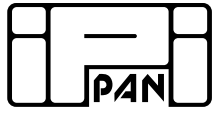
Wstępna weryfikacja typologii i strategii anotacji koreferencji w tekstach polskich

Maciej Ogrodniczuk, Katarzyna Głowińska,
Magdalena Zawisławska, Mateusz Kopeć, Agata Savary



INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. Jana Kazimierza 5, 01-248 Warszawa

Seminarium „Przetwarzanie języka naturalnego”
Instytut Podstaw Informatyki PAN

5 marca 2012, Warszawa

Spis treści

Opowiemy:

1. jak rozumiemy pojęcie koreferencji,
2. o projekcie CORE (którego celem jest wykrywanie koreferencji w tekstach polskich) – jego założeniach, zadaniach i spodziewanych wynikach,
3. o tym, co już zrobiliśmy w ramach rozpoznawania terenu: narzędziu regułowym, eksperymentach z narzędziami RARE i BART,
4. o sposobach ewaluacji jakości systemów do wykrywania koreferencji,
5. o tworzeniu korpusu nawiązań – instrukcji anotacyjnej, środowisku anotacyjnym,
6. o pierwszych wnioskach z rozpoczętego właśnie procesu anotacji.

Słowniczek pojęć

Anafora: relacja między **wystąpieniami**, której istotą jest uczytelnienie danego wyrażenia w tekście.

Słowniczek pojęć

Anafora: relacja między **wystąpieniami**, której istotą jest uczytelnienie danego wyrażenia w tekście.

Referencja: odwołanie do obiektu pozatekstowego.

Słowniczek pojęć

Anafora: relacja między **wystąpieniami**, której istotą jest uczytelnienie danego wyrażenia w tekście.

Referencja: odwołanie do obiektu pozatekstowego.

Koreferencja: relacja między wystąpieniami, której istotą jest odwołanie się do tego samego obiektu pozatekstowego.

Słowniczek pojęć

Anafora: relacja między **wystąpieniami**, której istotą jest uczytelnienie danego wyrażenia w tekście.

Referencja: odwołanie do obiektu pozatekstowego.

Koreferencja: relacja między wystąpieniami, której istotą jest odwołanie się do tego samego obiektu pozatekstowego.

Wystąpienia koreferencyjne tworzą **klaster**.
Wystąpienia, które nie są koreferencyjne, są **singletonami**.

Anafora i koreferencja nie są tym samym!

Anafora bez koreferencji:

(1) *Chcę być architektem i będę nim!*

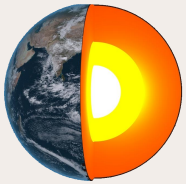
Koreferencja bez anafory:

(2) *W niedzielę w powietrze wzbił się jeden z największych samolotów pasażerskich świata, Boeing 747-8. Boeing 747-8 będzie konkurentem innego olbrzyma – Airbusa A380.*

(3) *Duszą towarzystwa był zięć Kowalskich.*
Młody prawnik właśnie wrócił ze Stanów.

CORE w pigułce

Projekt CORE



CORE

Komputerowe metody identyfikacji nawiązań w tekstach polskich

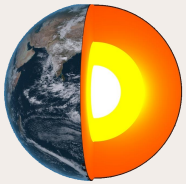
---

Projekt jest finansowany przez Narodowe Centrum Nauki (nr kontraktu 6505/B/T02/2011/40).

Ramy czasowe: kwiecień 2011 – kwiecień 2014.

CORE w pigułce

Projekt CORE



CORE

Komputerowe metody identyfikacji nawiązań w tekstach polskich

---

Projekt jest finansowany przez Narodowe Centrum Nauki (nr kontraktu 6505/B/T02/2011/40).

Ramy czasowe: kwiecień 2011 – kwiecień 2014.

Główny cel projektu:

Opracowanie technik, narzędzi oraz zasobów do automatycznej identyfikacji nawiązań w języku polskim o jakości porównywalnej z uzyskiwaną dla innych języków.

Zespół projektu CORE

Wykonawcy:

- **Maciej Ogrodniczuk** – kierownik projektu,
- **Barbara Dunin-Kęplicz** – nadzór merytoryczny i doradztwo w zadaniach lingwistycznych,
- **Adam Przepiórkowski** – ekspertyza lingwistyczna i informatyczna, współpraca projektowa z NKJP,
- **Agata Savary** – ekspertyza dotycząca anotacji korpusu, jednostek nazewniczych i wielowyrazowych oraz narzędzi do anotacji koreferencji,
- **Łukasz Dębowski** – udział w tworzeniu i rozbudowie narzędzi statystycznych,

Zespół projektu CORE

Wykonawcy:

- **Mateusz Kopeć** – główny informatyk, autor środowiska anotacyjnego,
- **Katarzyna Głowińska** – ekspertyza lingwistyczna oraz w zakresie organizacji pracy anotacyjnej,
- **Magdalena Zawisławska** – ekspertyza lingwistyczna i semantyczna, organizacja pracy anotacyjnej, superanotacja korpusu nawiązań,
- **Piotr Batko**, **Anna Grzeszak**, **Emilia Kubicka**, **Paulina Rosalska**, **Sebastian Żurowski** – anotacja koreferencji.

Główne zadania

Opracowanie:

1. typologii nawiązań na bazie dostępnych prac lingwistycznych,
2. korpusu nawiązań, identyfikującego możliwie szeroki zbiór typów nawiązań występujących w Narodowym Korpusie Języka Polskiego,
3. narzędzi informatycznych:
   - metod automatycznej identyfikacji nawiązań,
   - narzędzi realizujących te metody,
   - ocena efektywności różnych algorytmów (statystycznego i regułowego) w porównaniu z narzędziami podobnych typów dla innych języków.

Plan pracy

Zadanie 1	Określenie zakresu reprezentacji nawiązań, przygotowanie instrukcji dla anotatorów
Zadanie 2	Uzupełnienie typologii na bazie wyników projektu LUNA
Zadanie 3	Stworzenie korpusu nawiązań na bazie NKJP
Zadanie 4	Ewaluacja algorytmów i narzędzi obcojęzycznych do wykrywania nawiązań (BART, RARE, Reconcile itp.; ocena ich przydatności dla języka polskiego)
Zadanie 5	Stworzenie prototypu narzędzia regułowego do wykrywania nawiązań dla języka polskiego
Zadanie 6	Stworzenie prototypu narzędzia statystycznego do wykrywania nawiązań; ew. narzędzie hybrydowe
Zadanie 7	Ewaluacja powstałych narzędzi

Plan pracy

Zadanie 8	Rozbudowa narzędzi o moduł wykorzystujący sieć semantyczną (plWordNet) do poprawy jakości wykrywania nawiązań (hiperonimia i hiponimia; antonimia do wykluczania potencjalnych nawiązań)
Zadanie 9	Rozbudowa narzędzi o moduł wykorzystujący dostępne narzędzia i bazy faktów (Wikipedia) do poprawy jakości wykrywania nawiązań
Zadanie 10	Badanie stopnia ew. poprawy jakości streszczeń wynikającej z użycia modułu koreferencyjnego (współpraca z projektem ATLAS)
Zadanie 11	Analiza i weryfikacja uzyskanych wyników wraz z publikacją monografii

Typologia koreferencji w CORE

Interesują nas wyłącznie grupy nominalne (rozumiane szerzej niż w NKJP) powiązane relacjami:

- **identyczności** (ang. *identity of reference*),

Typologia koreferencji w CORE

Interesują nas wyłącznie grupy nominalne (rozumiane szerzej niż w NKJP) powiązane relacjami:

- **identyczności** (ang. *identity of reference*),
- **quasi-identyczności** (ang. *near-identity*), która może być realizowana np. poprzez:
  - rozmycie cech obiektu (ang. *neutralization*):

    (4) *Nie widziała „Przeminęło z wiatrem", ale czytała je.*
  - skupienie się na określonej cesze obiektu, jego aspekcie czasowym itp. (ang. *refocusing*):

    (5) *Warszawa jest pięknym miastem,*
    ale przedwojenna Warszawa była jeszcze piękniejsza.
  - ew. inne środki, które chcemy dopiero wykryć, np.

    (6) *Zdjął z półki wino i włożył je do koszyka.*

    (7) *Wyjął wino z lodówki i wypił je z gwinta.*

Jeszcze o związkach koreferencji i anafory

Wyrazem koreferencji jest m.in. istnienie w tekście wskaźników nawiązania:

- jawnych,
- zerowych – w CORE ograniczonych do tzw. podmiotów niezrealizowanych:

(8) *Janek widział dziś w teatrze Michała.*
ØMiał długą, rozwichrzoną brodę.

Interesuje nas:

- zarówno anafora, jak i katafora:

(9) — *Chodzi mi o to centrum handlowe.*
— *O Arkadię ci chodzi?*

- quasi-anafora:

(10) *Dwa lata pracowałam z Maryśką, ale już mam dość kretynki!*

Koreferencyjne środki wyrazu

Koreferencja może być wyrażana za pomocą różnych środków leksykalnych, gramatycznych i stylistycznych; oprócz zaimków wskazujących i osobowych są to np.

- synonim:

  (11) *Przed gabinetem lekarza zawsze kolejki.*
  — Bo doktor to taki miły, zagada, pożartuje.

- uogólnienie:

  (12) *Owczarki niemieckie dobrze pilnują posesji.*
  Psy te nie wpuszczą obcego za bramę.

- uszczegółowienie:

  (13) *Sklep z ubraniami znajduje się w centrum miasta.*
  Butik jest otwarty od 10 do 18.

Typologia koreferencji w CORE – cd.

Nie oznaczamy natomiast nawiązań:

- pośrednich (ang. *bridging anaphora*) – do części obiektu, elementu zbioru, nawiązań typu klasa/egzemplarz:

  (14) *Dom był maleńki. Mała kuchnia, mały pokoik.*

- kwantyfikowanych (ang. *bound anaphora*):

  (15) *Każdy prelegent musi przedstawić swój artykuł.*

- eliptycznych:

  (16) *Kupił pudełko czekoladek, ale niewiele Ø już zostało.*

- predykatywnych:

  (17) *Ewa jest nauczycielką.*

- między obiektami „tego samego rodzaju” (ang. *identity of sense*):

  (18) *Człowiek, który dał kwiaty swojej żonie, był milszy niż ten, który odmówił kupienia ich swojej.*

Strategie anotacyjne

Anotujemy:

- wszystkie frazy zagnieżdżone:

  (19) *Dyrektor departamentu firmy...*

- koordynację, grupy nominalne połączone przyimkiem:

  (20) *Jan z Marią przyszli na obiad.*
  Oni są przemili, zwłaszcza Maria.

- nieciągłości:

  (21) *Tylko takie książki kupuję, które mają dużo obrazków.*

Dla każdego wystąpienia oznaczamy głowę semantyczną,
a dla każdego klastra – wyrażenie dominujące:

- wymagane w każdym klastrze,
- wybrane spośród wystąpień lub podane przez anotatora.

Granice frazy nominalnej

Centrum frazy nominalnej jest rzeczownik, zaimek rzeczowny lub skrót. Podrzędnikami mogą być:

- rzeczowniki:

  (22) *kolega brata*
  (23) *malarz pejzażysta*

- przymiotniki i imiesłowy, np.:

  (24) *duży czerwony tramwaj*
  (25) *nadchodzące zmiany*

- frazy liczebnikowe, np.:

  (26) *zabójca pięciu kobiet*

- kubliki, np.:

  (27) *prawie cud*

Granice frazy nominalnej

Inaczej niż w NKJP podrzędnikami mogą być także:

- frazy przyimkowo-nominalne, np.

  (28) *ustawa o podatku dochodowym*

- zdania względne, np.:

  (29) *dziewczyna, o której rozmawiamy*

  nadrzędnikami zaś także:

- liczebniki, np.

  (30) *trzy rowery*

- przymiotniki (w przypadku elipsy rzeczownika), np.:

  (31) *[Podobają mi się] te niebieskie.*

Granice frazy nominalnej

Inne frazy nominalne:

- frazy opisujące daty lub godziny, np.

  (32) *12 lipca br.*
  (33) *12.07.1998*

- frazy nominalne skoordynowane, także ze spójnikiem przecinkowym, np.

  (34) *dyrektor i sekretarka*
  (35) *albo Jan, albo Maria*
  (36) *krzesło, stół i fotel*

Nowa gramatyka do wykrywania wystąpień

Gramatyka dostosowana do potrzeb projektu:

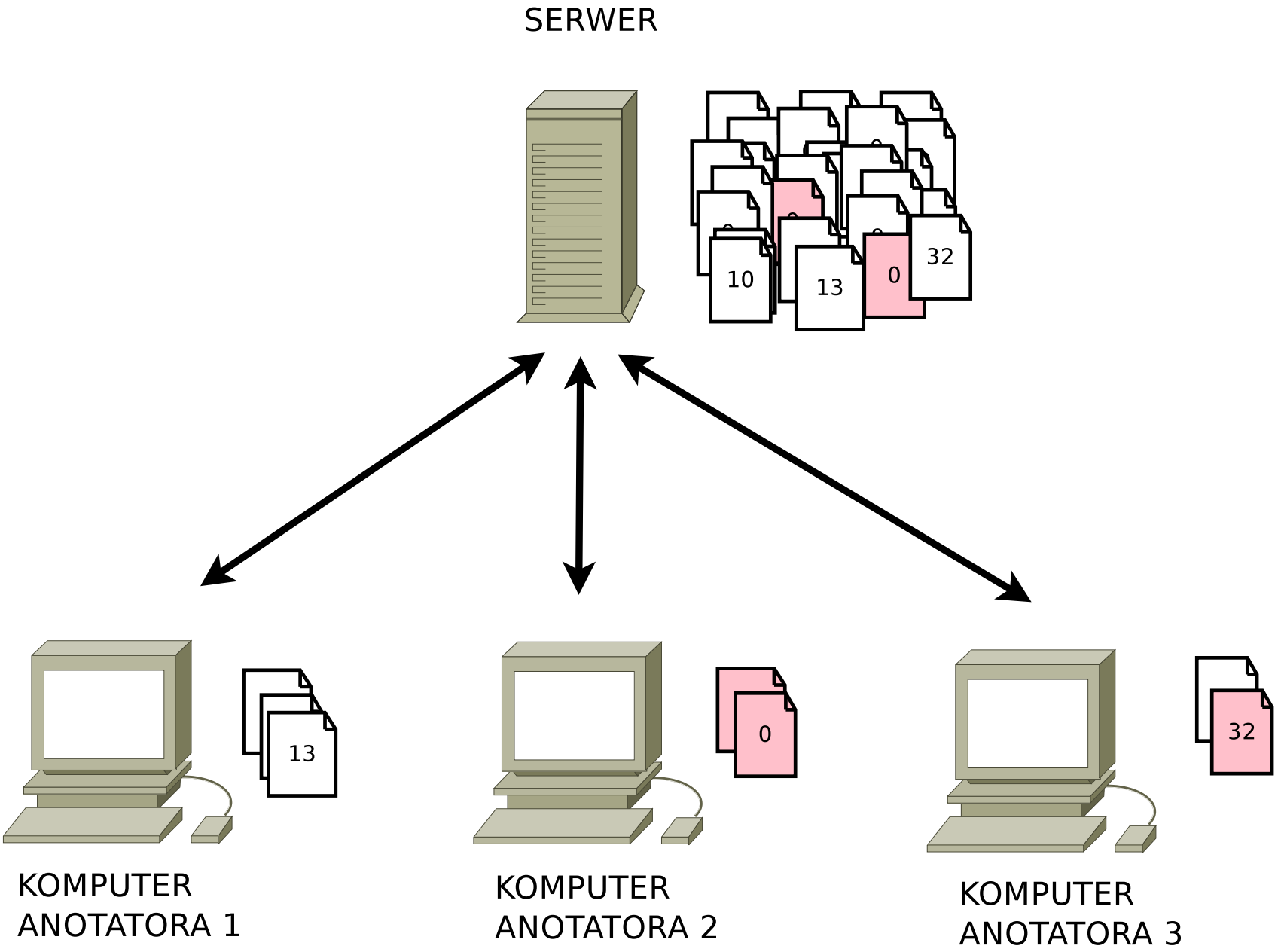
- ma wykrywać jedynie frazy nominalne (usunięcie innych grup, wydobycie fraz nominalnych z fraz przyimkowo-nominalnych),
- ma szerzej definiować frazę nominalną (zmiana fraz liczebnikowych na nominalne, włączenie zdań względnych),
- ma widzieć frazy zagnieżdżone (przebudowa całej gramatyki).

Środowisko anotacji

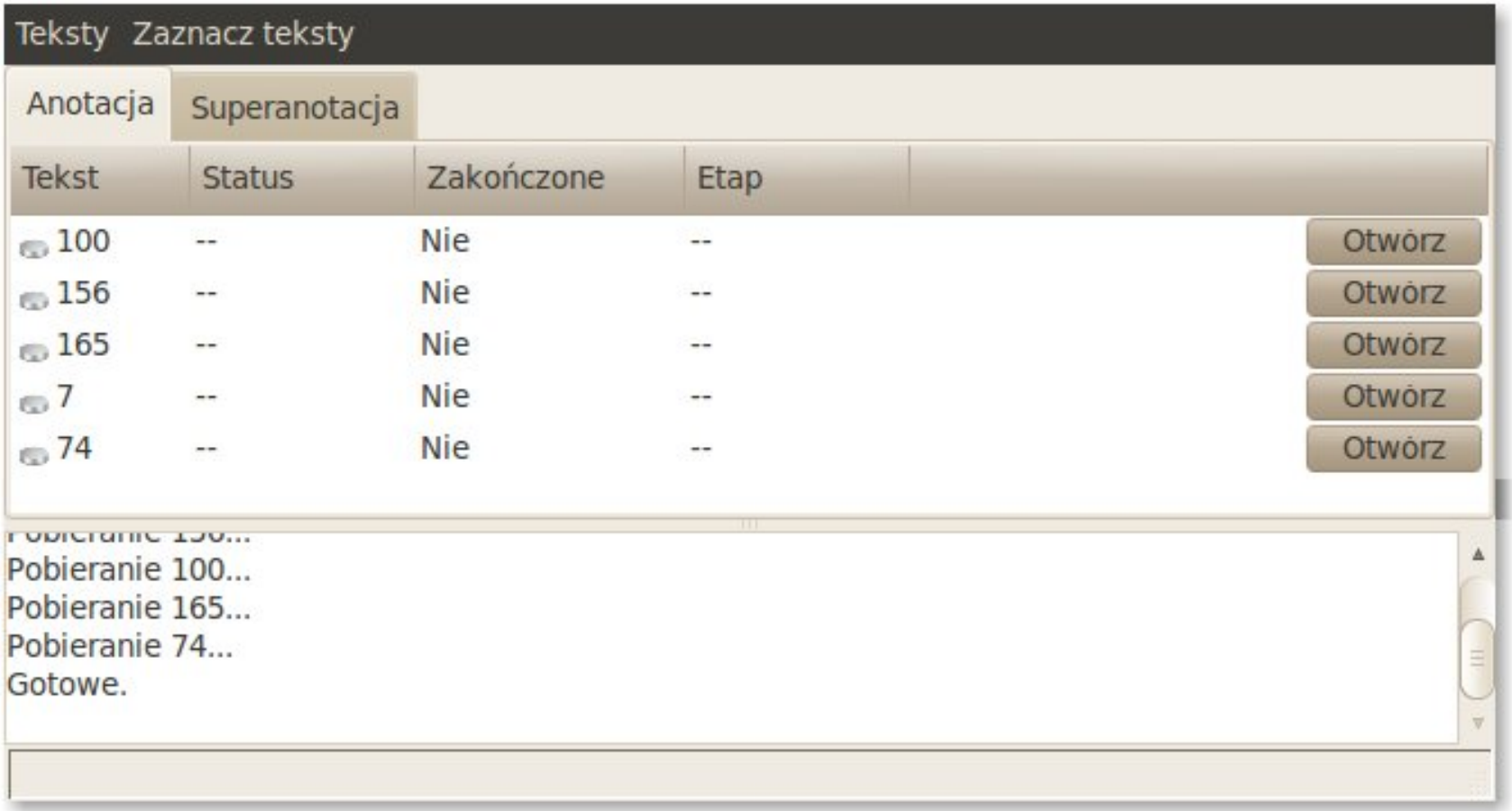
Składa się z dwóch programów:

- **manager** – do wymiany tekstów między własnym komputerem a serwerem,
  - powstały do dystrybucji tekstów podczas anotacji nazw własnych NKJP,
  - zmodyfikowany na potrzeby projektu,
- **MMAX2** – do faktycznej anotacji pojedynczych tekstów,
  - jedna z wielu dostępnych aplikacji do anotacji tekstu,
  - program desktopowy,
  - zmodyfikowany na potrzeby projektu.

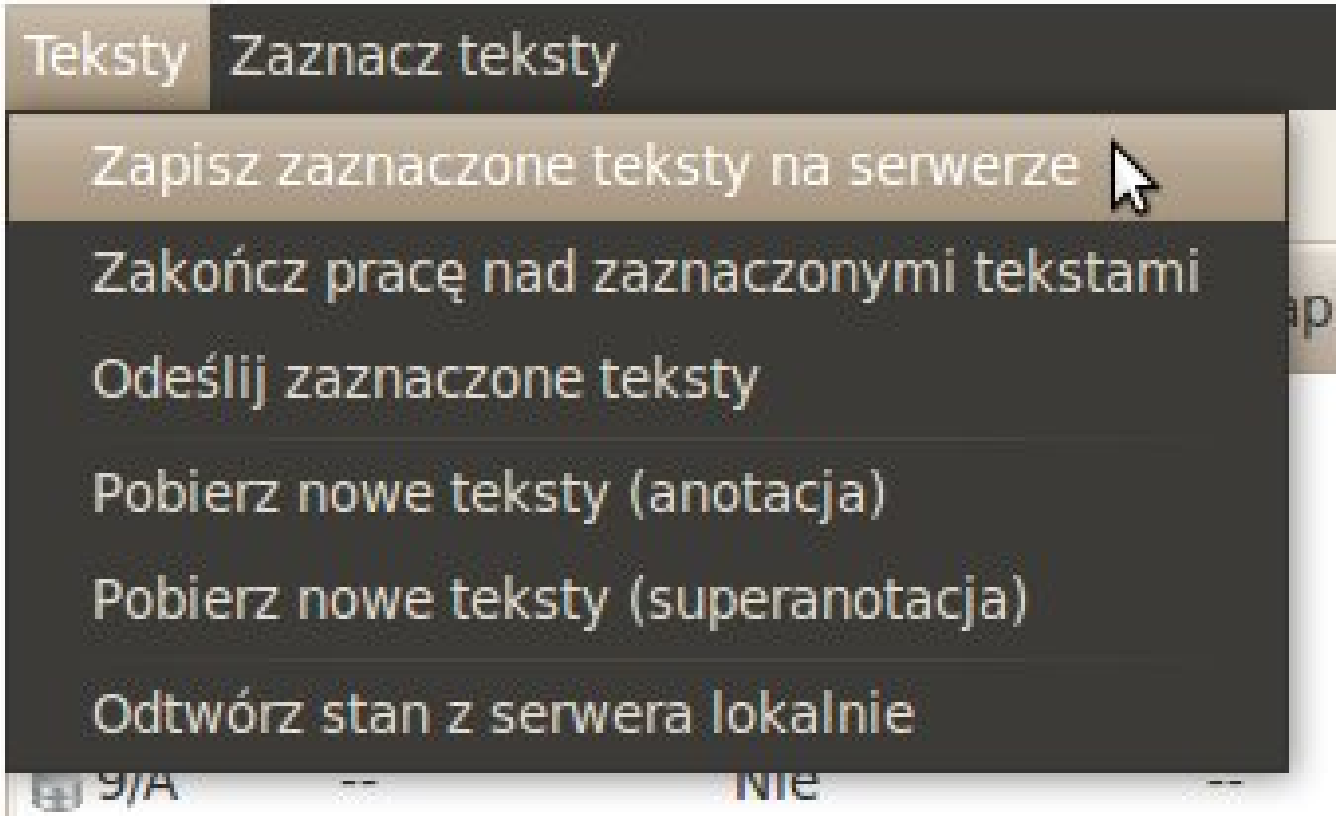
Manager: rola



Manager: interfejs użytkownika



Manager: operacje

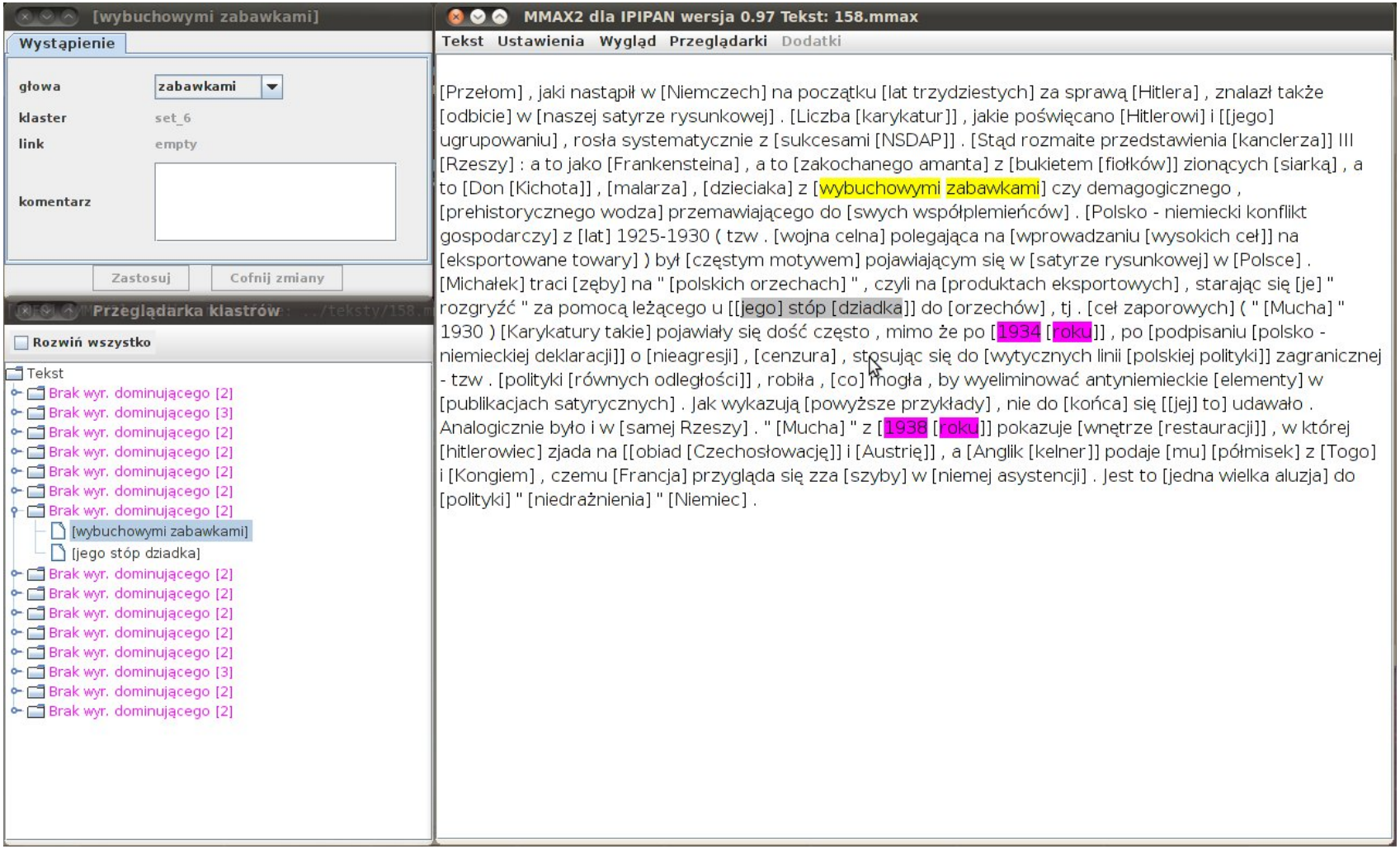


Manager: założenia

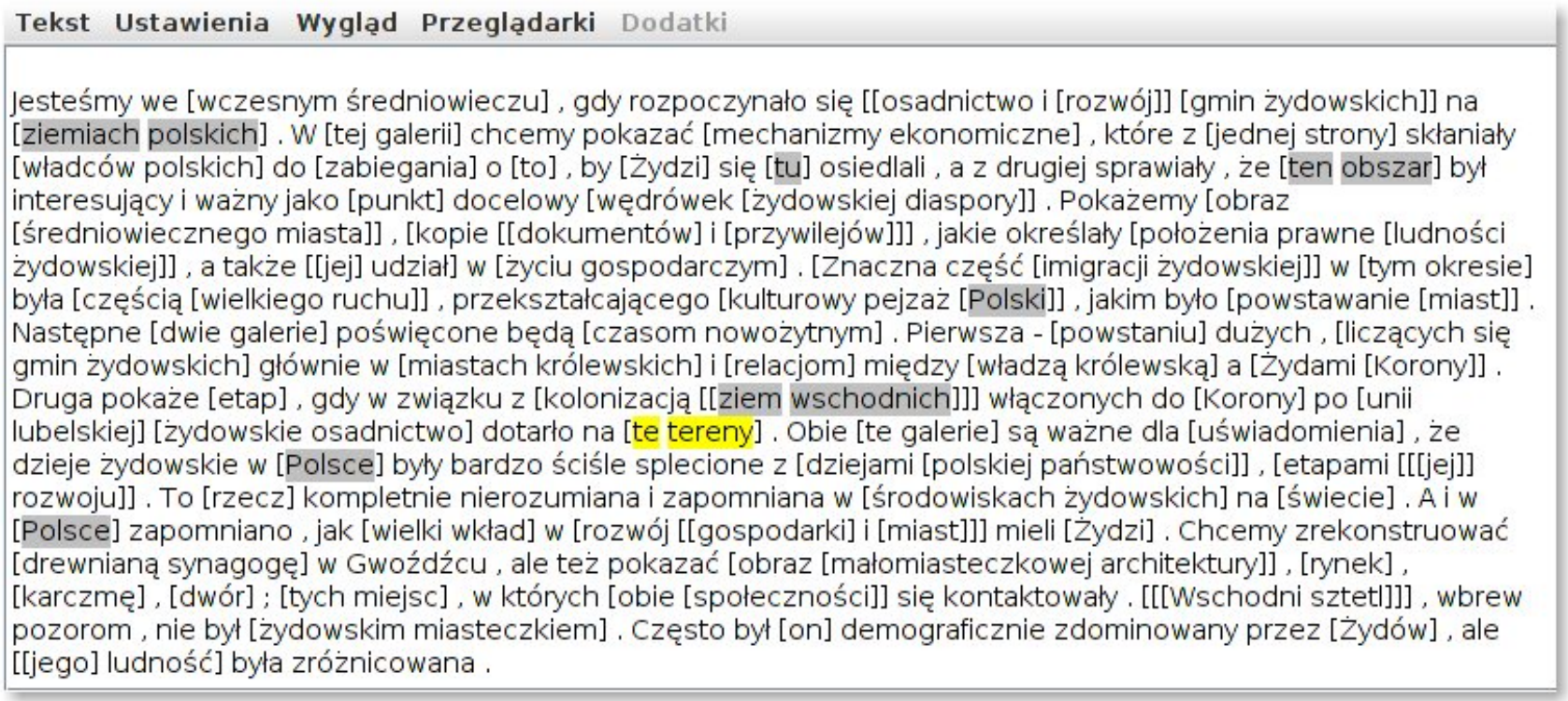
Manager zapewnia:

- losowe przyznawanie tekstów do anotacji anotatorom,
- kontrolę ilości tekstów anotowanych w danym momencie przez anotatora,
- kontrolę wersji wszystkich tekstów,
- automatyczne przekazywanie tekstów do superanotacji.

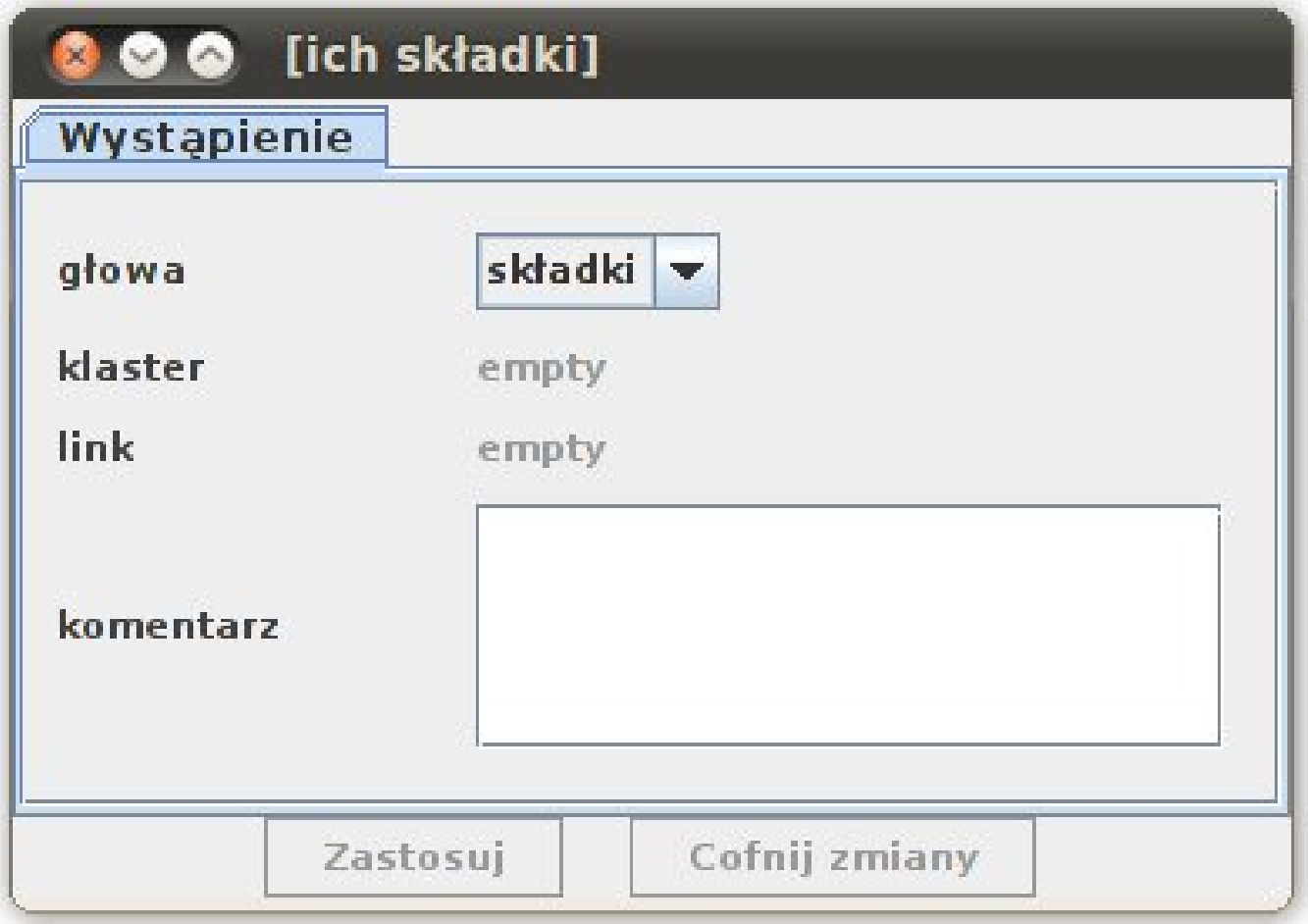
MMAX2: interfejs użytkownika



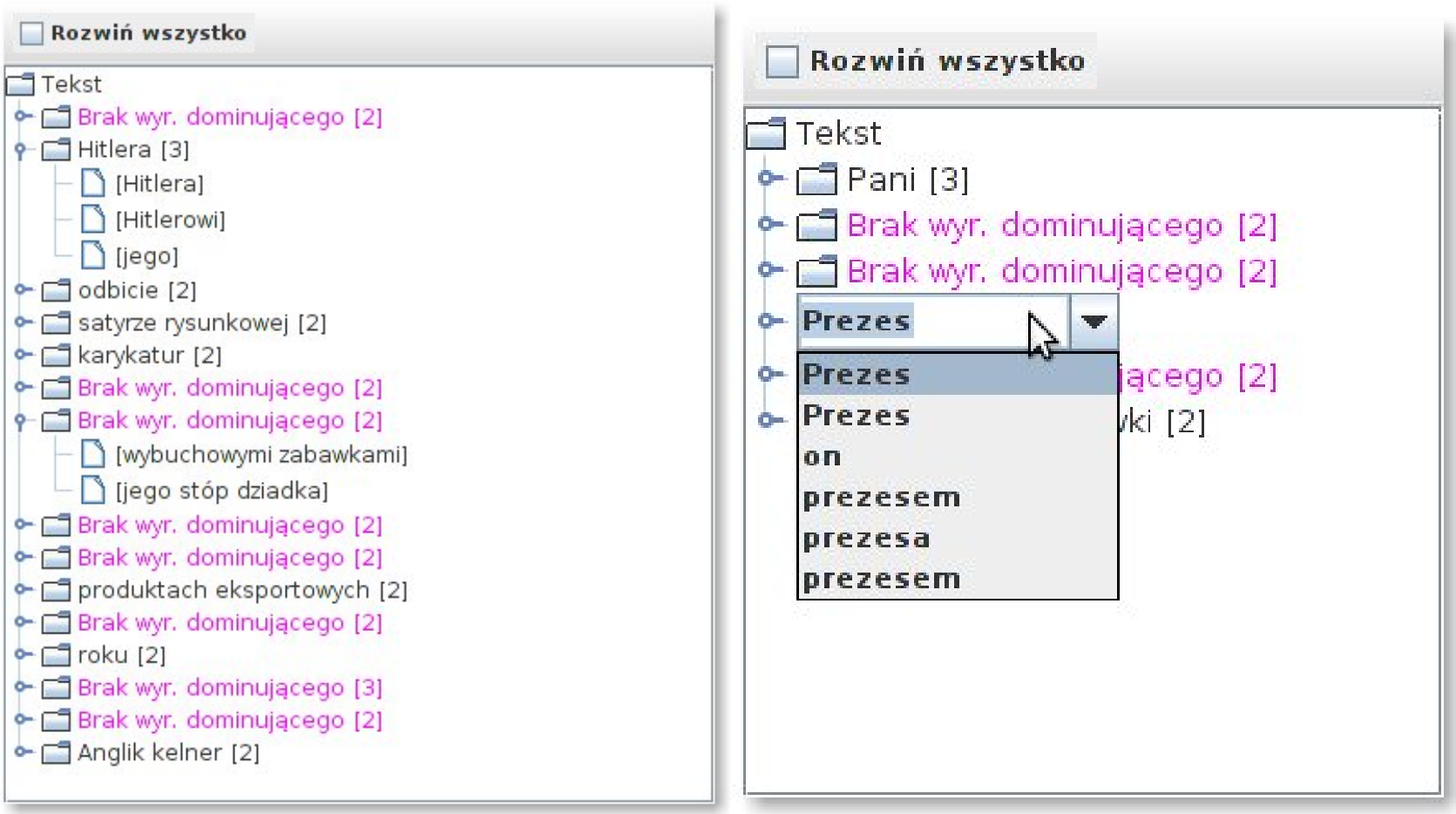
MMAX2: główne okno



MMAX2: atrybuty wystąpienia



MMAX2: przeglądarka klastrów



Praca z programem MMAX2

Dwie fazy pracy:

1. korekta wystąpień:
   - usunięcie wystapień błędnie wykrytych przez automat,
   - poprawienie błędnie oznaczonych granic wystąpień,
   - dodanie brakujących wystąpień
   - ustawienie prawidłowej głowy każdego wystąpienia.
2. korekta klastrów wystąpień:
   - usunięcie z klastrów wystąpień dodanych błędnie,
   - dodanie do istniejących klastrów wystąpień, które powinny się w nich znaleźć,
   - połączenie osobnym linkiem wystąpień quasi-identycznych,
   - oznaczenie wyrażeń dominujących.

Droga do automatycznej preanotacji

Stworzone narzędzia/zasoby:

1. Korpusik do ewaluacji
2. Prosty system regułowy do rozpoznawania klastrów
3. Moduł automatycznie rozpoznający wystąpienia

Korpusik do ewaluacji

Dane testowe:

- 15 fragmentów tekstów,
- 1737 wystąpień,
- 1262 klastrów,
- średnia wielkość klastra: 1,37 wystąpienia.

Rozmiar klastra	1	2	3	4	5	6	7..10	11..27
Liczba klastrów	1079	88	43	20	9	6	2..5	1

Miary ewaluacji systemów wykrywania koreferencji

4 główne miary

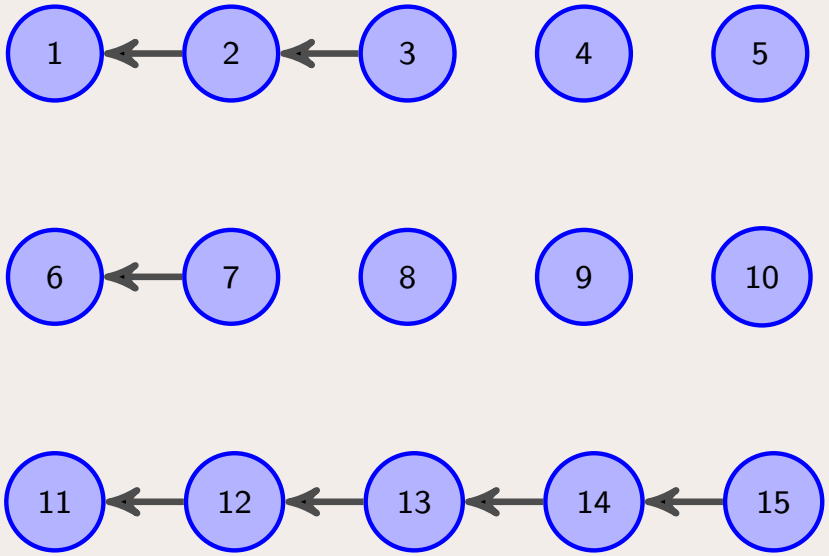
- MUC,
- B^3,
- CEAF,
- BLANC.

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

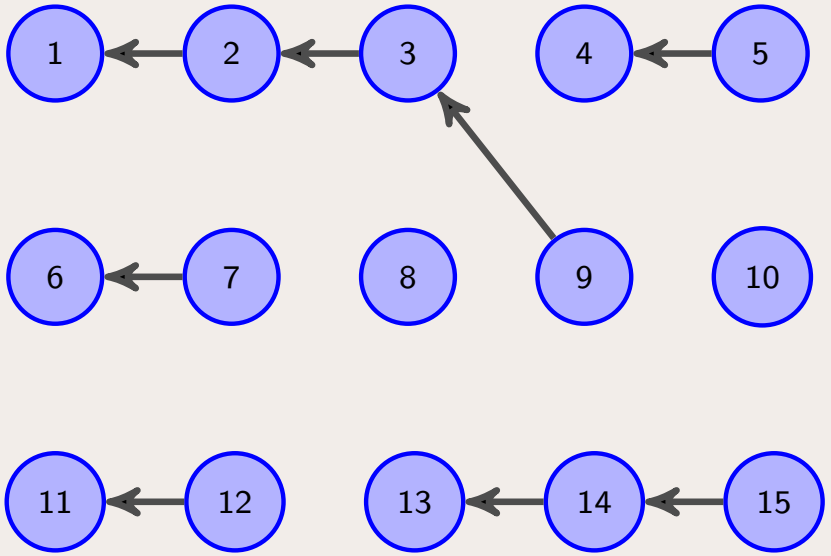
Idea

Rozpatrywanie minimalnej liczby brakujących/nadmiarowych połączeń w grupach.

Złoty standard



Wynik systemu



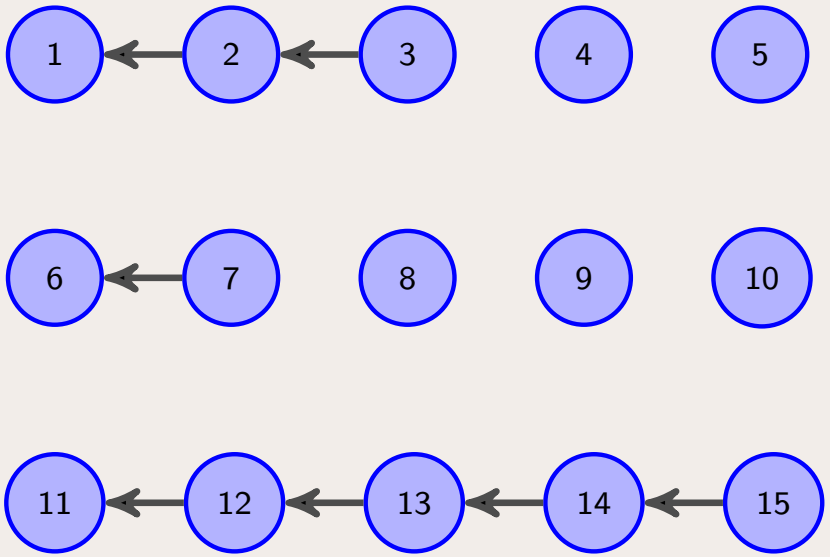
Kompletność: $\frac{6}{7} = 0,86$

Dokładność: $\frac{6}{8} = 0,75$

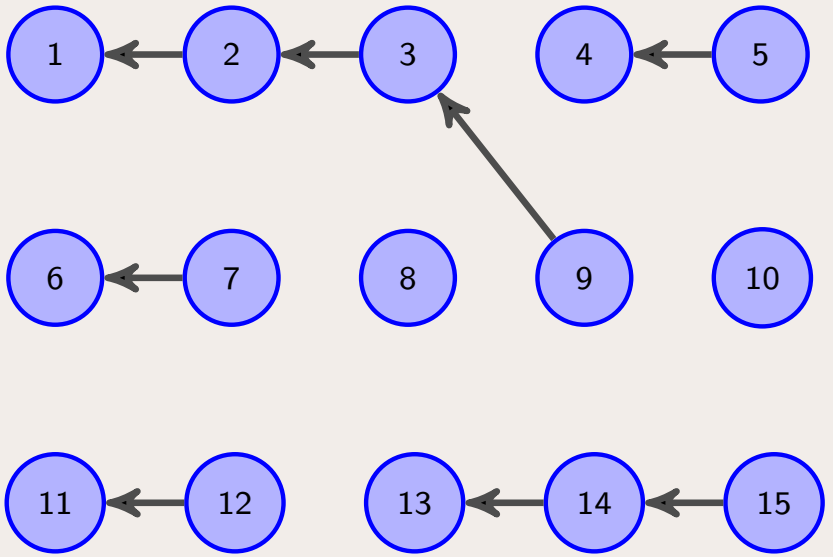
Idea

Dokładność, jak i kompletność obliczane są jako średnia z naturalnych wyników dla każdego wystąpienia.

Złoty standard



Wynik systemu



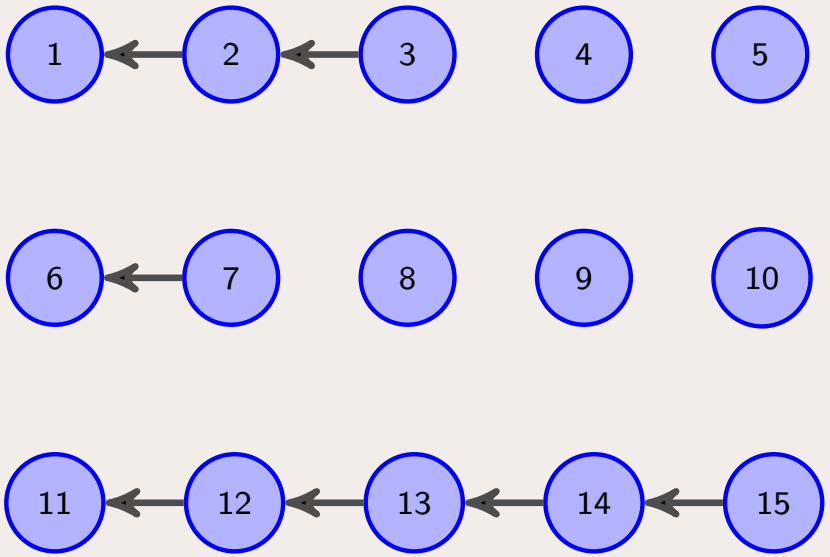
Kompletność: $\frac{63}{75} = 0,84$

Dokładność: $\frac{50}{60} = 0,833$

Idea

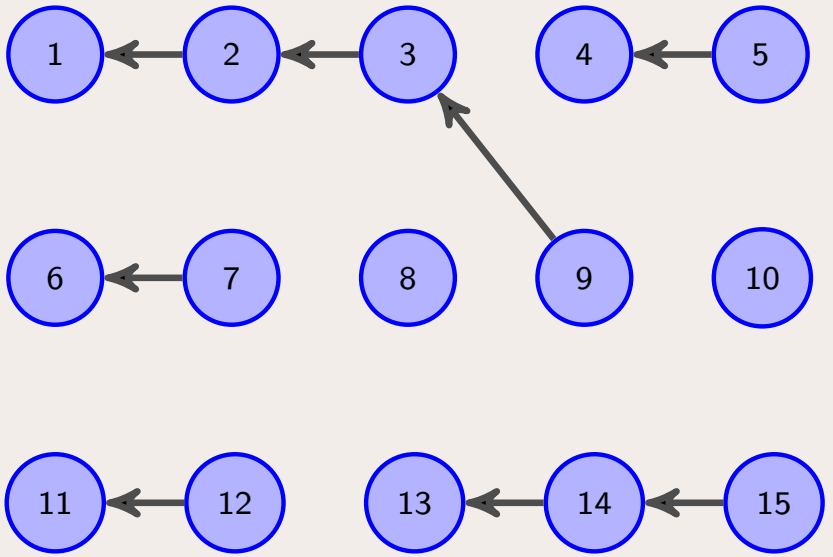
Przypisanie jeden-do-jednego klastrów wynikowych i złotego standardu podczas ich porównywania.

Złoty standard



Kompletność: $\frac{11}{15} = 0,73$

Wynik systemu



Dokładność: $\frac{11}{15} = 0,73$

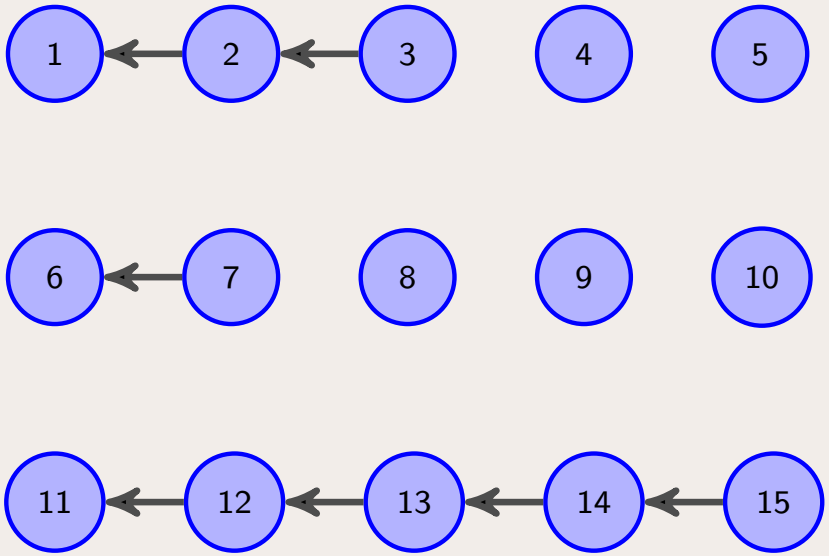
		SYS	
		Koreferentne	Niekoreferentne
GOLD	Koreferentne	rc	wn
	Niekoreferentne	wc	rn

Wynik	Koreferencja	Niekoreferencja	
P	$P_c = \frac{rc}{rc + wc}$	$P_n = \frac{rn}{rn + wn}$	$BLANC\text{-}P = \frac{P_c + P_n}{2}$
R	$R_c = \frac{rc}{rc + wn}$	$R_n = \frac{rn}{rn + wc}$	$BLANC\text{-}R = \frac{R_c + R_n}{2}$
F1	$F_c = \frac{2P_cR_c}{P_c + R_c}$	$F_n = \frac{2P_nR_n}{P_n + R_n}$	$BLANC\text{-}F = \frac{F_c + F_n}{2}$

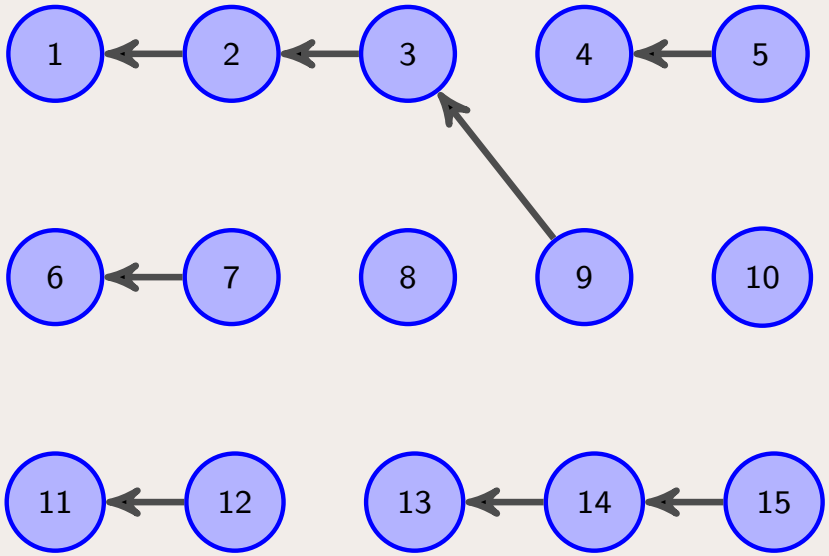
Idea

Liczenie połączenia oraz braku połączenia między wszystkimi parami wystąpień.

Złoty standard



Wynik systemu



Kompletność: $\frac{139}{182} = 0,76$

Dokładność: $\frac{149}{186} = 0,80$

Eksperyment 1: system regułowy detekcji klastrów

Wykorzystujemy nieliczne „bogate” cechy lingwistyczne (na bazie założeń Haghighiego i Kleina):

1. **ograniczenia składniowe** (wykluczanie zagnieżdżonych grup nominalnych),
2. **filtry składniowe** (eliminacja niezgodności składniowej głów),
3. **filtry semantyczne** (na bazie Słowosieci),
4. **wybór** (na podstawie wag).

Efekt:

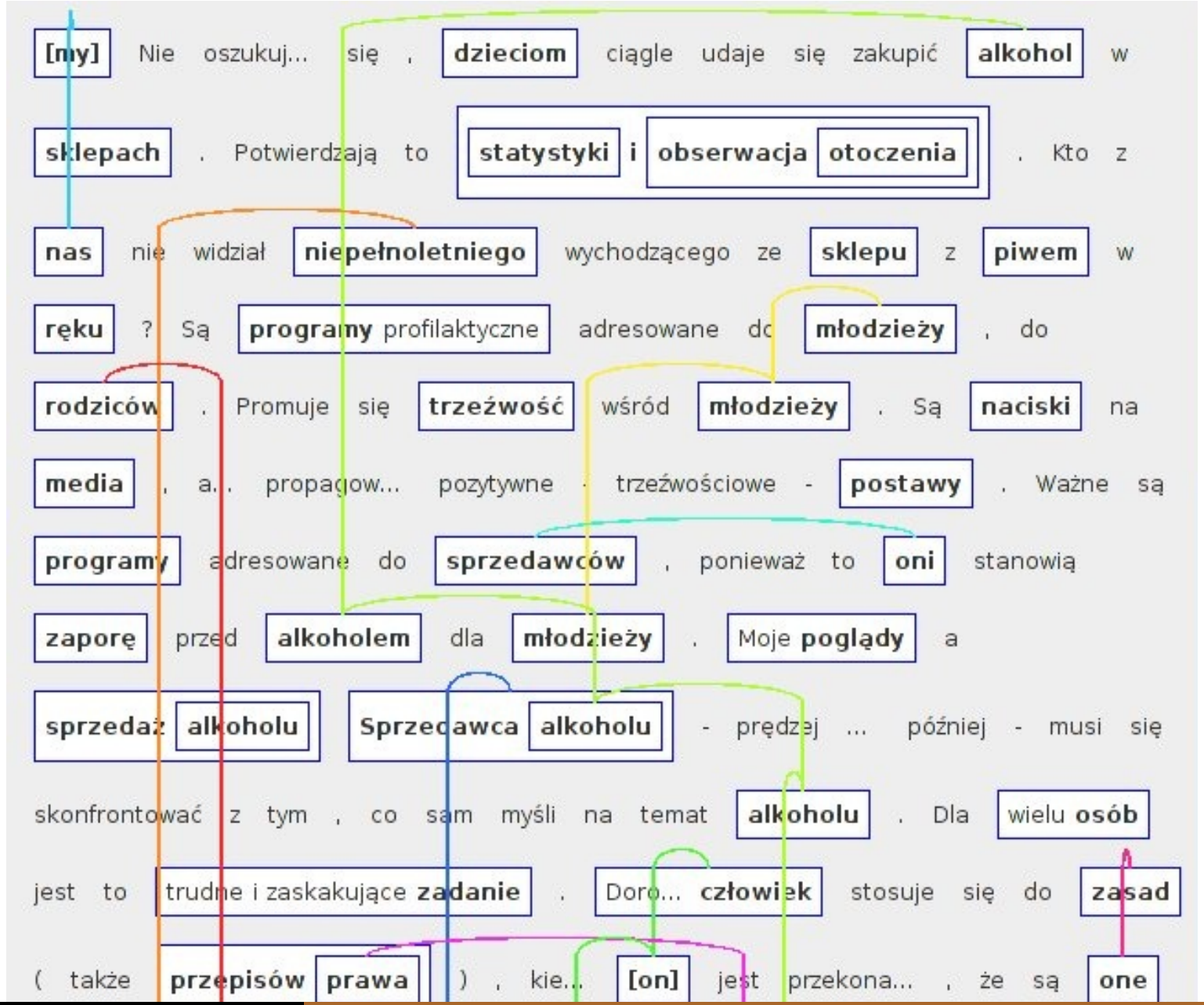
- zaprezentowany na konferencji DAARC 2011,
- wykorzystany do preanotacji aktualnie tworzonego korpusu.

Eksperyment 1: zestaw reguł

Użyte reguły:

1. eliminujemy wystąpienia o niezgodnym **rodzaju/liczbie** głów,
2. eliminujemy **zagnieżdżenia** fraz rzeczownikowych,
3. promujemy wystąpienia o zgodnych **lematach** głów,
4. badamy **wordnet**: synonimy, hiperonimy, alternimy i fuzzynimy,
5. promujemy zgodne **zaimki**.

Eksperyment 1: prototyp wizualizacji



Eksperyment 1: wyniki

Zbiór reguł	**MUC**			**CEAF**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	–			**93,10%**	67,64%	78,35%
All-sing. + head m.	50,73%	61,16%	55,46%	84,22%	79,14%	81,60%
5 rules	**75,36%**	59,46%	66,48%	78,62%	87,42%	82,79%
4 rules (no wordnet)	74,73%	**65,13%**	**69,60%**	83,45%	**88,36%**	**85,84%**

	B^3			**BLANC**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	72,65%	**100,00%**	84,16%	50,00%	49,18%	49,58%
All-sing. + head m.	84,17%	90,05%	87,01%	69,64%	**84,54%**	74,97%
5 rules	**90,56%**	82,56%	86,37%	**81,99%**	78,39%	80,08%
4 rules (no wordnet)	90,35%	86,66%	**88,47%**	81,94%	83,92%	**82,90%**

Eksperyment 2: Moduł automatycznie rozpoznający wystąpienia

Użyte narzędzia i zasoby:

- **Pantera/Morfeusz SGJP** — przejęcie informacji morfoskładniowej,
- **Spejd/gramatyka Kasi Głowińskiej** — frazy rzeczownikowe,
- **NERF** — nazwy własne.

Efekt:

W połączeniu z systemem regułowym:

- zaprezentowany na konferencji LTC 2011,
- wykorzystany do preanotacji korpusu.

Eksperyment 2: wyniki (bez anafory zerowej)

Zbiór reguł	**MUC**			**CEAF**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	–			**85,93%**	58,15%	69,36%
All-singl. + head m.	58,24%	**48,08%**	52,68%	76,61%	69,42%	72,84%
5 rules	**65,20%**	43,32%	52,05%	71,49%	70,59%	71,03%
4 rules (no wordnet)	64,43%	47,34%	**54,58%**	75,70%	**71,60%**	**73,59%**

	B^3			**BLANC**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	69,58%	**80,92%**	74,82%	50,00%	46,45%	48,16%
All-singl. + head m.	81,15%	71,14%	**75,81%**	53,95%	**79,34%**	55,54%
5 rules	**82,64%**	65,91%	73,33%	54,20%	72,48%	55,86%
4 rules (no wordnet)	82,42%	69,24%	75,26%	**54,26%**	77,60%	**56,03%**

Wykrywanie wystąpień *in vitro*

- Z zerową anaforą: R: 83,82%, P: 78,71%, F1: 81,18%
- Bez zerowej anafory: R: 88,86%, P: 78,71%, F1: 83,48%

Eksperyment 3: system „statystyczny”

Cel:

Stworzyć system do wykrywania wystąpień i koreferencji w tekstach polskich oparty na zasadach maszynowego uczenia się (z wykorzystaniem jednego z dostępnych środowisk).

Działania:

- wybór systemu: BART,
- dostosowanie go do języka polskiego.

Efekt:

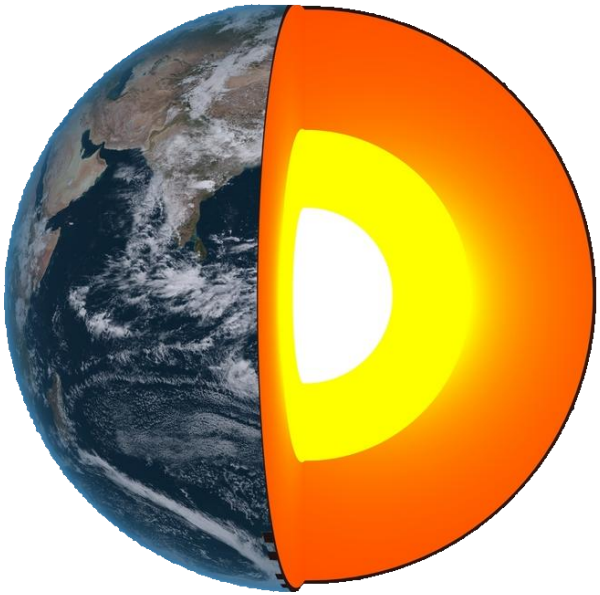
- będzie prezentowany na konferencji LREC 2012,
- wyniki systemu bliskie rezulatom systemu regułowego.

Współpraca międzyprojektowa

Możliwości współpracy:

- **NKJP** – dołączenie koreferencji jako kolejnego poziomu anotacji,
- **ATLAS** – udział w zadaniu dotyczącym streszczania (którego częścią jest identyfikacja nawiązań),
- **CLARIN** – włączenie narzędzia do identyfikacji nawiązań w infrastrukturę webserwisową,
- **CESAR** – zamieszczenie korpusu nawiązań i powstałych narzędzi w repozytorium META-SHARE.

Dziękujemy!



CORE