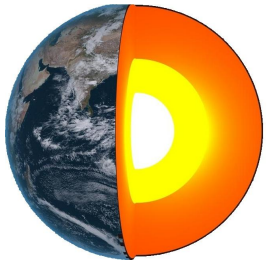
Polski Korpus Koreferencyjny – wersja wstępna

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

Seminarium "Przetwarzanie języka naturalnego", Instytut Podstaw Informatyki PAN, 3 grudnia 2012

Spis treści

O projekcie



CORE

Komputerowe metody identyfikacji nawiązań w tekstach polskich

Projekt finansowany przez Narodowe Centrum Nauki (nr kontraktu 6505/B/T02/2011/40), 2011-2014.

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

Podstawowe pojęcia

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

Wystąpienia koreferencyjne tworzą **klaster**.

Przedmiotem opisu są grupy nominalne powiązane relacją **identyczności** lub **niepełnej identyczności**.

Warszawa *jest pięknym miastem,*
ale ***przedwojenna Warszawa*** *była jeszcze piękniejsza.*

Zdjął z półki ***wino*** *i włożył* ***je*** *do koszyka.*
Wyjął ***wino*** *z lodówki i wypił* ***je*** *z gwinta.*

Stan projektu

- Korpus koreferencyjny prawie gotowy (zgodny z założeniami)
- Dwa prototypy narzędzi: regułowe i statystyczne
- Trwają badania nad:
  - definicją pojęcia identyczności
  - rolą relacji nadawca/odbiorca w procesie ustanawiania i odczytywania koreferencji
  - wykorzystaniem logiki 4-wartościowej do weryfikacji stopnia wiedzy o nawiązaniach

Schemat anotacji

Zakres anotacji:

- nawiązania bezpośrednie + quasi-anafora
- wskaźniki jawne i zerowe (tylko podmiot niezrealizowany)
- frazy zagnieżdżone, nieciągłe, frazeologizmy

  [[Jan Kowalski], [syn [Juliusza]], [ojciec [trojga dzieci]] [słowo [honoru]]
- szeroko pojęte frazy nominalne:

  m.in. z nadrzędnikiem liczebnikowym, ze zdaniem względnym, z frazą przyimkową

Schemat anotacji cd.

Zadania anotatora:

- oznaczenie wystąpień (fraz nominalnych)
- połączenie w klastry identycznościowe
- utworzenie linków quasi-identycznościowych
- wskazanie głowy semantycznej
- wyznaczenie wyrażenia dominującego

Wyrażenie dominujące

- wybrane spośród istniejących fraz:

dominanta: *planowana likwidacja kolejowych połączeń pospiesznych w obrębie miasta Rybnika i Subregionu Zachodniego Województwa Śląskiego*

klaster: *dokonywana redukcja połączeń, planowanej likwidacji kolejowych połączeń pospiesznych w obrębie miasta Rybnika i Subregionu Zachodniego Województwa Śląskiego*

dominanta: *anioły*

klaster: *aniołów, istot duchowych, niecielesnych, które Pismo Święte nazywa zazwyczaj aniołami, aniołami, aniołów, aniołowie*

- wpisane przez anotatora:

dominanta: *oni*

klaster: *usunęli, uratowali*

dominanta: *popis Hawranka*

klaster: *tę jego sztuczkę, ten popis*

Wyrażenie dominujące – problemy

Dominanta: *powołanie spółki zarządzającej zakładowym portem*

Klaster: *powołaniu spółki zarządzającej zakładowym portem, utworzeniu spółki "Port Morski w Policach sp. z o.o."*

Dominanta: *firma ojca*

Klaster: *drugą cukiernię, zwaną "Pod Filarami", firmie*

Dominanta: *dziewczynka*

Klaster: *dziesięcioletnią dziewczynkę, rannemu dziecku, Sanda, jej, dziecka, córkę, ją, dziewczynki, ją, dziecko, Sanda, Sandę, ją, dziesięciolatka, dziewczynkę, pogryzione dziecko*

Dominanta: *biurowiec przy Mysiej 5*

Klaster*: nowoczesny biurowiec z parkingiem podziemnym i szybkobieżnymi windami, biurowca przy Mysiej 5, jego, on, budynku, Nowy budynek, nowego gmachu, budynku, gmachem, nowym biurowcu, on, tego obiektu*

Organizacja korpusu

- Długość tekstów: 250-350 segmentów
- Tekst – ciąg kolejnych akapitów losowo wybrany z losowego tekstu z "dużego" NKJP
- Zrównoważenie jak w NKJP (14 typów tekstów)

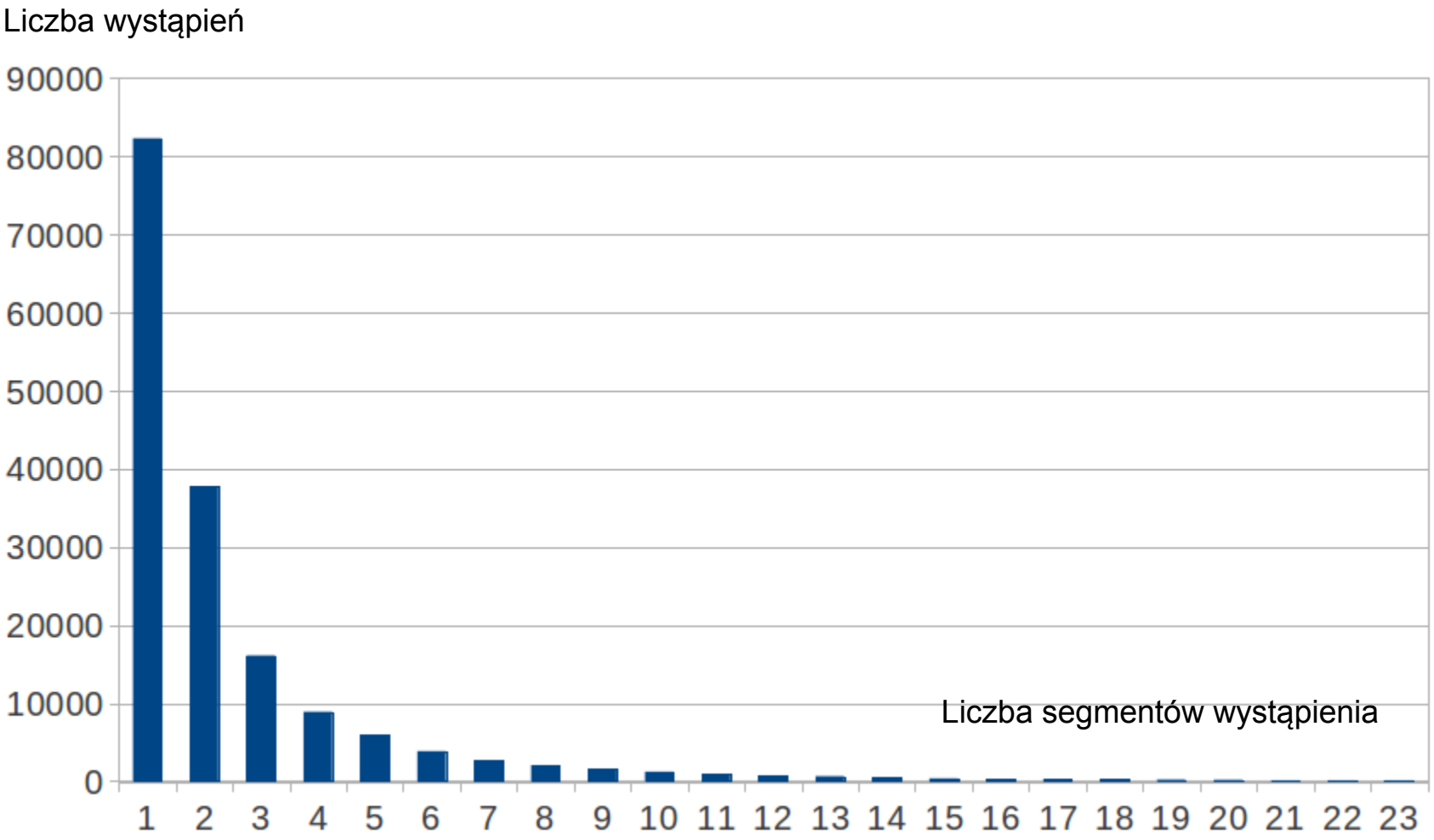
- Schemat anotacji: 1 anotator, 1 superanotator
- Automatyczna preanotacja

- Praca (super)anotatora przy użyciu 2 programów:
  - manager
  - mmax

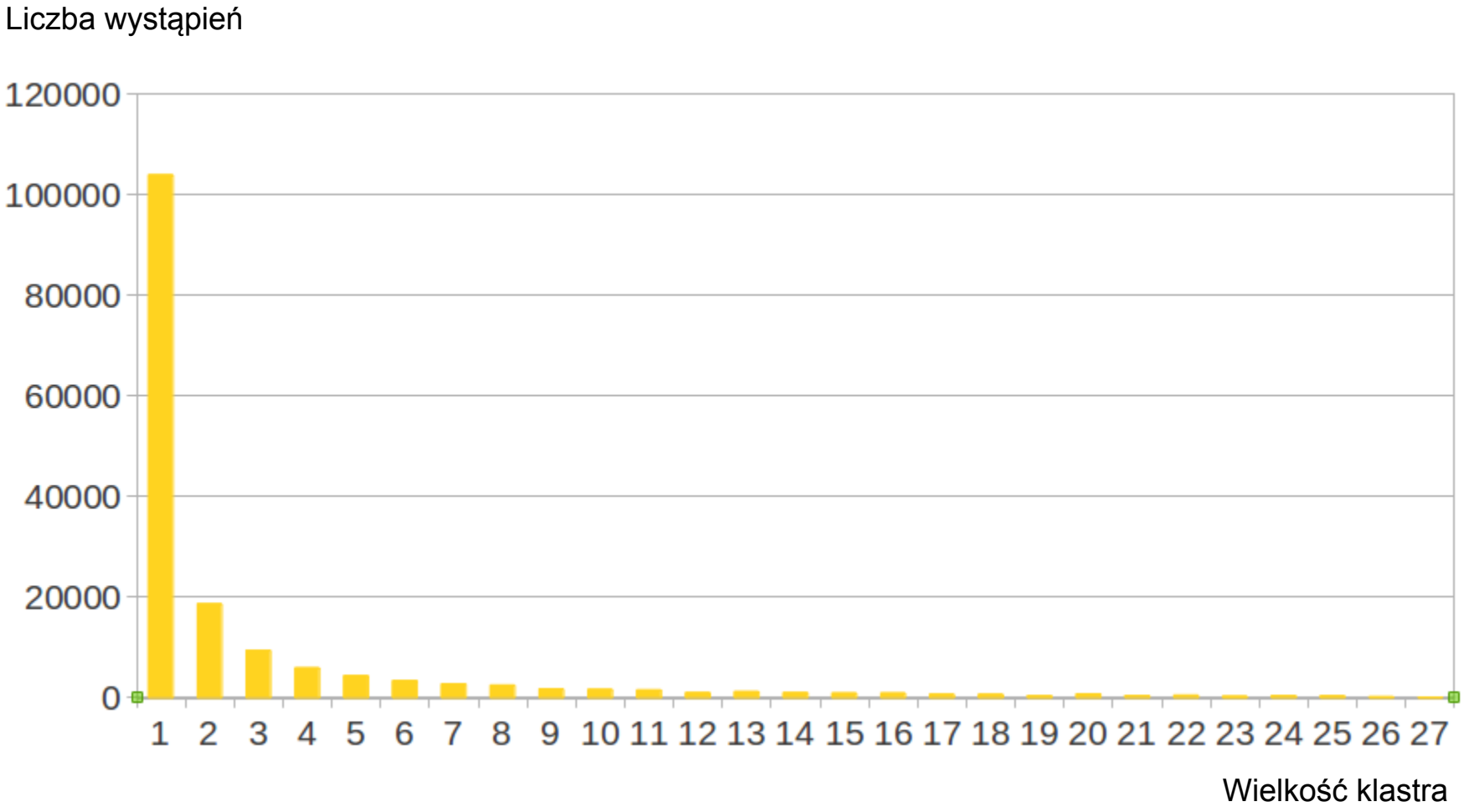
Stan bieżący

	Anotacja	Superanotacja
Teksty	1773	505
Segmenty	503985	144314
Wystąpienia	167765	48608
Klastry	17326	5192
Linki quasi	4288	1354
Stopień ukończenia	100%	28,5%

Wielkość wystąpień



Wielkość klastrów



Koreferencja a specyfika polszczyzny

The Procurator's cheek twitched and he said quietly:
`Bring in the accused.‘
At once two legionaries escorted a man of about twenty-seven from the courtyard, under the arcade and up to the balcony, where they placed him before the Procurator's chair. The man was dressed in a shabby, torn blue chiton. His head was covered with a white bandage fastened round his forehead, his hands tied behind his back. There was a large bruise under the man's left eye and a scab of dried blood in one corner of his mouth. The prisoner stared at the Procurator with anxious curiosity.

The Master and Margarita by Mikhail Bulgakov, Translated from the Russian by Michael Glenny, Printed in Great Britain by Collins Clear-Type Press London and Glasgow © 1967

Koreferencja a specyfika polszczyzny

Procuratorowi skurcz wykrzywił policzek. Powiedział cicho:
– Wprowadźcie oskarżonego.
Natychmiast dwóch legionistów wprowadziło między kolumny z ogrodowego placyku dwudziestosiedmioletniego człowieka i przywiodło go przed tron procuratora. Człowieków odziany był w stary, rozdarty, błękitny chiton. Na głowie miał biały zawój przewiązany wokół rzemykiem, ręce związano mu z tyłu. Pod jego lewym okiem widniał wielki siniak, w kąciku ust miał zdartą skórę i zaschłą krew. Patrzył na procuratora z lękliwą ciekawością.

Michał Bułhakow, *Mistrz i Małgorzata*, tłum. Irena Lewandowska, Witold Dąbrowski, Czytelnik, Warszawa 1988.

Właściwości gramatyczne polszczyzny, które wpływają na koreferencję

1. Swobodny szyk zdania
2. Tzw. *podmiot domyślny* i wyrażanie koreferencji jedynie za pomocą końcówki osobowej czasownika
3. Brak systemu rodzajników

Koreferencja a współklasyfikacja

Hipoteza:
Jeśli N jest nazwą, to dowolne dwa użycia nazwy N mają to samo odniesienie przedmiotowe.

Ten chłopiec *cały czas rozmawia podczas lekcji, a* ***tamten chłopiec*** *jest grzeczny.*

– *Pożycz mi* ***długopis****.*
– *A co się stało z* ***twoim długopisem****?*

Pojęcie *współklasyfikacji* (ang. *co-classification*):
wyrażenia są powiązane ze sobą na mocy faktu, że odnoszą się do różnych referentów, choć należących do tej samej klasy (Kunz 2009:32)

Przykład

*Każdy szanujący się poseł ma **asystenta**. **Asystentami** są z reguły ludzie młodzi, ale nie brakuje również szczerze zaangażowanych emerytów. Poglądy polityczne **asystenta** powinny być zbieżne z linią szefa. Pracują jako wolontariusze tak jak Marek Hajbos, **asystent** Zyty Gilowskiej. Poseł Adam Bielan (rzecznik PiS) na przykład płaci **asystentom** za wysyłanie korespondencji. Obecny minister sprawiedliwości Grzegorz Kurczuk zaczynał partyjną działalność jako **asystent** Izabelli Sierakowskiej. W ministry poszedł też były **asystent** Józefa Oleksego Lech Nikolski. Posłowie nie poprzestają na jednym **asystencie**.*

Koreferencja a współekstensja

Czy dane wyrażenia są koreferencyjne, w sytuacji
gdy jednostki leksykalne łączy jakaś relacja semantyczna?

Pojęcie *współekstensji* (ang. *co-extension*):
co najmniej dwa wyrażenia odnoszą się do czegoś,
co znajduje się w tym samym, ogólnie rozumianym polu
pojęciowym (relacje semantyczne typu mero/holonimia,
klasa/egzemplarz, hipo/hiperonimia, antonimia)
(Kunz 2009:32-34).

Trudno ustalić, czy to identyczność referentów, czy różne
obiekty połączone jakimś typem podobieństwa.

Przykłady 1-2

*[…] **mity** są niezastąpionym narzędziem dla psychologa, usiłującego prześledzić wzorce ludzkich zachowań. Wysiłki archeologów, religioznawców, antropologów doprowadziły z jednej strony do porzucenia eurocentrycznego spojrzenia na **mitologię** i odkrycia ...*

*... – Od czasu **okupacji**... – Ale tu je masz z powrotem, w metryce, i musisz ich używać w urzędowych papierach – powiedział oschle dyrektor i podsunął mi nowy blankiet do wypełnienia. – Kiedy to jest stara metryka, którą mi odtworzono zaraz po **wojnie**.*

Przykład 3

*Teraz **pożar** zwalił się na budy kupieckie na Velia Carinae, pochłonął je jednym łykiem, łapczywie i prędko, po czym uderzył zaraz szeroką ścianą **ognia** na skupisko suburskie. Wielopiętrowe domy stawały w **płomieniach** jedne za drugimi. Z ogarniętych **ogniem** insul ludzie nie mieli czasu uciekać. […] Zresztą i tym, co znaleźli się na dole, w ciasnych, krętych uliczkach pełnych szalejącego **ognia** i duszącego **dymu** niełatwo było uciekać. Krzycząc i nawołując się rozpaczliwie ludzie biegali tam i z powrotem, nie znajdując dla siebie wyjścia z **morza ognia**. Wpół oszaleli, w płonących tunikach, pędzili na oślep przed siebie, wpadali w **płomienie** i ginęli. Podcięte **ogniem** insule poczęły się walić. Osuwały się z łoskotem, zawalając ciasne uliczki. **Snopy iskier** buchały w niebo.*

Koreferencja a pojęcie identyczności

Identyczność wg Wierzbickiej (Wierzbicka 2010: 61)

- tożsamość, którą można wyrazić za pomocą określnika *ten sam*, uznanego przez Wierzbicką za uniwersalną i elementarną jednostkę semantyczną

Identyczność wg Fauconniera i Turnera (za: Libura 2010: 98)

- podstawowa istotna relacja (ang. *vital relation*)
- nie musi wynikać z percepcyjnego podobieństwa między obiektami (np. niemowlę i ta sama osoba jako dorosły człowiek)

Identyczność wg Recasens (Recasens i in. 2011: 1142-1143)

- zjawisko stopniowalne (pełna identyczność, brak identyczności oraz przypadki pośrednie, określane jako prawie identyczność (ang. *near identity*)

Problemy z identycznością w tekście

- *Twój ojciec był największy – powiedziała jakaś nieznana starsza pani, potrząsając ręką Aleksandra – syna **Gassmana**. – Będzie mi go brakować jako **aktora**, ale zwłaszcza będę tęsknił za **ojcem** – powiedział.*
- *Nie widziała **„Przeminęło z wiatrem”**, ale **je** czytała.*
- *Staliśmy, patrząc na dwa portrety **Królowej Elżbiety**. Na tym z lewej była **ona** ubrana jako cesarzowa Indii. Na tym z prawej **królowa** nosiła elegancką, błękitną suknię.*

Problemy z identycznością w tekście

W miejscu dawnej jezdni ryją buldożery. Bez problemu można dojechać ul. Bandurskiego, a że nawierzchnia ***Retkińskiej*** *była znana jako jedna z najbardziej dziurawych w mieście, nikt nawet specjalnie nie skarży się na utrudnienia w ruchu.* ***Nowa Retkińska*** *będzie miała i sygnalizację u zbiegu z ul. Krzemieniecką, i chodnik (spory odcinek* ***ulicy*** *był go całkowicie pozbawiony).*

Różnice w zasobie wiedzy pozajęzykowej między nadawcą i odbiorcą

Wówczas od przywozu samochodu nabytego na szczególnych warunkach nie płaci się ***podatku od towarów i usług*** *[…]*

3) Deklaracje. Deklaracje podatkowe w zakresie dokonywanych nabyć za okresy miesięczne ***podatku VAT****.*

Różnice w zasobie wiedzy pozajęzykowej między nadawcą i odbiorcą

*Winicjuszowi, który miał ochotę rozpytać Ursusa o ojczysty kraj **Ligii**, słowa te sprawiły pewną przyjemność, albowiem rozmowa z człowiekiem wolnym, jakkolwiek prostym, mniejszą przynosiła ujmę jego rzymskiej i patrycjuszowskiej godności niż rozmowa z niewolnikiem, w którym ni prawo, ni obyczaj nie uznawały ludzkiej istoty. – Toś ty nie Aulusów? – spytał. – Nie, panie. Ja służę **Kallinie**, jako służyłem jej matce, ale po dobrej woli.*

Różnice w zasobie wiedzy pozajęzykowej między nadawcą i odbiorcą

Jedynym moim pożywieniem w ostatnich trzech dniach były węglowodany – powiedział ***trzykrotny złoty medalista****. Sportowcy niemieccy są ogromnie zaskoczeni wiadomością o pozytywnym wyniku testu antydopingowego* ***Mühlegga****.*

Podsumowanie

Problemy z identyfikacją wyrażeń koreferencyjnych wynikają z zaburzeń na dwóch płaszczyznach:

1. **Językowej:**
    - Koreferencja jest związana z właściwościami systemu gramatycznego danego języka
    - Inne semantyczne zależności pomiędzy wyrażeniami (ang. *co-classification* oraz *co-extension*) są wyrażane za pomocą tych samych wykładników co koreferencja
2. **Pozajęzykowej: koreferencja wyznacza relację identyczności na poziomie pojęciowym/doświadczeniowym:**
    - Brak jasnej definicji identyczności i duża złożoność tego procesu na poziomie neurologicznym, percepcyjnym i mentalnym.
    - Różnice w zasobie wiedzy nadawcy i odbiorcy.

Problemy procesu anotacji i superanotacji

Układ: dwóch anotatorów + superanotator

1. Za

- Sprawdza poziom trudności zadania.
- Sprawdza, czy zadanie jest wykonywane.

2. Przeciw

- Równoległa anotacja nie eliminuje błędów, ale je mnoży.
- Rozstrzyganie kolizji jest niewystarczające.
- Konieczność równoległego przetwarzania informacji spowalnia proces superanotacji i generuje błędy.

Co dalej?

- Udostępnienie korpusu w repozytorium META-SHARE (projekt CESAR)
- Eksperymenty anotacyjne
- Eksperymenty projekcyjne
- Użycie danych korpusu do tworzenia narzędzi automatycznych

Dziękujemy!

- **Piotr Batko** – anotacja koreferencji,
- **Łukasz Dębowski** – udział w tworzeniu i rozbudowie narzędzi statystycznych,
- **Barbara Dunin-Kęplicz** – nadzór merytoryczny i doradztwo w zadaniach lingwistycznych,
- **Maria Głąbska** – anotacja koreferencji,
- **Katarzyna Głowińska** – ekspertyza lingwistyczna oraz w zakresie organizacji pracy anotacyjnej,
- **Anna Grzeszak** – anotacja koreferencji,
- **Mateusz Kopeć** – główny informatyk, autor środowiska anotacyjnego,
- **Emilia Kubicka** – anotacja koreferencji,
- **Barbara Masny** – anotacja koreferencji,
- **Maciej Ogrodniczuk** – kierownik projektu,
- **Adam Przepiórkowski** – ekspertyza lingwistyczna i informatyczna, współpraca z NKJP,
- **Paulina Rosalska** – anotacja koreferencji,
- **Agata Savary** – ekspertyza dotycząca anotacji korpusu, jednostek nazewniczych i wielowyrazowych oraz narzędzi do anotacji koreferencji, wsparcie w zakresie organizacji pracy anotacyjnej,
- **Magdalena Zawisławska** – ekspertyza lingwistyczna i semantyczna, organizacja pracy anotacyjnej, superanotacja korpusu nawiązań,
- **Sebastian Żurowski** – anotacja koreferencji.