Automatyczne wykrywanie koreferencji nominalnej w języku polskim

Maciej Ogrodniczuk

Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Seminarium IPI PAN

27 listopada 2017, Warszawa

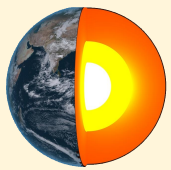
Spis treści

Opowiem:

- o pojęciu koreferencji (referencji, anafory...),
- o dotychczasowych pracach nad koreferencją w polszczyźnie i na świecie,
- o przyjętym modelu koreferencji i jego weryfikacji,
- o korpusie koreferencyjnym,
- o implementacji systemów do wykrywania koreferencji,
- o sposobach ewaluacji,
- o wynikach ewaluacji systemów dla języka polskiego,
- o dalszych planach badań.

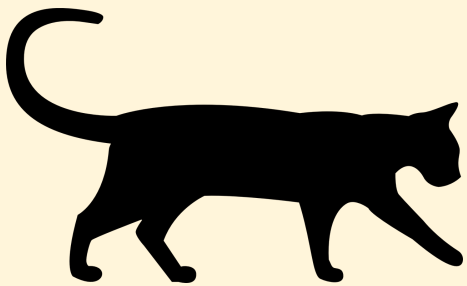
O projektach

CORE:



Komputerowe metody identyfikacji nawiązań w tekstach polskich (MNiSW, 40. konkurs na granty na badania własne, 2011–14)

COTHEC:



Ujednolicona teoria koreferencji w języku polskim i jej korpusowa weryfikacja (NCN, OPUS 8, 2015–18)

Referencja, anafora, koreferencja

Słowniczek pojęć:

- referencja,
- świat mentalny,
- koreferencja,
- koreferencyjne środki wyrazu,
- dyskurs / metatekst,
- anafora,
- wzmianki,
- klastry koreferencyjne (a nie łańcuchy koreferencyjne).

Motywacja

Dlaczego to jest ważne?

- referencja to ważny składnik analizy struktury tekstu,
- wcześniejsze badania prowadzone w czasach przedkomputerowych $\rightarrow$ potrzeba weryfikacji tekstowej, ewaluacji na szerszą skalę,
- wcześniejsze badania ograniczone tematycznie,
- zadanie ważne z perspektywy narzędziowej,
- weryfikacja ciekawych hipotez:
  - jakie środki językowe udaje się dekodować?
  - czy w językach fleksyjnych jest łatwiej?
  - jak referencja wpływa na dyskurs?

Metodologia i zakres badań

Metoda korpusowa:

- próbkowanie rzeczywistych tekstów językowych z reprezentatywnego zbioru,
- ręczne oznaczenie korpusu relacjami koreferencyjnymi zgodnie z wypracowaną taksonomią i instrukcją anotacyjną,
- tworzenie narzędzi do automatycznego wykrywania wzmianek i relacji koreferencyjnych metodami komputerowymi,
- ewaluacja jakości algorytmów na części testowej korpusu z wykorzystaniem pewnych miar jakościowych.

Koreferencja wewnątrzdokumentowa:

- dokument jako jednostka badań.

Koreferencja na świecie

Największe korpusy relacji referencyjnych:		
OntoNotes 5.0	angielski / chiński / arabski	1,6M / 950K / 300K
ANCORA-CO	hiszpański / kataloński	400K
Copenhagen Dependency Treebank	duński / angielski / niemiecki / włoski / hiszpański	5 x 100K
ANCOR	francuski	453K
GECCo	angielski / niemiecki	365K / 372K
PCC	polski	540K

Komputerowe implementacje koreferencji

Spacer przez historię:

- algorytmy regułowe, mechanizmy wnioskowania, pierwsze reprezentacje wiedzy ogólnej i lingwistycznej reguły składniowe (Hobbs 1976), teoria centrowania (Grosz 1977, Sidner 1979),
- propozycje rozwiązań o ograniczonym wykorzystaniu reprezentacji wiedzy (Mitkov 1990),
- algorytmy uczenia maszynowego (Connolly 1994, Soon 1999),
- nowe metryki: B^3 (Bagga 1998), CEAF (Luo 2005),
- stanfordzki system sit (Raghunatan 2010, Lee 2011),
- połączenie architektury sit z uczeniem maszynowym (Chen 2012, Ratinov 2012, Lee 2016),
- modele neuronowe (Wiseman 2015, Clark i Manning 2016).

Wydarzenia w temacie koreferencji

Konferencje, warsztaty, konkursy:

- MUC (Message Understanding Conference) Coreference Tasks (1995 i 1997),
- konferencja DAARC (1996–2011),
- SemEval Task: Coreference Resolution in Multiple Languages (2010),
- Modelling Unrestricted Coreference in OntoNotes (2011–12),
- Event Nugget Detection and Coreference (2015–16),
- CORBON (2016–2017),
- CRAC 2018.

Relacje referencyjne w polszczyźnie

Przegląd badań teoretycznych:

- Klemensiewicz (1937, 1968): hierarchia wskaźników nawiązania; Topolińska (1976, 1984), Paduczewa (1992): typologie grup referencyjnych: nieokreślone dla mówiącego, uniwersalne, ...
- Pisarkowa (1969): odpowiedniki nominalne pojawiają się, gdy mogłyby zawieść tradycyjne środki ujednoznaczniania,
- Grzegorczykowa (1996): znaczenie współdzielenia między nadawcą i odbiorcą wspólnego zasobu wiedzy,
- Dobrzyńska (1996): tekst ubogi w zależności referencyjne jest zwykle także ubogi stylistycznie,
- Dunin-Kęplicz (1989): analiza struktury dyskursu za pomocą metod formalnych,
- Marciniak (2002): próba implementacji elementów HPSG z interpretacją zaimków osobowych.

Analiza automatyczna w polszczyźnie

Wyłącznie częściowo:

- Mitkov (1997): wykrywanie poprzedników anaforycznych na podstawie zestawu ważonych cech,
- Kulików i in. (2004): wykrywanie zaimkowych relacji anaforycznych na podstawie zgodności liczby/rodzaju w odległości co najwyżej dwóch zdań,
- Abramowicz i in. (2006): wykrywanie koreferencji za pomocą badania podobieństwa napisów miarą Jaro-Winklera,
- Filak (2006): detektor anafory zaimkowej dla systemu GATE wykorzystujący drzewa decyzyjne,
- Broda i in. (2012): koreferencja pomiędzy uogólnionymi frazami rzeczownikowymi a nazwami własnymi,
- Kopeć (2014), Kaczmarek i Marcińczuk (2015): wykrywanie anafory zerowej.

Model relacji

Elementy modelu:

- szeroka definicja referencji:
  - referencja przysługuje także stanom, sytuacjom, miejscom i określeniom czasu — bytom wykreowanym w świecie tekstu,
  - referencyjne są także zagnieżdżone części frazeologizmów,
- typy i granice wzmianek:
  - zagnieżdżenia: dyrektor departamentu firmy,
  - frazy przyimkowo-nominalne: ustawa o podatku dochodowym,
  - zdania względne: dziewczyna, o której rozmawiamy,
  - podmioty zerowe,
  - koordynacja, grupy nominalne połączone przyimkiem: Jan z Marią przyszli na obiad.
  - nieciągłości: Tylko takie książki kupuję, które mają dużo obrazków.

Metodologia anotacji

Decyzje anotacyjne:

- anotacja szeregowa czy równoległa?
- preanotacja?
- długość tekstu?
- wybór tekstów?
- opracowanie instrukcji, faza rozpoznawcza, ...

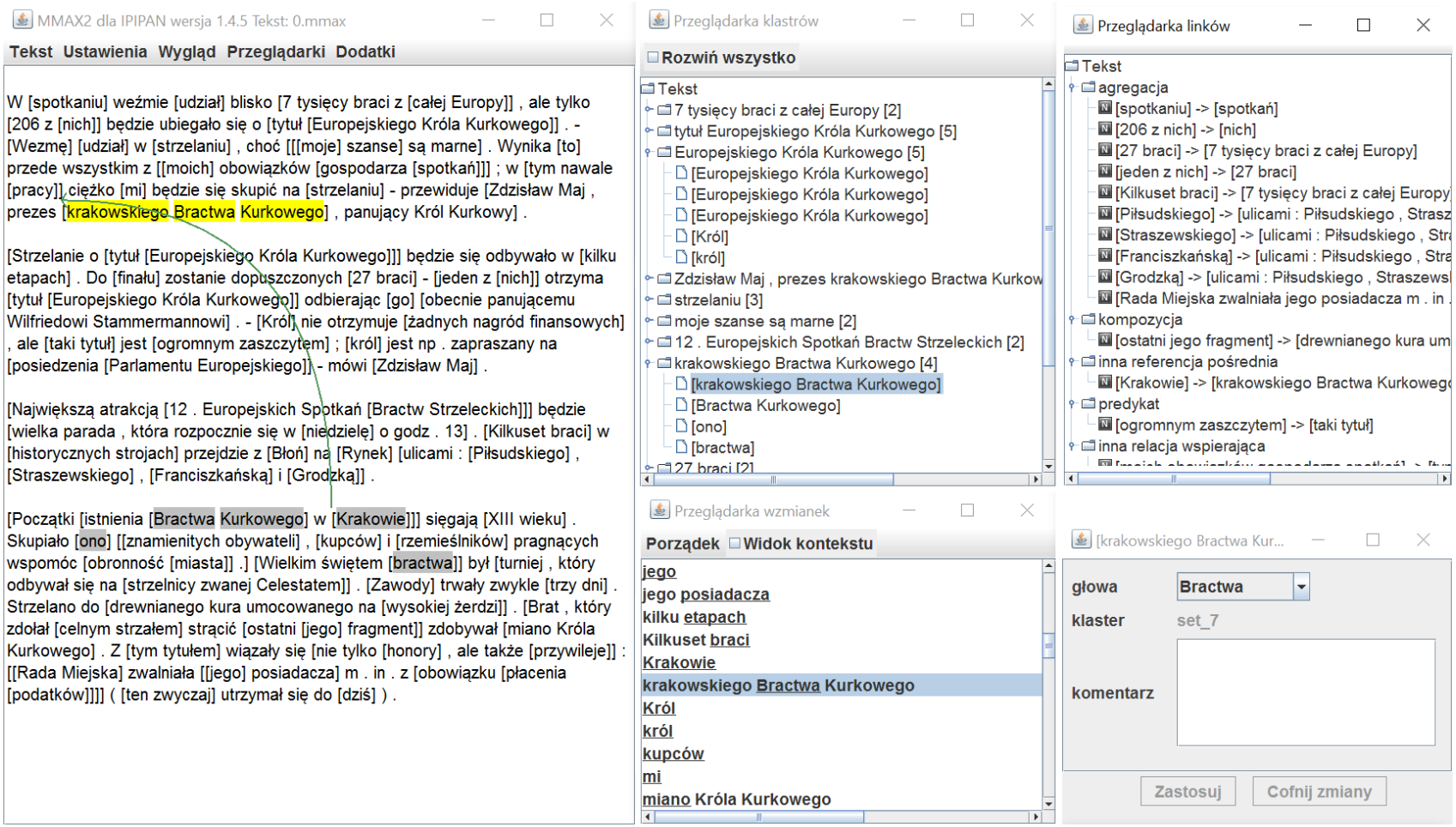
Anotacja koreferencji nominalnej

Praca anotatora:

- oznaczenie granic wzmianek,
- centra semantyczne fraz,
- powiązanie wzmianek w klastry,
- oznaczenie relacji dodatkowych,

w ramach poprawiania błędów preanotacji automatycznej.

Narzędzie anotacyjne: MMAX4REF



Korpus koreferencyjny

Teksty:

- „krótkie” — 250–350 segmentów, z NKJP,
- „długie” — 21 pełnych artykułów Rzeczpospolitej z 7 dziedzin (36K segmentów).

Podstawowe statystyki:

Tekstów	1 794
Segmentów	540 215
Wzmianek	180 432
Klastrów	128 273
– singletonowych	109 384
– pozostałych	18 889

Implementacja narzędzi

Zadania:

- wykrywanie wzmianek,
- łączenie wzmianek w klastry,
- ewaluacja.

Dwie strategie klastrowania:

- *mention-pair*: badana wzmianka dodawana jest do klastra wzmianki tekstowo wcześniejszej o najwyższym prawdopodobieństwie koreferencji,
- *entity-based*: badana wzmianka łączona jest z klastrem o najlepszej średniej ocenie (średnia z prawdopodobieństw par utworzonych z badanej wzmianki i pozostałych wzmianek w badanym klastrze),

Wykrywanie wzmianek: system regułowy

Elementy systemu:

- **Pantera/Morfeusz SGJP** — przejęcie informacji morfoskładniowej (`ppron12`, `ppron3`, `subst`, `depr`, `ger`),
- **Spejd/gramatyka NKJP z poprawkami** — frazy rzeczownikowe (`Noun`),
- **NERF** — nazwy własne z elementem rzeczownikowym,
- **wykrywanie podmiotów zerowych** (`fin`, `bedzie`, `aglt`, `praet`) w zdaniach bez grupy rzeczownikowej w mianowniku,
- **Walenty** — analiza wzmianek wchodzących w skład wielopozycyjnych schematów składniowych.

Wzmianki: system statystyczny

Etapy:

- rozpoznawanie głów (część mowy, liczba, głowa oznaczona przez Spejda, część nazwy własnej...),
- rozpoznawanie granic wzmianek: segment klasyfikowany jako część wzmianki na podstawie zestawu cech kandydata, głowy i pary,
- osobne wykrywanie podmiotów zerowych.

Wykrywanie wzmianek: metody ewaluacji

Dwie strategie:

- EXACT — wynik porównania wzmianek uwzględniający ich granice,
- HEAD — wynik porównania wyłącznie głów wzmianek.

Wykrywanie wzmianek: wyniki ewaluacji

EXACT:

Konfiguracja	P	R	F_1
Bazowy system regułowy	67,07%	67,19%	67,13%
+ Walenty	69,59%	67,85%	68,71%
System statystyczny	**73,60%**	**69,45%**	**71,47%**

HEAD:

Konfiguracja	P	R	F_1
Bazowy system regułowy	88,68%	89,37%	89,02%
+ Walenty	90,02%	88,30%	89,15%
System statystyczny	**91,95%**	**90,81%**	**91,38%**

Wykrywanie koreferencji: system regułowy

Nieliczne „bogate” cechy lingwistyczne:

Stopień zgodności par wzmianek obliczany jest od bazowej wartości 0,5 poprzez reguły:

- wymuszające zgodność liczby i rodzaju,
- zapobiegające łączeniu wzmianki zagnieżdżonej z nadrzędną,
- promujące wzmianki nominalne o identycznej zlematyzowanej postaci tekstowej,
- promujące wzmianki zaimkowe w trzeciej osobie zgodne z rzeczownikowym poprzednikiem w mianowniku.

Wykrywanie koreferencji: system regułowy

Algorytm klastrujący:

- oblicza zgodność badanej wzmianki z wcześniejszymi klastrami,
- zgodność wzmianki i klastra to maksimum zgodności wzmianki ze wszystkimi wzmiankami w klastrze,
- wzmianka dołączona jest do klastra o najlepszej zgodności przekraczającej wartość progową,
- w przypadku równoważności klastrów — do klastra zawierającego wzmiankę tekstowo najbliższą.

Koreferencja: system statystyczny

Przykłady uczące:

```
dla każdej wzmianki m:
  dla każdej wzmianki n poprzedzającej m:
    jeśli m i n są koreferencyjne:
      dla każdej wzmianki o pomiędzy n i m (bez m):
        jeśli o i m są koreferencyjne:
          stwórz z pary (o, m) przykład pozytywny
        w przeciwnym razie:
          stwórz z pary (o, m) przykład negatywny
```

Koreferencja: system statystyczny

147 cech z pięciu grup:

- cech powierzchniowych (przekształcenia postaci tekstowej),
- cech składniowych (zgodność liczby/rodzaju),
- cech semantycznych (zgodność klasy semantycznej, synonimia, hiperonimia, zgodność synonimiczna nazw własnych na podstawie Wikipedii),
- cech metatekstowych (istotność wzmianki mierzona jej pozycją, wzajemne położenie wzmianek),
- cech anaforycznych (występowanie wzmianki o identycznym centrum semantycznym w poprzednim zdaniu, odległość między wzmiankami o tym samym centrum).

Koreferencja: system sitowy

Kaskada klasyfikatorów regułowych:

- łączenie wzmianek zawierających ten sam tekst / te same lematy,
- zgodnych z automatycznie utworzonym akronimem,
- zgodnych centrów semantycznych i słów znaczących w treści wzmianki z klastrem,
- zgodnych podmiotów zerowych,
- konstrukcji rzeczownikowych i zaimkowych na podstawie ich zgodności morfoskładniowej.

Koreferencja: system neuronowy

Cechy wektora treningowego dla pary wzmianek:

- wektorowe reprezentacje centrów semantycznych wzmianek,
- średnie wektorów odpowiadających pięciu słowom przed i po każdej wzmiance oraz słów je tworzących,
- typ wzmianki (rzeczownikowa, zaimkowa, zerowa, inna),
- odległość dzieląca wzmianki (w słowach i wzmiankach), obecność w tym samym zdaniu i akapicie...

Eksperymenty:

- zwiększenie rozmiaru wektorów z 50 do 300,
- większa liczba cech z wcześniejszych systemów,
- zwiększenie liczby przykładów negatywnych,
- badanie wszystkich wzmianek (nie tylko poprzedzających),
- połączenie najlepszego modelu sitowego z neuronowym.

Wykrywanie koreferencji: metody ewaluacji

3 główne miary:

- MUC (Villain et al. 1995),
- B^3 (Bagga and Baldwin 1998),
- CEAF (Luo 2005).

Wszystkie porównują samo łączenie w klastry, przy założeniu dobrze wykrytych wzmianek.

W praktyce stosuje się średnią z powyższych miar (CoNLL).

Ewaluacja wzmianek systemowych

Dwa warianty obliczeń:

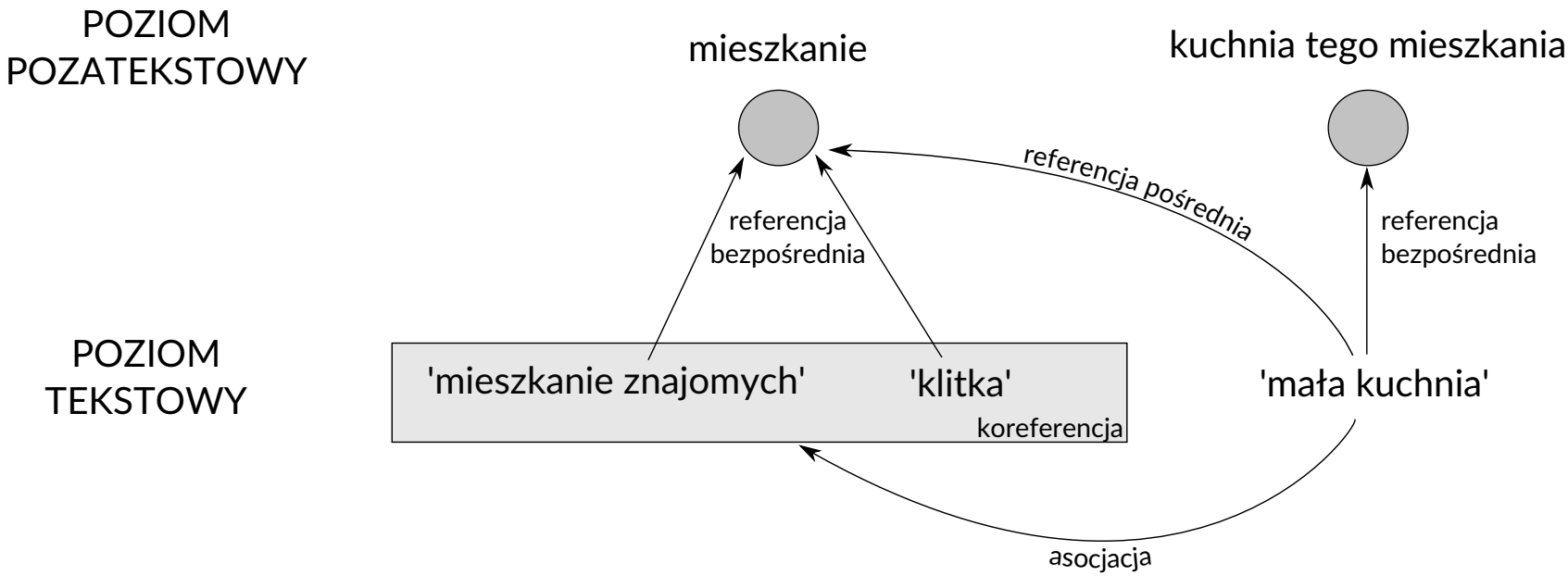
- INTERSECT — uwzględniane są tylko te wzmianki systemowe, które znajdują się w kluczu,
- TRANSFORM — zestawy wzmianek są przekształcane w celu zapewnienia identyczności zbiorów:
  - wzmianki z klucza, które nie mają odpowiedników systemowych są dodawane do wyników systemu jako singletony,
  - systemowe wzmianki singletonowe bez odpowiedników z klucza są usuwane,
  - systemowe wzmianki niesingletonowe (wchodzące w skład jakiegoś klastra koreferencyjnego) są dodawane do klucza jako singletony.

Wykrywanie koreferencji

Wyniki ewaluacji:

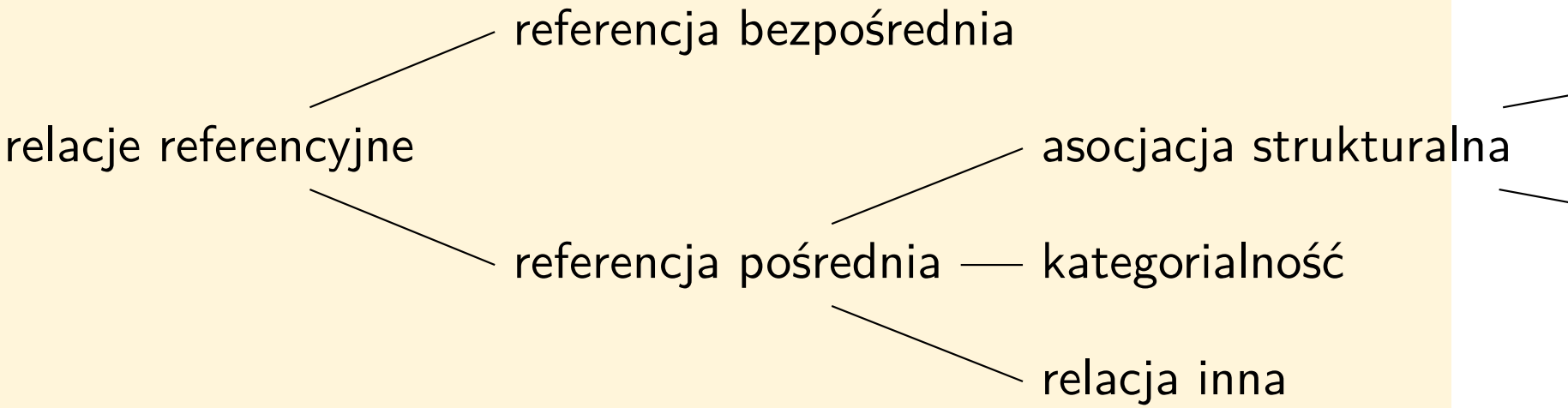
Rodzaj systemu	Miara F_1			
	MUC	B^3	CEAFE	CoNLL
System regułowy	57,63	81,76	80,00	73,13
System statystyczny	64,34	85,07	85,06	78,16
System sitowy	67,73	86,63	87,74	80,70
System neuronowy	67,23	86,59	86,89	80,24
System hybrydowy	**69,21**	**86,74**	**87,75**	**81,23**

Relacje bezpośrednie i pośrednie



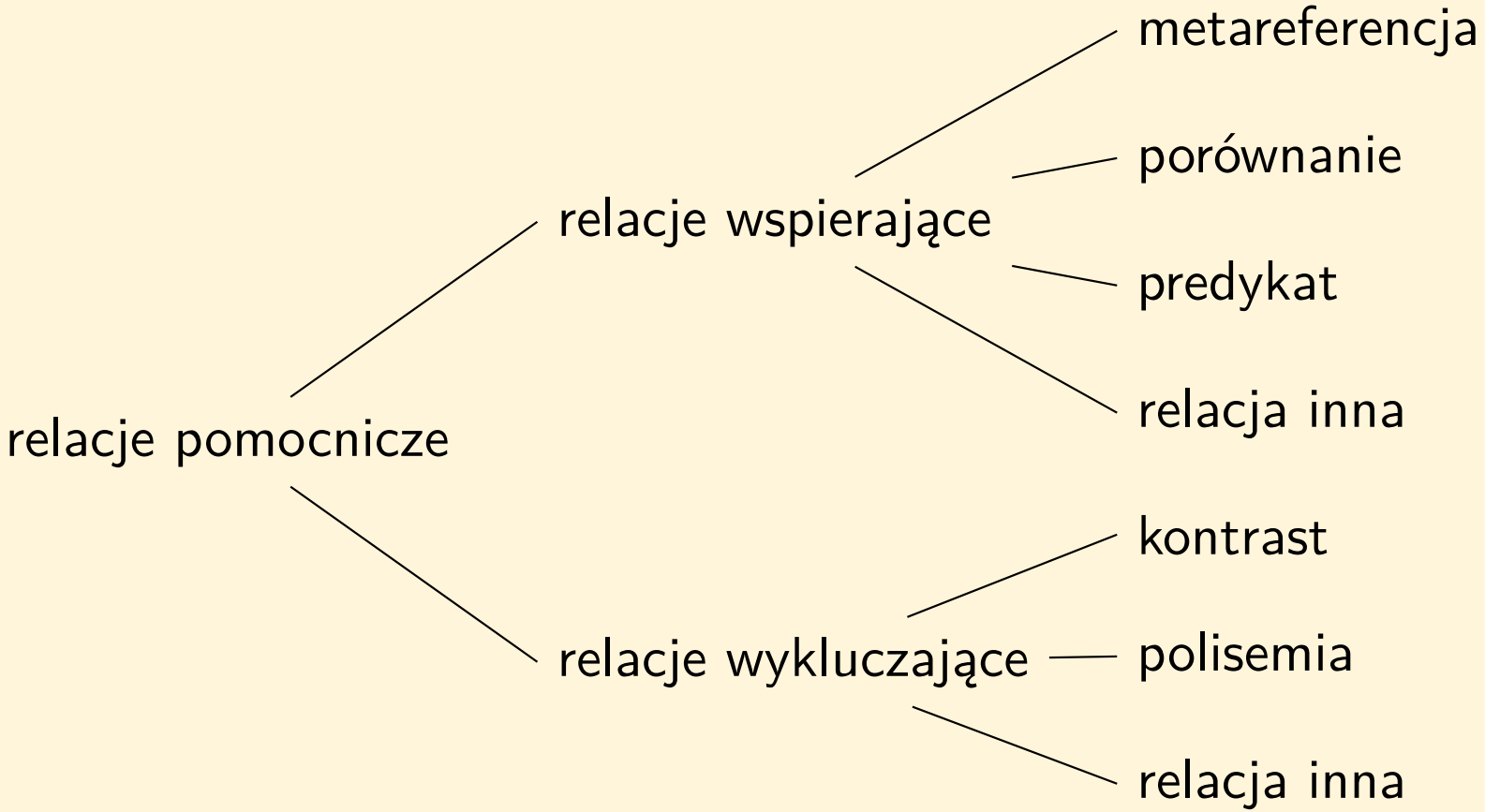
Typologia relacji referencyjnych

Relacje pośrednie:



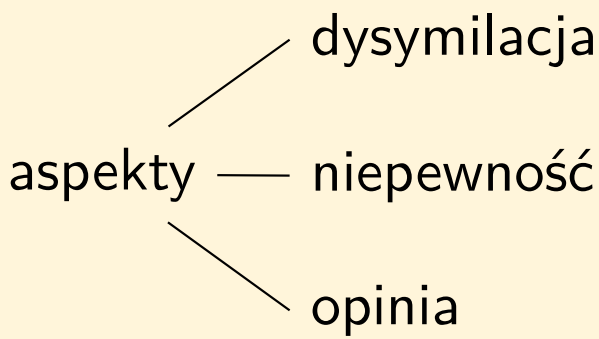
Typologia relacji referencyjnych

Relacje wspierające i wykluczające:



Typologia relacji referencyjnych

Aspekty relacji:



Koreferencja uniwersalna?

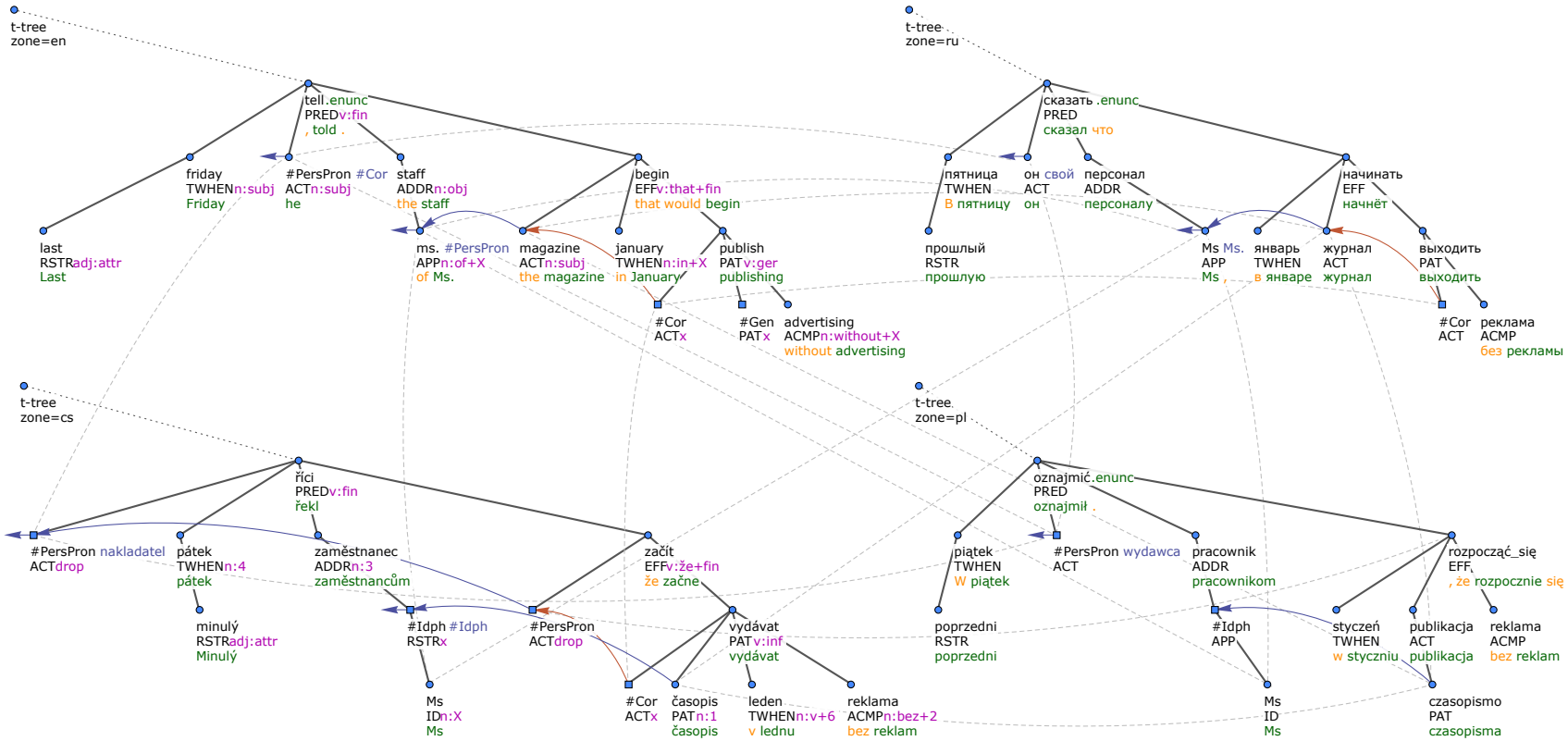
Porównanie opisu referencji w różnych językach:

- wcześniej: OntoNotes (2007–12), angielsko-niemiecki ParCor (2014), PCEDT 2.0 Coref (2016),
- PAWS 1.0: wielojęzyczny korpus równoległy tekstów angielskich, czeskich, polskich i rosyjskich,
- metoda:
  - automatyczne zrównoleglenie tekstów na poziomie słów,
  - parsowanie składniowe UD,
  - ręczna anotacja relacjami referencyjnymi na warstwie tektogramatycznej.
- rodzaje koreferencji:
  - gramatyczna (włączając zaimki względne i zwrotne),
  - tekstowa (grupy nominalne włącznie z podmiotami zerowymi, elipsy, koreferencja przysłówkowa, konstrukcje z rozdzielonym poprzednikiem),
  - odwołania do dłuższych fragmentów tekstu.

Anotacja w korpusie PAWS

EN: Last Friday, he told the staff of Ms. that the magazine in January would begin publishing without advertising.

RU: В прошлую пятницу он сказал персоналу Ms, что в январе журнал начнёт выходить без рекламы.



CS: Minulý pátek řekl zaměstnancům Ms., že časopis v lednu začne vydávat bez reklam.

PL: W poprzedni piątek oznajmił pracownikom Ms., że w styczniu publikacja czasopisma rozpocznie się bez reklam.

Wielojęzyczne porównanie relacji

Anotacja relacji metatekstowych w anotacji PDTB:

- jednostki dyskursywne połączone znacznikami,
- hierarchiczne relacje pomiędzy jednostkami jako nośniki spójności tekstu (np. Contingency $\rightarrow$ Cause $\rightarrow$ Reason: *Zabrał obraz, bo zdał sobie sprawę z jego wartości.*)
- relacje koreferencyjne – gdy nośnikiem spójności tekstu jest referent użyty w obu fragmentach.
- anotacja 6 tekstów TED (EN/DE/PL/PT/RU/TR).

W stronę badania spójności

Anotacja metatekstowa Korpusu Koreferencyjnego:

- metaoperatorami rozszerzającymi pojęcie znacznika dyskursywnego w sensie PDTB o partykuły, zaimki i ich kombinacje,
- relacjami czasowymi łączącymi zdarzenia za pomocą relatorów (wykładników jednoczesności, następstwa itp.),
- relacjami między zapisem zdarzeń komunikacyjnych a ich autorskimi kwalifikacjami (nazwami) w tekście,
- relacjami między pytaniami a odpowiedziami, z wyróżnieniem opcjonalnego zaimka pytajnego.

Dziękuję!

Prace nie udałyby się bez wsparcia:

- lingwistek — K. Głowińskiej, A. Savary, A. Wójcickiej, M. Zawisławskiej,
- informatyków — M. Kopcia, P. Morawieckiego i B. Nitonia,
- anotatorów — B. Alberskiego, A. Andrzejczuk, M. Głąbskiej, A. Grzeszak, A. Kostrowieckiej, E. Kubickiej, D. Lipińskiego, B. Milanowskiej, E. Pędzich, B. Pukalskiej, P. Rosalskiej, A. Sulicha, M. Szczyszka, D. Ziembickiego i S. Żurowskiego,
- ekspertów służących wiedzą i pomocą na różnych etapach prac: B. Dunin-Kęplicz, P. Batki, Ł. Degórskiego, Ł. Dębowskiego, Ł. Kobylińskiego, M. Lenarta, M. Marciniak, A. Mykowieckiej, A. Przepiórkowskiego, J. Waszczuka, M. Wolińskiego, A. Wróblewskiej,
- pozostałych członków Zespołu Inżynierii Lingwistycznej.