Morfologiczno-składniowe znakowanie korpusów tekstów w języku polskim – po co i jak?

Adam Radziszewski
Politechnika Wrocławska

17 lipca 2009

O czym będzie mowa

- Co daje korpus oznakowany?
  - Badanie języka a wyszukiwanie w tekście
  - Bez korpusu oznakowanego
  - Z korpusem oznakowanym
- Tager
  - Wieloznaczność gramatyczna
  - Jak to działa?
- Płaskie frazy składniowe (*chunking*)
- Dyskusja

Badanie języka a wyszukiwanie w tekście

- Badanie właściwości danego typu wyrażeń
  - Znaleźć wszystkie wystąpienia słowa *rok* w tekście (wszystkie formy – *rok*, *latami* itp.)
  - Znaleźć wszystkie zaimki osobowe
  - Wszystkie wyrażenia typu *przyimek rzeczownik przyimek*, by sprawdzić ile z nich jest złożonymi przyimkami (*ze względu na*)
  - Słowa niejednoznaczne między rzeczownikiem a czasownikiem (*piekło*, *nadzieje*)
- Badanie kontekstu wystąpienia wyrażeń
  - W jakim kontekście pojawiają się zaimki osobowe?
  - Czy słowo *jeden* może pełnić funkcję przedimka nieokreślonego? (interesują nas całe zdania)

Bez korpusu oznakowanego

1. Mamy książki / notatki (papierowe).
   Przeglądanie linijka po linijce.

2. Mamy tekst w komputerze.
   Wszystkie wystąpienia słowa *rok* – wypisujemy na kartce wszystkie możliwe formy i szukamy ich po kolei.
   Znajdujemy formę *lata*:
   Minęły dwa lata.
   Statek, który lata na księżyc.
   Jak znaleźć wyrażenia typu *przyimek rzeczownik przyimek*?
   Mniejsze problemy: podział na zdania, wielkość liter, różna liczba znaków pustych

Z korpusem oznakowanym (1)

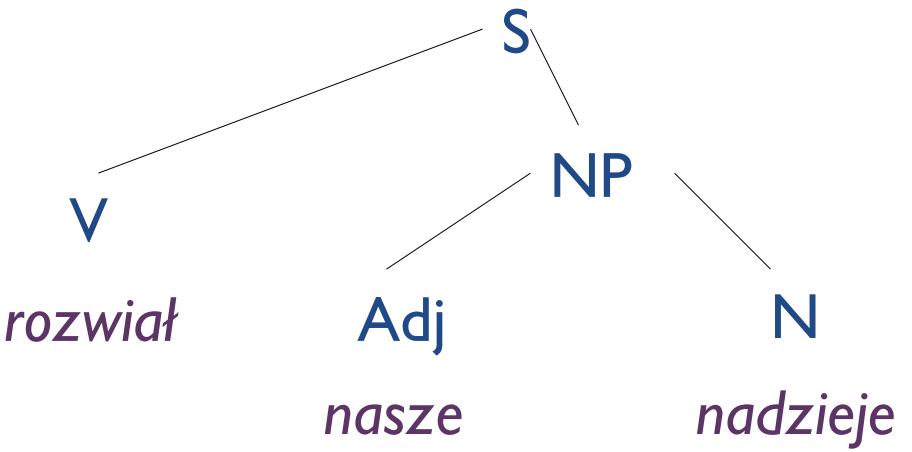
- Mamy ujednolicony zbiór tekstów
- Podział na zdania i wyrazy (abstrahujemy od ilości i rodzaju znaków pustych)
- Każdy tekst *może być* sklasyfikowany (gatunek, rok itp.)
- Program umożliwiający wyszukiwanie w korpusie (np. Poliqarp)
  - Można zadać wszystkie przykładowe zapytania
  - Zebrać wyniki w jednym miejscu
- Jak uzyskać oznakowany korpus ze zbioru własnych tekstów?

Z korpusem oznakowanym (2)

- Różne stopnie oznakowania korpusu
  - Podział na słowa (segmenty) i zdania
  - Klasyfikacja dokumentów (*metadane*)
  - Klasyfikacja słów
    - Części mowy (ogólniej: klasy słów)
    - Charakterystyka odmiany (np. przypadek)
    - Cechy składniowe (np. przyimek wymaga dopełniacza)
    - Formy hasłowe (lematy, *latają* → *latać*)
  - Opis niejednoznaczności
    - Słowa opisane odpowiednio do kontekstu
      Rozwiał nasze nadzieje/rzeczownik.
    - Opis niezależny od kontekstu
      Rozwiał nasze nadzieje/{rzeczownik, czasownik}

Z korpusem oznakowanym (3)

- Różne stopnie oznakowania korpusu (c.d.)
  - Pełny opis składniowy



  - Płytki opis składniowy

    [*Rozwiał*$_{V}$] [*nasze nadzieje*$_{NP}$]

Poliqarp (1)

- Narzędzie do wyszukiwania napisane w IPI PAN
- Wolne oprogramowania (licencja GPL – gwarancja, że zawsze będzie dostępny za darmo wraz z możliwością zmian i poprawek)
- Przetwarza korpusy oznakowane
  - Oznakowanie na poziomie słów
  - Niejednoznaczność
  - Podział na zdania i akapity
  - Wyszukiwanie po metadanych
- Język zapytań o dużej sile wyrazu
- Nieco kłopotliwy proces przygotowania korpusu

Poliqarp (2)

- Wyszukiwanie w ściągniętym korpusie IPI PAN
- Wyszukiwanie we własnym korpusie
- Interfejs sieciowy – korpus.pl
- DEMONSTRACJA

Tager

Adam Radziszewski
Politechnika Wrocławska

17 lipca 2009

O czym będzie mowa

- Co daje korpus oznakowany?
- **Tager**
  - Wieloznaczność gramatyczna
  - Jak ją rozwiązać?
  - TaKIPI
  - Błędy tagera
  - Jak to działa?
- Płaskie frazy składniowe (*chunking*)
- Dyskusja

Wieloznaczność gramatyczna (1)

- Interesuje nas klasyfikacja słów (segmentów)
  - Część mowy / klasa słowa
  - Cechy związane z odmianą (przypadek, liczba itp.)
  - Niektóre cechy składniowe
- W zależności od kontekstu, słowo może mieć różną interpretację
  - *Nic nie działa* / *Wytoczyli działa*
  - *Mamy siłę* / *Mamy nie ma* / *Młode mamy*
  - *To*: przymiotnik, rzeczownik, predykatyw, spójnik lub partykuła

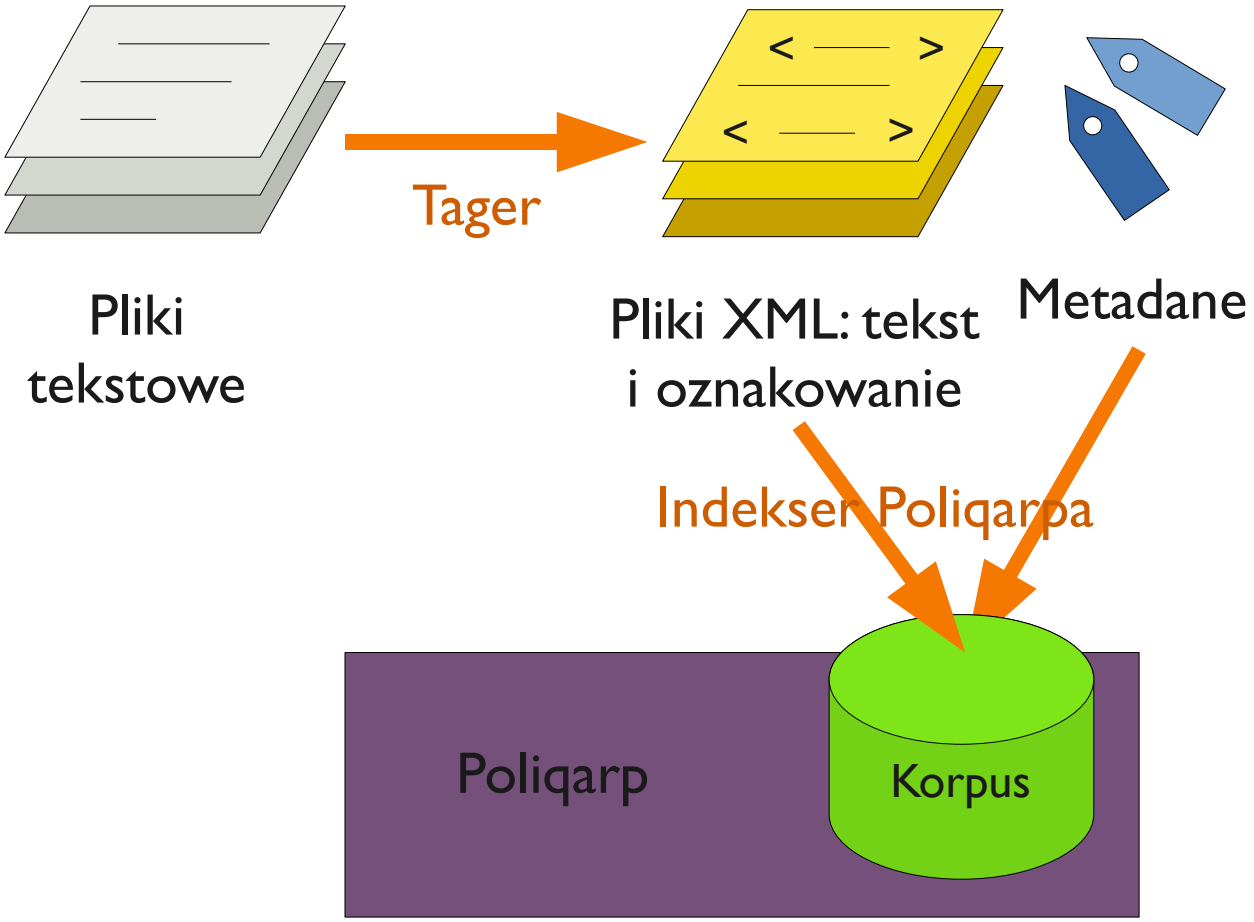
Wieloznaczność gramatyczna (2)

działa

działać	fin :	sg : ter : imperf
działo	subst :	sg : gen : n
	subst :	pl : nom.acc.voc : n
dziać	praet :	sg : f : imperf

- Dla człowieka rozstrzygnięcie tej niejednoznaczności nie stanowi problemu (o ile kontekst jest wystarczający)
- Podobnie działają programy zwane tagerami (dezambiguatorami, ang. *taggers, disambiguation engines*)

Po co nam tager?



Tager TaKIPI (1)

- Instytut Informatyki, Politechnika Wrocławska
- plwordnet.pwr.wroc.pl/g419/tagger
- Maciej Piasecki. *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*. Task Quarterly, 2007, 11, 151-167.
- Tager korzysta z analizatora Morfeusz:
  Marcin Woliński. *Morfeusz — a practical tool for the morphological analysis of Polish*. Proc. of Intelligent Information Processing and Web Mining

Tager TaKIPI (2)

- Tager przypisuje każdemu słowu znacznik (*tag*) i lemat
- Znaczniki są pozycyjne: subst : sg : gen : n
  rzeczownik dopełniacz
  liczba poj. r.nijaki
- Tager korzysta z analizatora morfologicznego *Morfeusz* (tj. leksykonu wszystkich możliwych znaczników)
- Proces jest dwuetapowy: bezkontekstowe przypisanie znaczników wg Morfeusza i ujednoznacznienie

Tager TaKIPI (3)

uda	*udać*	fin : sg : ter : perf
	udo	subst : sg : gen : n
		subst : pl : nom : n
		subst : pl : acc : n
		subst : pl : voc : n
szpaka	*szpak*	subst : sg : gen : m2
		subst : sg : acc : m2

Tager TaKIPI (3)

uda	*udać*	fin : sg : ter : perf
	udo	subst : sg : gen : n
		subst : pl : nom : n
		subst : pl : acc : n
		subst : pl : voc : n
szpaka	*szpak*	subst : sg : gen : m2
		subst : sg : acc : m2

Tager TaKIPI (3)

uda	*udać*	fin : sg : ter : perf
	udo	subst : sg : gen : n
		subst : pl : nom : n
		subst : pl : acc : n
		subst : pl : voc : n
szpaka	*szpak*	subst : sg : gen : m2
		subst : sg : acc : m2

Błędy tagera

- Tager osiąga dokładność 93,4%
- Średnio co piętnaste słowo jest źle oznakowane
- Są słowa, które z reguły są źle oznakowywane

Fragmenty z korpusu IPI pan:

Przyjdzie, ciach [*ciacho*:subst:pl:gen:n] *! – klasnął w ręce*

Wielu pyta [*pyta*:subst:sg:nom:f] *nas o takie rzeczy*

Jak to działa?

Płaskie frazy składniowe

Adam Radziszewski
Politechnika Wrocławska

17 lipca 2009

O czym będzie mowa

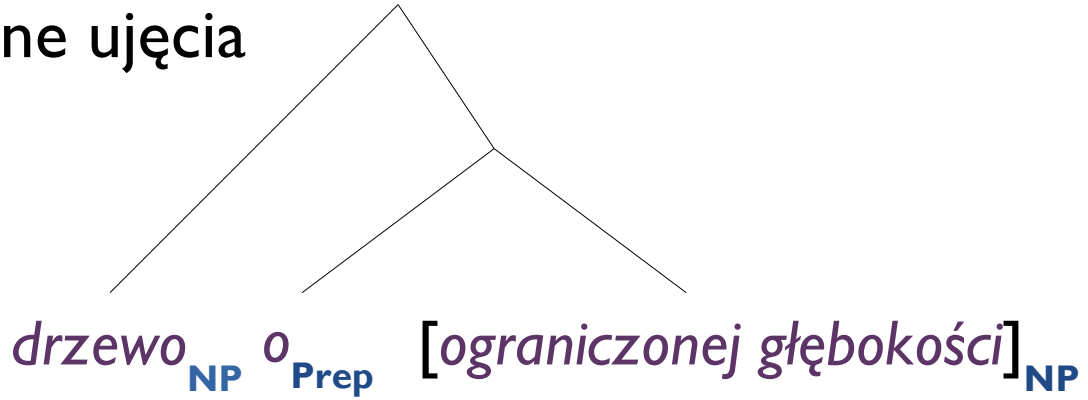
- Co daje korpus oznakowany?
- Tager
- **Płaskie frazy składniowe (*chunking*)**
  - Pełna analiza składniowa
  - Płytka analiza składniowa, *chunking*
  - Frazy rzeczownikowe w językach słowiańskich
  - Jakie frazy?
  - Problemy
- Dyskusja

Pełna analiza składniowa

- Pełna informacja o rozbiorze zdania
- Uwzględnia wszystkie rodzaje fraz / grup
- Każdy element zdania ma swoje miejsce w strukturze
- Pełna struktura, najczęściej drzewo rozbioru zdania
- Bardzo trudne zadanie
- Istniejące analizatory dają tysiące alternatywnych rozbiorów dla istniejących zdań
- Dla niektórych zdań nie potrafią dać żadnego
- Analiza jednego zdania może trwać bardzo długo

Płytka analiza składniowa

- Inżynierski kompromis
- Rezygnujemy z dokładności opisu na rzecz jego wiarygodności i wykonywalności
- Różne ujęcia



[*Kompletnie płasko*]$_{ADV}$, [ograniczony zestaw]$_{NP}$ [fraz]$_{NP}$

(*chunking*)

Chunking (1)

- Kompletnie płaskie frazy – *chunks* (*całostki*?)
- Oznaczanie całostek – *chunking* (*całostkowanie*?)
- Idea pochodzi od Stevena Abneya
  - [*I begin*] [*with an intuition*]: [*when I read*] [*a sentence*], [*I read it*] [*a chunk*] [*at a time*]
  - [*The girl*] *saw* [*a monkey*] [*with a telescope*]
- Ustalamy zbiór fraz, które nas interesują, niektórzy ograniczają się jedynie do rzeczownikowych (NP)
- Dla każdej frazy można zdefiniować zadanie całostkowania jako znalezienie początków i końców całostek

[*Taka praca*] *może zniweczyć* [*nawet najwybitniejsze talenty*]

Chunking (2)

- Dla wielu zastosowań jest to wystarczający stopień analizy
- Całostki NP pozwalają znaleźć nazwy organizacji, osób, miejsc (*byty nazwane*); podmioty i orzeczenia
- Przydatne w *wydobywaniu informacji z tekstu*
  - Duży zbiór tekstów, np. ogłoszeń o pracę
  - Zadanie wydobywania informacji, np. oferty dla tłumacza spełniające ograniczenia (np. miejsce pracy – Львів)
- Przydatne do pozyskiwania dalszej wiedzy lingwistycznej
- Może poprawić jakość tagera

Moje zadanie

- Określiłem wytyczne znakowania całostek (nieścisłe, często się zmieniają ☹)
- Ręcznie oznaczam całostki kilku typów
- Piszę program, który będzie się uczył reguł całostkowania na podstawie oznakowanego tekstu

Całostki NP a języki słowiańskie

- Chorwaci (Tadić), Serbowie (Nenadić) ograniczają zakres całostek NP do fraz uzgodnionych (ta sama liczba, rodzaj i przypadek)

[*Sztuka*] *utraciła* [*swoją moc pobudzającą*]
nom:sg:f acc:sg:f

- Ułatwia to tagowanie (skoro oznaczona całostka jest uzgodniona, wiadomo, które tagi zostawić)

[*Ministerstwo*] [*Edukacji Narodowej*] *i* [*Sportu*]
nom:sg:n gen:sg:f gen:sg:m

- Modyfikacja dopełniaczowa nie jest uwzględniana. Podobnie z koordynacją (choć są wyjątki)
- Mimo wszystko wydaje się sensowne zacząć od takich NP

Jakie frazy?

- Całostki NP czasem zawierają w sobie słowa nieodmienne

[*Kolejny krok naprzód*], [*rzecz całkiem łatwa*]

- Czasem rolę słów nieodmiennych pełnią frazy

[*zmiana* [*in plus*]], [*książki* [*w ogóle*] [*nie*] *czytane*]

- Jest więc sens znakować złożone wyrażenia nieodmienne
  - Złożone przyimki (*ze względu na*)
  - Złożone partykuły/przysłówki??? (*na przykład*)
  - Złożone spójniki??? (*na ile*)
- Frazy liczebnikowe ([*rok*] [*tysiąc dziewięćset dwunasty*])

Problemy (1)

- Czy da się jakoś sensownie zdefiniować takie całostki?
  - Poparte teorią gramatyczną
  - Praktyczne i rozstrzygalne decyzje
- Co należy do NP?
  - *To nie moja wina* (*wina nie moja* / *to nie jest moja wina*)
  - Młodzi nie czytają (przymiotnik użyty nominalnie)
  - Moja wina / jego (gen) wina (brak uzgodnienia?)
  - *Nie tylko* [*ptaki*] *latają*

Problemy (2)

- Czy klasyfikacja słów i fraz nieodmiennych pomoże określić granice NP?
  - W KIPI słowa te należą do tzw. *kublików* (*partykuło-przysłówków*), spójników bądź przyimków
  - [*krok naprzód*], [*walka wręcz*],
  - *Jest to wręcz* [*przegrana walka*], *nie* [*moja wina*],
  - *stał się* [człowiekiem], [stawanie się] [człowiekiem]
- Być może dałoby się taka klasyfikację przeprowadzić półautomatycznie

Dyskusja