luźne, które nie mogą być frazami wymaganymi:

- fraza nominalna w wołaczu otoczona przecinkami,
- fraza werbalna, której centrum stanowi imiesłów przysłówkowy uprzedni lub współczesny, otoczona przecinkami,
- fraza zdaniowa typu **aż2, bo, chociaż, choćby, co, dopóki, gdy, gdyby, jeśli, podczas, ponieważ, zanim**.

Drugi eksperyment

	Świgra		ze zmianami	
	akc.	odrz.	akc.	odrz.
zdania	590	1449	544	1495
	29,0%	71,0%	26,7%	73,3%
drzewa	300		183	
czas (s)	0,44	0,2	1,29	0,57
kroki wyw.	603983	319901	868284	539021
razem drzew	12 705 836 679		1 326 161 857	
razem czas	1h 28m 13s		9h 58m 11s	

Zdania odrzucone zawierają 208, dla których analiza nie zakończyła się w 8 minut.

Przykłady ram wykrytych przez Świgrę

- Wolałbym obejrzeć pani stopę. (21 drzew)
  - woleć [InfP(perf)]
  - woleć [InfP(perf), NP(acc)]
  - woleć [InfP(perf), NP(dat), NP(acc)]
  - woleć [InfP(perf), NP(gen), NP(acc)]
  - woleć [InfP(perf), NP(loc), NP(acc)]

Jak odwzorować typy fraz GFJP?

typy fraz zadane przez Adama	GFJP
nominal phrase	**np**
numeral	brak
adjectival	**adjp**
prepositional-nominal	**prepnp**
prepositional-numeral	brak
prepositional-adjectival	brak
adverbial	**advp**
infinitival	**infp**
sentential introduced by a complementiser	?
sentential int. by an interrogative or relative phrase	?
oratio recta	brak
reflexive marker	?

Dalsze prace

- Potrzeba więcej wyników, żeby je przepuścić przez maszynerię statystyczną.
- Trzeba rozszerzyć gramatykę:
  - koordynacja wewnątrz fraz,
  - frazy liczebnikowe.
- Przydałaby się pełniejsza analiza morfologiczna (wprowadzenie modułu zgadującego?).
- Zbadać, czy podział na zdania podany w Korpusie jest wystarczająco dobry.
- Jak wywnioskować, która z wielu ram dla danego zdania jest dobra?