Polski Korpus Koreferencyjny – wersja 0.85

October 1, 2013

Mateusz Kopeć

Instytut Podstaw Informatyki Polskiej Akademii Nauk
m.kopec@ipipan.waw.pl

1 Wprowadzenie

Polski Korpus Koreferencyjny (PKK) jest dużym, ręcznie anotowanym korpusem ogólnie pojmowanej koreferencji w języku polskim. Jego wielkość jest porównywalna do warstwy anotacji anafory w korpusie KPWr [2], jednakże zakres pojmowania koreferencji jest szerszy (np. zakres anotacji nie jest ograniczony do głów semantycznych nazw własnych, lecz dotyczy także innych bytów wraz z dokładnym określeniem ich granic) i metodologia jego budowania zakłada sprawdzanie i poprawianie anotacji. PKK, o wielkości 540,000 tokenów, jest wśród największych korpusów koreferencyjnych na świecie, takich jak Tüba/DZ [3] dla języka niemieckiego, NAIST Text [5] dla japońskiego, OntoNotes 2.0 [16] dla angielskiego, arabskiego i chińskiego, Prague Dependency Treebank [11] dla czeskiego i ANCOR [10] dla francuskiego.

Cel powstania PKK to wsparcie lingwistycznych badań zjawiska koreferencji oraz rozwoju zaawansowanych narzędzi przetwarzania języka polskiego, a w szczególności automatycznych narzędzi wykrywania koreferencji. Szczegółowe instrukcje anotacyjne stworzone na potrzeby anotacji korpusu i prezentujące analizę zjawiska koreferencji mogą być przydatne w kolejnych projektach. Zbadana została możliwość wykorzystania konceptów prawie-identyczności, wyrażeń dominujących i semantycznego podejścia do identyczności odwołania (ang. identity-of-reference), z czego wnioski można wykorzystać do zbudowania podobnego korpusu, szczególnie dla innego języka o bogatej odmianie wyrazów.

Niniejszy dokument opisuje zawartość tego (zrównoważonego) korpusu, proces jego powstawania i dostępność jego danych.

2 Baza tekstowa korpusu

PKK składa się z dwóch podkorpusów:

- 1773 tekstów „krótkich", czyli o długości 250-350 segmentów [1] , będących wybranymi fragmentami dłuższych dokumentów (wybranymi jako pewna liczba pełnych, kolejnych akapitów z pojedynczego źródła).
- 21 tekstów „długich" – pełnych tekstów o długości 1000-4000 segmentów.

Wierzymy, że powyższy skład korpusu pozwoli na badanie wpływu długości polskich tekstów na naturę więzów koreferencyjnych w nich zawartych.

2.1 Teksty krótkie

Teksty krótkie są wybranymi losowo fragmentami losowo wybranych (z zachowaniem zrównoważenia typów) tekstów z Narodowego Korpusu Języka Polskiego – NKJP [17]. Ekstrakcja fragmentów z NKJP przebiegała w następujący sposób – najpierw losowany był tekst z NKJP, a następnie spośród wszystkich możliwych ciągów kolejnych akapitów o długości 250-350 segmentów w tym tekście, jeden taki ciąg był losowo wybierany i wyciągany w celu stworzenia tekstu w PKK. Tekst z NKJP, z którego wyciągnięto już jeden fragment, nie był brany pod uwagę w dalszej części ekstrakcji. Losowanie fragmentów tekstów do ekstrakcji przebiegało tak długo, aż osiągnięta została założona wielkość korpusu. Teksty były pobierane z NKJP w formie czystego tekstu.

Typy tekstów krótkich w PKK odpowiadają typom NKJP, również zrównoważenie korpusu pod kątem tekstów różnych typów odpowiada zrównoważeniu ręcznie anotowanego podkorpusu NKJP (ze względu na rozmiar nazywanego milionowym), opisanemu w [17]. Liczba tekstów i segmentów przypadających na każdy typ jest umieszczona w tabeli 1.

Korpus zawiera łącznie 1773 teksty „krótkie", zawierające 31145 zdań, 503985 segmentów, co daje średnio ok. 284 segmenty/tekst i 18 zdań/tekst. Średnia długość zdania to ok. 16 segmentów.

2.2 Teksty długie

Źródłem tekstów jest tzw. Korpus „Rzeczpospolitej" (KR; [20]) - zbiór artykułów prasowych (w formacie HTML) pobranych z internetowego serwisu gazety o tym samym tytule. Do PKK wybrane zostały pełne artykuły z KR, liczące od 1000 do 4000 segmentów. Wybór został przeprowadzony w sposób losowy (z ręcznym odrzucaniem rozmów lub artykułów będących agregatem kilku krótszych tekstów), dane zostały wydobyte jako czysty tekst. Na podstawie metadanych w plikach HTML (atrybut `DZIAL`) określonych zostało 7 najczęstszych typów tekstów w KR, a następnie do PCC wybrano po 3 teksty każdego z nich. Typy tekstów długich w PKK przedstawione są w tabeli 2.

[1] Segmenty to jednostki o roli podobnej do słów, pozwalające na oddanie morfoskładniowych właściwości w języku polskim (oraz pewnych decyzji podjętych arbitralnie); np. aglutynant (*długo+śmy*), partykuła *-że* (*znasz+że*) itp. są uznawane za osobne segmenty. Szczegółowy opis znajduje się w punkcie 6.2.2 w [17].

Typ tekstów	# tekstów	# segmentów	%
Dzienniki	459	127840	25,36%
Pozostałe periodyki	406	117694	23,35%
Literatura piękna	288	80263	15,92%
Literatura faktu	96	27743	5,50%
Typ informacyjno-poradnikowy	100	27732	5,50%
Mówione konwersacyjne	83	25336	5,02%
Internetowe nieinteraktywne (statyczne strony)	63	17734	3.51%
Internetowe interaktywne (blogi, fora, usenet)	63	17694	3.51%
Inne teksty pisane	55	15190	3,01%
Mówione medialne	44	12806	2,54%
Quasi-mówione (protokoły sesji parlamentu)	43	12783	2,53%
Typ naukowo-dydaktyczny	35	10255	2,03%
Książki publicystyczne	19	5492	1,08%
Książka niebeletrystyczna nieklasyfikowana	19	5423	1,07%
Razem	1773	503985	100,00%

Table 1. Typy krótkich tekstów w PKK

Typ tekstów	# tekstów	# segmentów	%
Publicystyka, opinie	3	7078	19,53%
Prawo	3	5915	16,32%
Ekonomia	3	5843	16,13%
Kraj	3	5172	14,27%
Sport	3	4324	11,93%
Kultura	3	4113	11,35%
Nauka i technika	3	3789	10,46%
Razem	21	36234	100,00%

Table 2. Typy tekstów długich w PKK

Korpus zawiera łącznie 21 tekstów, 2078 zdań, 36234 segmenty, co daje średnio ok. 1725 segmentów/tekst i 99 zdań/tekst. Średnia długość zdania to ok. 17 segmentów.

3 Anotacja korpusu

3.1 Poziomy anotacji

Wyekstrahowane czyste teksty został przetworzone za pomocą narzędzi automatycznych na kilku poziomach: analizy morfoskładniowej przy użyciu Morfeusza [23], tagowania oraz segmentacji na zdania i tokeny za pomocą Pantery [1] oraz preanotacji wystąpień oraz klastrów koreferencyjnych przy pomocy Rulera [13]. Segmentacja została poprawiona ręcznie jedynie w wypadkach, gdy błędy narzędzia automatycznego uniemożliwiały poprawną anotację koreferencji.

3.2 Przebieg anotacji

Preanotowane teksty zostały sprawdzone przez anotatorów. W wypadku błędnej lub brakującej anotacji automatycznej, ich zadaniem było:

- oznaczyć granice wystąpień,
- wybrać głowy semantyczne wystąpień,
- oznaczyć relacje quasi-identyczności,
- pogrupować wystąpienia w relacji identyczności,
- przypisać wyrażenia dominujące grupom wystąpień.

W dalszej kolejności tekst został sprawdzony i poprawiony przez superanotatora.

Ze względu na konieczność przeprowadzenia badań zgodności anotatorów, 210 tekstów „krótkich" zostało przetworzonych w inny sposób: dwóch anotatorów niezależnie anotowało ten sam tekst, a superanotator łączył ich dwie wersje anotacji, tworząc ostateczną wersję tekstu.

Statystyki anotacji znajdują się w Tabeli 3.

Typ tekstów	# wystąpień	# linków prawie-identyczności	# klastrów singletonowych	# klastrów niesingletonowych
krótkie	167765	4288	103938	17326
długie	12442	421	7075	1256
razem	180207	4709	111013	18582

Table 3. Statystyki anotacji

3.3 Wytyczne anotacyjne

W pracy [12] przedstawione są główne aspekty zakresu i wytycznych anotacji w PKK. Zakres ten obejmuje wszystkie grupy nominalne (GN) wraz z zaimkami, gdyż różnica między GN a wystąpieniem jest trudna do wyróżnienia w ogólności. Jeśli chodzi o koreferencję, ograniczamy się do relacji, które nie mogą być wywiedzione bezpośrednio ze składni. Po pierwsze, predykaty nominalne (np. „Helena jest dyrektorką.") nie są nigdy łączone ze swoimi podmiotami (chociaż są wystąpieniami, jak wszystkie GN). Po drugie, inaczej niż w [8] i [11], apozycja jest postrzegana jako pojedyncze wystąpienie, a nie ciąg koreferentnych wystąpień (np. „Oskarżony, mąż ofiary, ojciec trojga dzieci został dowieziony do sądu."). Po trzecie, jak w [4], [11] i [19], oznaczamy nieciągłe GN jako wystąpienia (np. „To był delikatny, że tak powiem, temat."). Wreszcie, podobnie do [14,16,5] i [19], oznaczamy również podmioty zerowe, bardzo częste w języku polskim.

W korpusie występują dwie relacje: identyczność (grupująca wystąpienia w klastry, tj. klasy równoważności) oraz – eksperymentalna – prawie-identyczność, zaproponowana w [18]. Zdefiniowanie prawie-identyczności jest interesujące z tego względu, że pozwala spojrzeć na koreferencję jako na stopniowalną (a nie binarną) relację. Jak zostało to opisane w [12], liczba linków prawie-identycznościowych oraz zgodność anotatorów w tym aspekcie są zbyt niskie, aby uznać tą warstwę za spójnie zaanotowaną. W związku z pionierskim (w kontekście języka polskiego) charakterem naszego projektu, inne relacje niż identyczność i prawie-identyczność nie wchodzą w zakres anotacji: niebezpośrednia anafora (typu „bridging" lub „associative") i „discourse deixis" [4,11,7,15], elipsa (za wyjątkiem podmiotu zerowego), relacje predykatywne i związane [6], „split antecedent" [4], identyczność sensu [5], itp.

Poza anotowanie prawie-identyczności, inne oryginalne aspekty naszego schematu anotacji (zob. [12]) to:

- oznaczanie wyrażenia dominującego, tj. wyrażenia zawierającego najwięcej informacji semantycznych o referencie,
- oznaczanie głowy semantycznej, a nie składniowej.

3.4 Narzędzia anotacyjne

Podczas ręcznej anotacji korpusu wykorzystywane były dwa narzędzia. Pierwsze z nich to aplikacja DistSys, pozwalająca na zarządzanie procesem dystrybucji tekstów wśród anotatorów i superanotatorów, inspirowana procesem anotacji w NKJP [22]. Jest to aplikacja ogólnego użycia, możliwa do wykorzystania w innym zadaniu anotacyjnym, pod warunkiem, że zadanie wymaga rozdzielania plików pochodzących z centralnego serwera wśród anotatorów, którzy lokalnie anotują te pliki (używając innego programu), a po zakończeniu odsyłają z powrotem do centralnego repozytorium.

Drugim narzędziem jest MMAX, silnie zmodyfikowana wersja narzędzia do anotacji MMAX2 [9], która była wykorzystywana do anotacji pojedynczego tekstu (gdy już został pobrany przez anotatora za pomocą DistSys). MMAX2 jest narzędziem ogólnego zastosowania, które w celu uproszczenia i przyspieszenia anotacji zostało istotnie zmienione – wiele niepotrzebnych możliwości zostało usuniętych. Wiele nowych funkcjonalności zostało dodanych, w związku z wymaganiami zgłaszanymi przez anotatorów (na przykład możliwość cofnięcia ostatniej zmiany). Modyfikacje obejmują stworzony plugin do superanotacji, który wyświetla różnice między dwoma wersjami tego samego tekstu i pozwala na ich ujednolicenie. Różnice na są podzielone na poziomy: przykład superanotowania granic wystąpień jest pokazany na Fig. 1. Każdy wiersz prezentuje jedną różnicę między anotatorami A i B: pierwsza kolumna opisuje którego wystąpienia dotyczy różnica, druga prezentuje decyzję anotatora A, a trzecia anotatora B. W pierwszym wierszu widać, że anotator A zaznaczył wystąpienie "to" (znak plusa), podczas gdy drugie tego nie zrobił (znak minusa). Podwójne kliknięcie jednej z opcji powoduje jej wybranie jako poprawnej.

Obydwa narzędzia są dostępne na stronie:
`http://zil.ipipan.waw.pl/PolishCoreferenceTools`.



Fig. 1. Okno superanotacji w MMAX

4 Dostępność korpusu

Polski Korpus Koreferencyjny jest dostępny do pobrania na licencji Creative Commons Uznanie autorstwa 3.0 pod adresem: `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`. Korpus można pobrać w 3 formatach, opisanych poniżej. Dodatkowo istnieje możliwość przeglądania zawartości korpusu online (zob. Fig. 2) zwizualizowanej przy pomocy zmodyfikowanej wersji narzędzia do anotacji Brat (dla czytelności wizualizacji długich łańcuchów koreferencyjnych konieczne były pewne zmiany w programie).

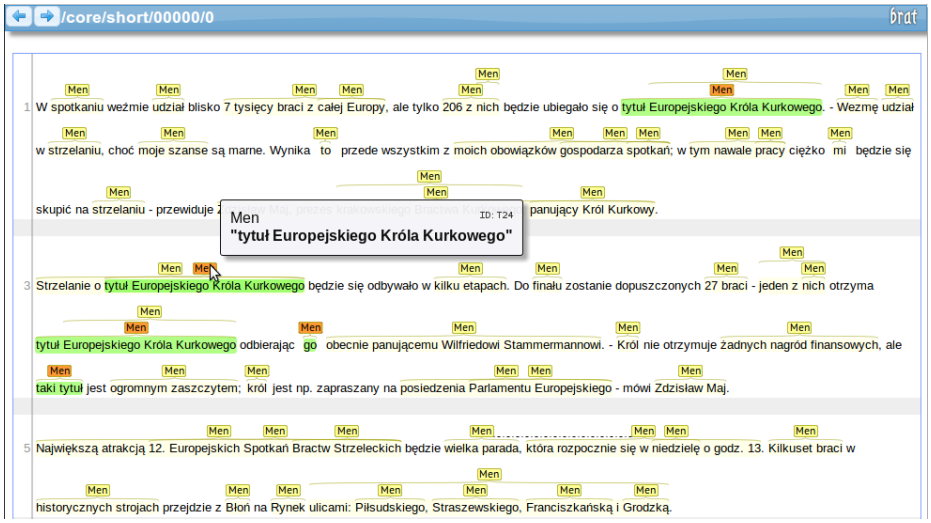


Fig. 2. Wizualizacja korpusu online

4.1 Format Brat

Brat to środowisko online do zbiorowej anotacji [21], które wykorzystuje prosty format danych, opisany pod adresem `http://brat.nlplab.org/standoff.html`. Pojedynczy tekst jest reprezentowany przez dwa pliki:

- plik z rozszerzeniem `.txt` – zawiera czysty tekst,
- plik z rozszerzeniem `.ann` z informacjami na temat wystąpień (oznaczonych jako ciągi znaków z poprzedniego pliku) i relacji między nimi (zarówno identyczności, jak i prawie-identyczności).

Plik z rozszerzeniem `.txt` składa się z 4 rodzajów wierszy, w których kolejne informacje rozdzielone są znakiem tabulacji:

1. wierszy zaczynających się od litery `T`, oznaczających wystąpienia. Przykład wiersza:

   ```
   T24 Mention 370 398 Jeorjos Aleksandros Mangakis
   ```

   Rozdzielone znakiem tabulacji pola tego wiersza to:
   - `T24` to identyfikator wystąpienia.
   - `Mention 370 398` opisuje pozycję wystąpienia – `370` to pozycja pierwszej a `398` ostatniej litery wystąpienia w tekście (czyli pliku `0.txt`). W przypadku nieciągłych wystąpień może się tu pojawić średnik, rozdzielający ciągłe fragmenty wystąpienia (np. `370 398;410 420` oznacza nieciagłe wystąpienie o dwóch ciągłych fragmentach: od pozycji `370` do `398` oraz od `410` do `420`.
   - `Jeorjos Aleksandros Mangakis` to tekstowa forma wystąpienia.

2. wierszy zaczynających się od litery `R`, oznaczających linki prawie-identycznościowe. Przykład wiersza:

   ```
   R4 Quasi Arg1:T337 Arg2:T628
   ```

   Rozdzielone znakiem tabulacji pola tego wiersza to:
   - `R4` to identyfikator linku.
   - `Quasi Arg1:T337 Arg2:T628` opisuje dwa argumenty: źródło i cel linku. `Arg1:T377` to źródło linku – wystąpienie o identyfikatorze `T337`, natomiast `Arg2:T628` to cel linku – wystąpienie o identyfikatorze `T628`.

3. wierszy zaczynających się od znaku `*`, oznaczających niesingletonowe klastry. Przykład wiersza:

   ```
   * Coref T361 T362
   ```

   Rozdzielone znakiem tabulacji pola tego wiersza to:
   - `*` symbolizuje relację równoważności.
   - `Coref T361 T362 T451` opisuje elementy relacji. `T361 T362 T451` to identyfikatory trzech wystąpień, formujących klaster. Liczba tych identyfikatorów może być dowolnie duża, zawsze jest ich co najmniej dwa.

4. wierszy zaczynających się od znaku `#`, opisujących głowy wystąpień oraz ich wyrażenia dominujące (w wypadku wystąpień niesingletonowych). Przykłady:

   ```
   #2 empty T2 Head: "braci"
   #3 empty T2 Dominant: "7 tysięcy braci z całej Europy"
   ```

   Rozdzielone znakiem tabulacji pola tego wiersza to:
   - `#2` or `#3` – identyfikator anotacji,
   - `empty T2` – `T2` to identyfikator wystąpienia, którego głowa lub wyrażenie dominujące jest opisywane.

– `Head: "braci"` lub `Dominant: "7 tysięcy braci z całej Europy"` – głowa lub wyrażenie dominujące, zawarte w cudzysłowie.

4.2 Format MMAX

Format MMAX jest opisany w instrukcji programu MMAX2 (`http://mmax2.net`). W jego zmodyfikowanej w PKK wersji pojedynczy tekst jest reprezentowany przez trzy pliki XML:

– plik z rozszerzeniem `.mmax`, przechowujący źródło tekstu (oryginalny identyfikator z NKJP lub KR) i jego typ,
– plik z końcówką `_words.xml`, zawierający tekst podzielony na słowa oraz ich informacje morfoskładniowe,
– plik z końcówką `_mentions.xml` z informacjami na temat wystąpień (opisanych jako zbiory słów z poprzedniego pliku), wraz z relacjami identyczności i prawie-identyczności pomiędzy nimi.

Plik z rozszerzeniem `.mmax` w tagu `title` atrybut `val` przechowuje informacje na temat tego, skąd wyekstrahowany został fragment tekstu. Przykładowo wartość:

```
Paragraphs: p-57,p-58,p-59,p-60
from NKJP text with id: IPIPAN_1301919980826
```

oznacza, że dany tekst pochodzi z tekstu NKJP o identyfikatorze `IPIPAN_1301919980826`, a dokładniej składa się z jego 4 akapitów o identyfikatorach: `p-57`, `p-58`, `p-59` i `p-60`. Druga istotna informacja to typ tekstu zgodny z wcześniej opisaną typologią – znajduje się on w atrybucie `val` tagu `catRef`.

Plik z końcówką `_words.xml` zawiera warstwę tekstową, wraz z podziałem tekstu na słowa (tagi `word`).

Każde słowo ma tutaj przypisany identyfikator w postaci atrybutu `id`, część mowy w atrybucie `ctag` oraz inne informacje morfoskładniowe w atrybucie `msd`.

Opcjonalne atrybuty słowa to `hasNps`, `lastInSent`, `lastInPar`, które mają wartość `true` gdy odpowiednio: przed słowem nie występuje spacja, słowo jest ostatnie w zdaniu, słowo jest ostatnie w akapicie.

Plik z końcówką `_mentions.xml` zawiera listę wszystkich oznaczonych w tekście wystąpień (tagi `markable`).

Każde wystąpienie w atrybucie `span` ma opisane, które słowa z pliku `0_words.xml` obejmuje. Atrybut `span` to po prostu lista oddzielonych przecinkami identyfikatorów słów (atrybutu `id` w tagu `word`). Dla skrócenia zapisu wprowadzono też dwie kropki (..), które oznaczają, że wystąpienie obejmuje wszystkie słowa

począwszy od tego z identyfikatorem, który pojawia się przed .., aż do słowa o identyfikatorze pojawiającym się po .. (włącznie).

Atrybut `mention_head` zawiera głowę semantyczną wystąpienia w postaci tekstowej.

Atrybut `near_identity` może mieć wartość `empty`, co oznacza, że wystąpienie nie zostało połączone linkiem quasi (wychodzącym), bądź też wartość będącą identyfikatorem `id` wystąpienia, które jest quasi-identyczne z nim (jest celem linku).

Atrybut `mention_group` również może mieć wartość `empty`, co oznacza, że wystąpienie nie jest w żadnym klastrze identycznościowym (jest singletonem), bądź też wartość będącą identyfikatorem klastra (np. `set_5`). Wszystkie wystąpienia w tym pliku, które mają taką samą wartość atrybutu `mention_group`, są w tym samym klastrze identycznościowym.

Atrybut `dominant` zawiera wyrażenie najlepiej opisujące klaster, do którego należy dane wystąpienie, w postaci tekstu. Występuje tu redundancja w przypadku klastrów wieloelementowych, ponieważ każde wystąpienie z osobna posiada wpisaną tą samą wartość atrybutu.

4.3 Format TEI

Format TEI w PKK jest rozszerzeniem formatu TEI z Narodowego Korpusu Języka Polskiego [17]. Pojedynczy tekst jest reprezentowany przez standardowe pliki (umieszczone w oddzielnym katalogu, oznaczonym identyfikatorem tekstu):

- `text_structure.xml`,
- `ann_segmentation.xml`,
- `ann_morphosyntax.xml`,
- `header.xml`,

oraz dodatkowe (w stosunku do NKJP):

- `ann_mentions.xml`,
- `ann_coreference.xml`.

Plik `text_structure.xml` zawiera wyłącznie tekst podzielony na akapity. Plik `ann_segmentation.xml` to podział na zdania i segmenty, `ann_morphosyntax.xml` zawiera informacje o morfoskładni. Plik `header.xml` zawiera tylko informację o typie tekstu (tag `catRef`, atrybut `target`, zgodnie z tabelami 1 i 2) oraz skąd tekst pochodzi, czyli identyfikator oryginalnego tekstu i identyfikatory oryginalnych akapitów (tag `title`).

Plik `ann_mentions.xml` zawiera wystąpienia (reprezentowane przez znaczniki `<seg>`), każde z nich jest zbiorem segmentów z warstwy morfoskładniowej, czyli zawiera wskaźniki na segmenty z pliku `ann_morphosyntax.xml` (podobnie do anotacji nazw własnych w NKJP). Fig. 3 przedstawia wystąpienie *"umiejętność logicznego myślenia"* oznaczone jako 3 wskaźniki do warstwy morfoskładni, z których jeden jest semantyczną głową wystąpienia (co jest zakodowane w tagu `<f name="semh">` zawierającego się w `<fs name="mention">`). Każde Wystąpienie jest poprzedzone komentarzem z jego formą ortograficzną.

Plik `ann_mentions.xml` odwzorowuje również strukturę tekstu, wystąpienia są pogrupowane w zdania i akapity poprzez tagi `<s>` i `<p>`.

Plik `ann_coreference.xml` zawiera informacje o niesingletonowych klastrach identycznościowych oraz linkach prawie-identycznościowych, zakodowanych jako grupy wskaźników do warstwy wystąpień w pliku `ann_mentions.xml`. Każda grupa jest poprzedzona komentarzem zawierającym ortograficzne formy wystąpień będących w relacji. Fig. 4 pokazuje kodowanie dwóch relacji: klastra identycznościowego (zawierającego wystąpienia `mention_8` i `mention_14`, o typie `ident`) oraz relację prawie-identyczności (między wystąpieniami `mention_30` i `mention_5`, o typie `near-ident`). W wypadku indentyczności, kodowanie obejmuje wyrażenia dominujące klastra (tag `<f>` o wartości `"dominant"` atrybut`name`). W wypadku prawie-identyczności, grupa ma zawsze dwa wystąpienia składowe, a jedno z nich posiada przy wskaźniku informację, że jest źródłem relacji (atrybut `type` tagu `<ptr>`, wartość `source`).

References

1. Szymon Acedański. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer, 2010.
2. Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*, Istanbul, Turkey, 2012. ELRA.
3. Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. Recent developments in linguistic annotations of the tüba-d/z treebank, 2004.
4. Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. A unified representation for morphological, syntactic, semantic, and referential annotations. In *IN ACL WORKSHOP ON FRONTIERS IN CORPUS ANNOTATION II: PIE IN THE SKY, ANN ARBOR*, 2005.
5. Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 132–139, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
 xmlns:nkjp="http://www.nkjp.pl/ns/1.0">
<TEI>
  <teiHeader/>
  <text>
    <body>
      <p xml:id="mentions_p-1" corresp="morph_1-p">
        <s xml:id="mentions_p-1.1-s" corresp="morph_1.1-s">
          ...
          <!-- umiejętność logicznego myślenia -->
          <seg xml:id="mention_8">
            <fs type="mention">
              <f name="semh"
                 fVal="ann_morphosyntax.xml#morph_1.1.23-seg"/>
            </fs>
            <ptr target="ann_morphosyntax.xml#morph_1.1.23-seg"/>
            <ptr target="ann_morphosyntax.xml#morph_1.1.24-seg"/>
            <ptr target="ann_morphosyntax.xml#morph_1.1.25-seg"/>
          </seg>
          ...
          <!-- Europejskiego Króla Kurkowego  -->
          <seg xml:id="mention_9">
            <ptr type="nonhead" target="morph_1.1.46-seg"/>
            <ptr type="head" target="morph_1.1.47-seg"/>
            <ptr type="nonhead" target="morph_1.1.48-seg"/>
          </seg>
          ...
        </s>
        <s xml:id="mentions_p-1.2-s" corresp="morph_1.2-s">
          ...
          <!-- to -->
          <seg xml:id="mention_11">
            <ptr type="head" target="morph_1.2.2-seg"/>
          </seg>
          ...
        </s>
      </p>
      <p xml:id="mentions_p-2" corresp="morph_2-p">
      ...
      </p>
      ...
    </body>
  </text>
</TEI>
</teiCorpus>
```

Fig. 3. Kodowanie wystąpień w `ann_mentions.xml`

```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:xi="http://www.w3.org/2001/XInclude"
  xmlns:nkjp="http://www.nkjp.pl/ns/1.0">
  <xi:include href="NKJP_1M_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <p>
          <!-- umiejętność logicznego myślenia; umiejętności te -->
          <seg xml:id="coreference_0">
            <fs type="coreference">
              <f name="type" fVal="ident"/>
              <f name="dominant"
                 fVal="umiejętność logicznego myślenia"/>
            </fs>
            <ptr target="ann_mentions.xml#mention_8"/>
            <ptr target="ann_mentions.xml#mention_14"/>
          </seg>
          ...
          <!-- filharmonia; nowa filharmonia -->
          <seg xml:id="coreference_2">
            <fs type="coreference">
              <f name="type" fVal="near-ident"/>
            </fs>
            <ptr target="ann_mentions.xml#mention_5"/>
            <ptr type="source" target="ann_mentions.xml#mention_30"/>
          </seg>
          ...
        </p>
      </body>
    </text>
  </TEI>
</teiCorpus>
```

Fig. 4. Identyczność i prawie-identyczność zakodowana w `ann_coreference.xml`

6. Frederik Coppens Walter Daelemans Veronique Hoste Geert Kloosterman Anne-Marie Mineur Joeri Van Der Vloet Iris Hendrickx, Gosse Bouma and Jean-Luc Verschelde. A coreference corpus and resolution system for dutch. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.
7. I. Korzen and Buch-Kromann M. Anaphoric relations in the copenhagen dependency treebanks. In *Proceedings of DGFS workshop*, pages 83–98, 2011.
8. Linguistic-Data-Consortium.
9. Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.
10. Judith Muzerelle, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol. ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 555–563, Les Sables d'Olonne, France, 2013.
11. A. Nedoluzhko, Mírovský J., Ocelák R., and Pergler J. Extended coreferential relations and bridging anaphora in the prague dependency treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, page 1–16.
12. Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Maosong Sun, Min Zhang, Dekang Lin, and Haifeng Wang, editors, *12th China National Conference on Computational Linguistics (12th CCL) and the 1st International Symposium on Natural Language Processing based on Naturally Annotated Big Data (1st NLP-NABD)*, volume 8202 of *Lecture Notes in Computer Science*, pages 97–108. Springer-Verlag, Berlin, Heidelberg, 2013.
13. Maciej Ogrodniczuk and Mateusz Kopeć. End-to-end coreference resolution baseline system for Polish. In Zygmunt Vetulani, editor, *Proceedings of the Fifth Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 167–171, Poznań, Poland, 2011.
14. P. Osenova and Simov K. Btb-tr05: Bultreebank stylebook. bultreebank version 1.0., 2004.
15. Massimo Poesio and Ron Artstein. Anaphoric annotation in the arrau corpus. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.
16. Sameer S. Pradhan, Lance Ramshaw, Ralph M. Weischedel, Jessica MacBride, and Linnea Micciulla. Unrestricted coreference: Identifying entities and events in ontonotes. In *ICSC*, pages 446–453. IEEE Computer Society, 2007.
17. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw, 2012.

18. Marta Recasens, Eduard Hovy, and M. Antònia Martí. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138 – 1152, 2011.
19. Marta Recasens and M.Antònia Martí. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44(4):315–345, 2010.
20. Korpus rzeczpospolitej. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.
21. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics.
22. Jakub Waszczuk, Katarzyna Glowinska, Agata Savary, Adam Przepiórkowski, and Michal Lenart. Annotation tools for syntax and named entities in the national corpus of polish. *IJDMMM*, 5(2):103–122, 2013.
23. Marcin Woliński. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin, 2006.