Maciej Ogrodniczuk

Automatyczne wykrywanie nominalnych zależności referencyjnych w polskich tekstach współczesnych

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

Automatyczne wykrywanie nominalnych zależności referencyjnych w polskich tekstach współczesnych

Maciej Ogrodniczuk

Automatyczne wykrywanie nominalnych zależności referencyjnych w polskich tekstach współczesnych

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

Recenzenci:
prof. dr hab. Włodzimierz Gruszczyński, prof. dr hab. Adam Pawłowski

Redaktor prowadzący:
Karolina Kozakowska

Korekta:
Monika Szewczyk, Magdalena Zawisławska

Projekt okładki i stron tytułowych:
Anna Gogolewska

Ilustracja na okładce:
Nongkran_ch/iStock

Skład i łamanie w systemie LaTeX:
Maciej Ogrodniczuk

Publikacja finansowana przez Instytut Podstaw Informatyki PAN.

ISBN 978-83-235-3622-2 (druk) ISBN 978-83-235-3630-7 (PDF online)

Praca powstała w wyniku realizacji projektu badawczego 2014/15/B/HS2/03435 finansowanego ze środków Narodowego Centrum Nauki.

Wydawnictwa Uniwersytetu Warszawskiego
00-497 Warszawa, ul. Nowy Świat 4
e-mail: wuw@uw.edu.pl
księgarnia internetowa: www.wuw.pl

Wydanie 1, Warszawa 2019

Spis treści

Przedmowa **13**

Informacja o finansowaniu prac **15**

Podziękowania **17**

1. Założenia badawcze **19**
1.1. Referencja, koreferencja, anafora, asocjacja 19
1.2. Motywacja 21
1.3. Cele badawcze 22
1.4. Zakres badań 23
1.5. Metodologia 24

2. Od ujęć teoretycznych do dekodowania relacji referencyjnych **27**
2.1. Pojęcie i zakres referencji 27
2.2. Klasyfikacje typów wzmianek i relacji referencyjnych 29
2.2.1. Klemensiewicz 30
2.2.2. Topolińska 31
2.2.3. Paduczewa 32
2.2.4. Clark i inne klasyfikacje zagraniczne 33
2.3. Cechy relacji referencyjnych 35
2.4. Projekty korpusowe 38
2.5. Komputerowe implementacje modelu referencji 41
2.6. Metody ewaluacji 48
2.6.1. Miara MUC 50
2.6.2. Miara B^3 51
2.6.3. Miara CEAF 52
2.6.4. Miara BLANC 53

3. Model relacji referencyjnych **55**
3.1. Świat tekstu i własność referencji 55
3.2. Typy i granice wzmianek 56

3.3. Relacje tekstowe i pozatekstowe 58
3.4. Typologia referencyjna 59
3.4.1. Koreferencja 61
3.4.2. Referencja pośrednia 61
3.4.3. Relacje wspierające 63
3.4.4. Relacje wykluczające 63
3.4.5. Aspekty 64

4. Korpus zależności referencyjnych 67
4.1. Wybór tekstów 67
4.2. Wybór strategii anotacyjnej 69
4.2.1. Liczba i profil anotatorów 69
4.2.2. Anotacja szeregowa a anotacja równoległa 71
4.2.3. Preanotacja 72
4.2.4. Superanotacja automatyczna 73
4.3. Prace anotacyjne 75
4.3.1. Faza rozpoznawcza 75
4.3.2. Anotacja koreferencji nominalnej 76
4.3.3. Anotacja ogólnych zależności referencyjnych 78
4.4. Narzędzia anotacyjne 79
4.5. Zgodność anotatorów 83
4.5.1. Wzmianki 83
4.5.2. Klastry koreferencyjne 84
4.5.3. Pozostałe relacje 85
4.6. Korekta błędów 87
4.7. Udostępnienie korpusu 88
4.7.1. Format SemEval/CoNLL 89
4.7.2. Format MMAX 91
4.7.3. Format TEI 94
4.7.4. Format narzędzia BRAT i wersja online korpusu 96
4.7.5. Wyszukiwarka korpusowa 101
4.8. Statystyki korpusowe 101
4.8.1. Własności tekstów 101
4.8.2. Własności wzmianek 103
4.8.3. Statystyka relacji referencyjnych 107

5. Implementacja 113
5.1. Wykrywanie wzmianek . . . 113
5.1.1. System regułowy . . . 114
5.1.2. System statystyczny . . . 115
5.2. Wykrywanie koreferencji . . . 117
5.2.1. System regułowy . . . 117
5.2.2. System statystyczny . . . 118
5.2.3. System sitowy . . . 120
5.2.4. System neuronowy . . . 122
5.2.5. System hybrydowy . . . 125
5.3. Dekodowanie relacji pośrednich i pomocniczych . . . 125

6. Ewaluacja szczegółowa 127
6.1. Wykrywanie wzmianek . . . 127
6.2. Wykrywanie koreferencji . . . 128
6.2.1. Wzmianki idealne . . . 128
6.2.2. Wzmianki systemowe . . . 129
6.3. Wykrywanie wybranych zależności pośrednich i pomocniczych . . . 130
6.4. Analiza błędów . . . 130
6.4.1. Błędy wykrywania wzmianek . . . 133
6.4.2. Błędy wykrywania koreferencji . . . 135
6.4.3. Analiza relacji pośrednich . . . 136

7. Perspektywy badań 139
7.1. W stronę koreferencji uniwersalnej . . . 139
7.2. Model Penn Discourse Treebank . . . 142
7.3. Anotacja metatekstowa . . . 146

Podsumowanie 149

English summary 153

Bibliografia 161

Skorowidz 187

Skorowidz terminów angielskich 189

Wykaz powstałych narzędzi i zasobów 191

Table of contents

Preface **13**

Funding information **15**

Acknowledgements **17**

1. The point of departure **19**
1.1. Reference, coreference, anaphora, association 19
1.2. Motivation 21
1.3. Research objectives 22
1.4. Scope of work 23
1.5. Methodology 24

2. From theoretical perspective to decoding of referential relations **27**
2.1. The concept and scope of reference 27
2.2. Classifications of mention types and referential relations 29
2.2.1. Klemensiewicz 30
2.2.2. Topolińska 31
2.2.3. Paduczewa 32
2.2.4. Clark and other foreign classifications 33
2.3. Features of referential relations 35
2.4. Corpus projects and automated resolution 38
2.5. Computer-based implementations of reference 41
2.6. Evaluation methods 48
2.6.1. MUC metric 50
2.6.2. B^3 metric 51
2.6.3. CEAF metric 52
2.6.4. BLANC metric 53

3. Typology of referential relations **55**
3.1. Discourse world and referential properties 55
3.2. Mention types and borders 56

3.3. Textual relations vs. out-of-text reference 58
3.4. Referential typology 59
3.4.1. Coreference 61
3.4.2. Indirect reference 61
3.4.3. Supporting relations 63
3.4.4. Excluding relations 63
3.4.5. Facets 64

4. Corpus of referential relations **67**
4.1. Text selection 67
4.2. Annotation strategy 69
4.2.1. Number and profile of annotators 69
4.2.2. Serial vs. parallel annotation 71
4.2.3. Pre-annotation 72
4.2.4. Automated adjudication 73
4.3. Annotation phases 75
4.3.1. Preparatory phase 75
4.3.2. Annotation of nominal coreference 76
4.3.3. Annotation of referential relations 78
4.4. Annotation tools 79
4.5. Annotator agreement 83
4.5.1. Mentions 83
4.5.2. Coreference clusters 84
4.5.3. Other relations 85
4.6. Error correction 87
4.7. Corpus availability 88
4.7.1. SemEval/CoNLL format 89
4.7.2. MMAX format 91
4.7.3. TEI format 94
4.7.4. BRAT format and online corpus version 96
4.7.5. Corpus search engine 101
4.8. Corpus statistics 101
4.8.1. Textual properties 101
4.8.2. Mention statistics 103
4.8.3. Coreference clusters 107

5. Implementation **113**
5.1. Mention detection . . . 113
5.1.1. Rule-based mention detection . . . 114
5.1.2. Statistical mention detection . . . 115
5.2. Coreference resolution . . . 117
5.2.1. Rule-based coreference resolution . . . 117
5.2.2. Statistical coreference resolution . . . 118
5.2.3. Sieve-based coreference resolution . . . 120
5.2.4. Deep network-based coreference resolution . . . 122
5.2.5. Hybrid system . . . 125
5.3. Decoding associative and auxiliary relations . . . 125

6. Evaluation **127**
6.1. Mention detection . . . 127
6.2. Coreference resolution . . . 128
6.2.1. Gold mentions . . . 128
6.2.2. System mentions . . . 129
6.3. Detection of selected indirect relations . . . 130
6.4. Error analysis . . . 130
6.4.1. Mention detection errors . . . 133
6.4.2. Coreference resolution errors . . . 135
6.4.3. Analysis of bridging relations . . . 136

7. Research perspectives **139**
7.1. Towards Universal Coreference . . . 139
7.2. Penn Discourse Treebank model . . . 142
7.3. Discourse-based annotation . . . 146

Conclusions **149**

English summary **153**

Bibliography **161**

Glossary **187**

Glossary of English terms **189**

Implemented tools and resources **191**

Przedmowa

Niniejsza książka jest wynikiem interdyscyplinarnych (lingwistyczno-informatycznych) badań nad automatycznym dekodowaniem relacji referencyjnych w tekstach polskich. Głównych celem tych badań było stworzenie komputerowego modelu zależności tego rodzaju oraz implementacja wykrywających je narzędzi. Opisywane prace były prowadzone pod moim kierownictwem w Zespole Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN od 2011 r. i finansowane ze środków Ministerstwa Nauki i Szkolnictwa Wyższego oraz Narodowego Centrum Nauki w ramach dwóch grantów badawczych.

Już w momencie wnioskowania o pierwszy projekt wielu kolegów, także z zagranicy, przekonywało mnie, że temat komputerowego dekodowania referencji nie jest już popularny w światowej nauce, w szczególności ze względu na spore trudności w przekroczeniu progu 70–80% miary F_1 (w zależności od języka), co w opinii niektórych możliwe byłoby tylko przy uwzględnieniu tzw. wiedzy ogólnej, wciąż trudno kodyfikowalnej w systemach komputerowych. Dodatkowy problem stanowił zamiar koncentracji prac na języku polskim, niszowym z globalnej perspektywy naukowej. Wątpliwości te potwierdziła zresztą nieudana próba nakłonienia badaczy z innych krajów do udziału w zadaniu wykrywania referencji dla polszczyzny na dostarczonych danych postawionym uczestnikom współorganizowanego przeze mnie warsztatu CORBON (*Coreference Resolution Beyond OntoNotes*) w 2016 r. Mimo wielu sygnałów wstępnego zainteresowania tematem, bariera językowa okazała się zbyt wysoka lub wyniki uzyskiwane standardowymi metodami zbyt słabe, by je zaprezentować.

Dekoder zależności referencyjnych stanowił jednak ważny element, którego brakowało w zestawie podstawowych narzędzi językowych powstałych w ostatnich latach dla polszczyzny. Mogłyby z niego w oczywisty sposób skorzystać algorytmy automatycznego streszczania (np. w celu zastępowania wyrażeń niepełnoznacznych), tłumaczenia komputerowego (do ujednoznaczniania wariantów tłumaczeń) czy analizy metatekstowej. Jednocześnie w ciągu ostatnich lat nastąpił intensywny rozwój nowych, efektywnych metod komputerowych, a zaspokojenie „pierwszych potrzeb" w dziedzinie polskiej inżynierii lingwistycznej umożliwiło skoncentrowa-

nie prac na bardziej wymagających problemach z pogranicza składni i semantyki oraz referencji oraz dyskursu (metatekstu).

W związku z tym, że w języku polskim zagadnienie przetwarzania relacji referencyjnych w ujęciu ogólnym nie było dotąd systematycznie badane metodami lingwistyczno-informatycznymi, praca ta stanowi pierwszą skondensowaną próbę komputerowego opisu referencji nominalnej w języku polskim oraz przedstawienie sposobu implementacji narzędzi do jej wykrywania. Zgodnie z aktualnymi trendami wykorzystuję do tego celu podejście korpusowe, z ręczną anotacją konstrukcji referencyjnych, pozwalające zarówno na weryfikację zaproponowanej teorii na rzeczywistych danych, jak i tworzenie narzędzi automatycznych metodami maszynowego uczenia, a następnie ocenę jakości powstałych narzędzi za pomocą standardowych miar ewaluacyjnych.

Książka podzielona jest na części odpowiadające głównym blokom tematycznym pracy korpusowo-informatycznej. Po przedstawieniu założeń (rozdział 1) oraz stanu obecnej wiedzy teoretycznej i praktycznej w zakresie, w jakim była przydatna w pracach algorytmicznych (rozdział 2), prezentuję stworzony na ich potrzeby model relacji referencyjnych (rozdział 3), użyty następnie w procesie anotacyjnym o szczegółowo określonych ramach, który doprowadził do powstania korpusu zależności referencyjnych (rozdział 4). Dane korpusu posłużyły następnie do stworzenia kilku wariantów narzędzi do automatycznego wykrywania referencji (rozdział 5), a ich jakość została oceniona zgodnie z dostępnymi metrykami (rozdział 6). Perspektywa dalszych badań (rozdział 7) została zaprezentowana w szerszym kontekście modelowania relacji metatekstowych. Ostatni rozdział stanowi krótkie podsumowanie uzyskanych wyników.

Obecna publikacja prezentuje czytelnikowi polskiemu prace prowadzone w trakcie ośmiu lat, co wiąże się z dwiema konsekwencjami. Pierwszą z nich jest konieczność podsumowania wyników opisywanych już częściowo wcześniej, w monografii anglojęzycznej (Ogrodniczuk 2015) oraz licznych artykułach i publikacjach konferencyjnych. Drugą – potrzeba skondensowanego przedstawienia obszernego materiału. W celu ułatwienia lektury wszystkie fragmenty, mogące wymagać dokładniejszych objaśnień, zostały zaopatrzone w odesłania do wcześniejszych prac. Na końcu książki zamieszczono jej angielskie streszczenie przeznaczone dla czytelników zagranicznych.

Informacja o finansowaniu prac

Prace nad analizą relacji koreferencyjnych w polszczyźnie były prowadzone w projekcie badawczym „Komputerowe metody identyfikacji nawiązań w tekstach polskich” (CORE) finansowanym przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach 40. konkursu na granty na badania własne (dyscyplina naukowa N519 – Metody Komputerowe w Nauce; nr kontraktu: 6505/B/T02/2011/40; kwiecień 2011 – lipiec 2014).

Rozszerzone badania nad uogólnionymi relacjami referencyjnymi z komponentem nominalnym były prowadzone w projekcie badawczym „Ujednolicona teoria koreferencji w języku polskim i jej korpusowa weryfikacja” (COTHEC) finansowanym przez Narodowe Centrum Nauki w konkursie OPUS 8 (obszar badawczy: HS – Nauki Humanistyczne, Społeczne i o Sztuce; panel HS2 – Kultura i twórczość kulturowa; nr kontraktu: 2014/15/B/HS2/03435; luty 2015 – lipiec 2018).

Badania koreferencji w dyskursie zostały rozpoczęte w projekcie „Structuring Discourse in Multilingual Europe” (TextLink) finansowanym przez Komisję Europejską w ramach akcji COST IS1312 (moduł Individuals, Societies, Cultures and Health; kwiecień 2014 – kwiecień 2018), w szczególności podczas krótkiej misji naukowej autora (Short Term Scientific Mission) w School of Informatics na uniwersytecie w Edynburgu (luty–kwiecień 2016).

Anotacja relacji metatekstowych na materiale korpusu zależności referencyjnych została rozpoczęta w projekcie CLARIN-PL dotyczącym konstrukcji infrastruktury badawczej i realizowanym w ramach wspólnego międzynarodowego przedsięwzięcia pn. CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure i finansowanego w postaci kosztów wkładu krajowego na mocy decyzji MNiSW nr DIR/WK/2016/02 (lipiec 2016 – czerwiec 2018).

Podziękowania

Dziękuję obu zespołom projektowym, które wzięły udział w realizacji opisanych w tej książce prac, w szczególności:

– lingwistkom – Katarzynie Głowińskiej, Agacie Savary, Alicji Wójcickiej, Magdalenie Zawisławskiej;

– informatykom – Zbigniewowi Gawłowiczowi, Mateuszowi Kopciowi, Pawłowi Morawieckiemu i Bartłomiejowi Nitoniowi;

– anotatorom – Bartłomiejowi Alberskiemu, Annie Andrzejczuk, Marii Głąbskiej, Annie Grzeszak, Agnieszce Kostrowieckiej, Emilii Kubickiej, Dawidowi Lipińskiemu, Barbarze Milanowskiej, Ewelinie Pędzich, Barbarze Pukalskiej, Paulinie Rosalskiej, Adrianowi Sulichowi, Michałowi Szczyszkowi, Danielowi Ziembickiemu i Sebastianowi Żurowskiemu;

– redaktorom, korektorom i tłumaczom – Filipowi Skwarskiemu, Monice Szewczyk, Joannie Wieruckiej i Justynie Żurkowskiej–Paciorek;

– ekspertom służącym wiedzą i pomocą na różnych etapach prac – Barbarze Dunin-Kęplicz, Piotrowi Batce, Łukaszowi Degórskiemu, Łukaszowi Dębowskiemu, Łukaszowi Kobylińskiemu, Michałowi Lenartowi, Małgorzacie Marciniak, Agnieszce Mykowieckiej, Adamowi Przepiórkowskiemu, Jakubowi Waszczukowi, Marcinowi Wolińskiemu, Alinie Wróblewskiej;

– pozostałym członkom Zespołu Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN tworzącym życzliwą atmosferę pracy w jednej z najlepszych grup lingwistyczno-komputerowych w Polsce.

Dziękuję Rodzicom, Żonie i Synom, bez których wszystko wyglądałoby inaczej.

1.

Założenia badawcze

1.1. Referencja, koreferencja, anafora, asocjacja

Tworząc i analizując wypowiedzi, stale odnosimy się do rzeczy, które znamy. Zjawisko to nazywamy **referencją** (ang. *reference*), czyli aktem odwołania się do rzeczywistości pozajęzykowej za pomocą środków językowych użytych w wypowiedzi. Obiekty, które przywołujemy, nie muszą oczywiście pochodzić ze świata rzeczywistego – wystarczy, by należały do mentalnego **świata tekstu** (świata dyskursu, ang. *discourse world*) stworzonego na potrzeby komunikacji językowej. Na podobnej zasadzie odwołujemy się do stanów, zdarzeń, czynności, miejsc, czasu i innych zjawisk pozatekstowych (w dalszej części wywodu używam określenia „obiekt" dla wszystkich typów bytów mogących podlegać referencji).

Wyrażenia referencyjne, za pomocą których tworzymy odwołania w tekście, nazywam **wzmiankami** (ang. *mentions*). W skład wzmianki wchodzą, oprócz jej **centrum semantycznego** (ang. *semantic head*; rozdział 4.3.2), także jego wszystkie podrzędniki, zgodnie z założeniem o konieczności zapewnienia wzmiance semantycznej precyzji (np. wyrażenie *samochód, który potrącił moją żonę* jest znaczeniowo pełniejsze niż samo jego centrum *samochód*). Zasadniczo odniesienia do obiektów realizowane są jako uogólnione konstrukcje nominalne, ale czasem wzmianką może być także dłuższy fragment tekstu, np. opisujący pewną sytuację.

Wzmianki odpowiadające obiektom przywołanym w tekście tylko raz nazywam **singletonami** (ang. *singleton*). Kiedy odwołanie następuje wielokrotnie, pomiędzy fragmentami wypowiedzi o wspólnym odniesieniu zachodzi zjawisko **koreferencji** (ang. *coreference*); zbiór takich odwołań nazywam **klastrem koreferencyjnym** (ang. *coreference cluster*). W literaturze funkcjonuje także nazwa *łańcuch koreferencyjny* (ang. *coreference chain*), moim zdaniem błędnie sugerująca sekwencyjność wzmianek, która nie zawsze zachodzi; np. w sytuacji realizacji odwołania za pomocą powtórzenia nazwy, do interpretacji następnika nie jest wymagane odwołanie do poprzednika.

Ze względów stylistycznych kolejne odwołania są zwykle realizowane za pomocą innych środków językowych niż proste powtórzenie – jeśli odnosimy się do wcześniej wymienionego obiektu, np. często przylatującej do ogrodowego karmnika charakterystycznej sikorki, możemy użyć wyrażenia bliskoznacznego z użytym wcześniej (*sikora*, *bogatka*), hiperonimu (*ptak*), zaimka (*ona*), neologizmu (*słoninkożerca*), nazwy własnej (*Krzywodziobek*), czy nawet wyrażenia idiolektalnego zrozumiałego tylko dla domowników (*ten nasz wróbel*). Koreferencja jest więc zjawiskiem posługującym się środkami znacznie wykraczającymi poza czystą składnię i semantykę, zachodzącym na poziomie całościowego rozumienia struktury tekstu (ang. *discourse*) i łączącym świat językowy z pozajęzykowym. Z tego powodu problem **dekodowania koreferencji** (ang. *coreference resolution*) jest uznawany za jeden z najtrudniejszych w przetwarzaniu języka naturalnego.

Interpretacja niektórych rodzajów wzmianek (np. zaimkowych) jest niemożliwa bez posłużenia się innym fragmentem tekstu i wówczas między powiązanymi fragmentami zachodzi wewnątrztekstowa relacja **anafory** (ang. *anaphora*) lub **katafory** (ang. *cataphora*), odpowiadająca odniesieniu do elementu pełnoznacznego następującego liniowo przed elementem niepełnoznacznym lub po nim. Posturzyńska-Bosko (2015) za Maillardem (1974) zjawiska te określa łącznie terminem **diafory** (ang. *diaphora*); termin ten nie jest jednak powszechnie stosowany, zatem dla uproszczenia używam dalej określenia „anafora" w znaczeniu diafory, sygnalizując rozróżnienie szczegółowe w razie potrzeby. Anafora jest zatem relacją wykorzystującą zestaw cech konotowanych przez powiązane wzmianki (niezależnie od ich denotacji), podczas gdy koreferencja zakłada zgodność denotacji (por. Topolińska 1977). Warto zwrócić uwagę, że referencja jako zjawisko na pograniczu tekstu i rzeczywistości pozajęzykowej jest jednak ogólniejsza i mentalnie wcześniejsza od anafory: autor wypowiedzi najpierw podejmuje decyzję o odwołaniu się danego obiektu, a następnie o użyciu środków językowych, za pomocą których zostanie ono zrealizowane, z uwzględnieniem uwarunkowań stylistycznych.

Biorąc pod uwagę odwołania pozatekstowe, oprócz **bezpośrednich** (ang. *direct reference*), w przypadku których wzmianka odnosi się jawnie do opisywanego obiektu, w tekście mogą wystąpić **odwołania pośrednie** (ang. *indirect reference*), nazywane też często **asocjacyjnymi** (ang. *associative anaphora*, *bridging*) czy rzadziej – **interreferencją** (ang. *interreference*, patrz Janssen 1980). Wzmianka odnosi się wówczas do danego obiektu za pośrednictwem innego, pozostającego z nim w określonej zależności (np. odwołanie bezpośrednie do schodów jest też odwołaniem pośrednim do konkretnego domu, w którym te schody się znajdują, a nie do jakiegoś innego domu).

W tekście mogą się też znajdować dodatkowe określenia wzmianki, które rozszerzają zakres odnoszących się do niej nazw. Mogą mieć one postać na przykład rzeczownika w narzędniku pełniącego funkcję predykatywną czy etykiety zawierającej dodatkową informację. Mimo że pomiędzy wzmianką a tak podaną informacją uzupełniającą nie zachodzi relacja koreferencji, interpretacja łączącej je relacji może być jednak bardzo pomocna w dekodowaniu dalszych odwołań.

1.2. Motywacja

Teoria referencji jest uważana za jeden z ważniejszych składników semantycznej analizy struktury tekstu. Temat ten jest obecnie przedmiotem badań wielu grup naukowych na całym świecie. Jakkolwiek problem nawiązań poruszany był w polskiej literaturze lingwistyczno-informatycznej już wielokrotnie, zjawisko to nie wydaje się jednak dostatecznie zbadane, co widać na przykładzie pojęcia koreferencji: część badaczy używa go zamiennie z anaforą (np. Marciniak 2001), jeszcze inni uznają za podrzędny w stosunku do anafory (np. Matysiak 2007, Broda i in. 2012a), co oznacza, że brakuje systematycznego opisu powszechnego i ważnego zjawiska w sposób możliwy do zastosowania w dalszych badaniach.

Istotną przesłankę do podjęcia badań lingwistyczno-komputerowych tego problemu stanowi to, że większość prac teoretycznych dla polszczyzny powstało w czasach przedkomputerowych, przez co istniejące teorie nie doczekały się jeszcze szeroko zakrojonej weryfikacji tekstowej. Wraz z rozwojem inżynierii lingwistycznej i dostępnością mocy obliczeniowej komputerów badania teoretyczne coraz częściej łączą się z praktycznymi, a podejście korpusowe zapewnia zarówno możliwość ewaluacji istniejących hipotez na szeroką skalę, jak i tworzenie nowych teorii na podstawie obszernych zbiorów danych językowych. Celem opisanych dalej badań jest zatem także weryfikacja obszernego, a niewykorzystywanego jeszcze w ten sposób materiału teoretycznego za pomocą metod lingwistyczno-komputerowych. Proponowane podejście wydaje się też ogólniejsze od dotychczasowych z jeszcze jednego powodu: zarówno częsta w literaturze analiza użyć anaforycznych (nie zapewniająca pełnego pokrycia zjawisk koreferencyjnych – patrz np. Data-Bukowska 2008), jak też jej ograniczenie do grup określonego typu (np. nazw własnych; patrz Maziarz i in. 2016) skłaniają do podjęcia badań nad zjawiskiem referencji w wymiarze ogólnym, na bogatym i dostępnym materiale korpusowym.

Również z perspektywy narzędziowej bieżący stan prac nad identyfikacją relacji referencyjnych wydaje się niewystarczający – wyniki osiągane przez narzędzia au-

tomatyczne są w dużej mierze efektem ich poprawnego działania dla częstych, ale prostych przypadków, w których do rozstrzygania zgodności wystarczą środki analizy powierzchniowej lub proste zależności morfoskładniowe, takie jak zgodność rodzaju i liczby gramatycznej. Z kolei możliwość zastosowania istniejących teorii ogólnych utrudnia ich częsta zależność od złożonych własności semantycznych czy pragmatycznych, takich jak konieczność wcześniejszej znajomości stanu kognitywnego autora wypowiedzi (Gundel i in. 1993) czy struktury dyskursu (Grosz i in. 1995), które dziś nie wydają się możliwe do zdekodowania za pomocą środków lingwistyczno-informatycznych.

Zadanie wydaje się też ważne z perspektywy krajowej – dla języka polskiego takich badań przed rokiem 2010 prawie nie prowadzono; o podejmowanych dotąd próbach piszę dokładniej w rozdziale 2.5. Sam komponent do dekodowania relacji referencyjnych jest także istotnym elementem warstwowego modelu przetwarzania języka, stanowiącym punkt wyjścia do bardziej złożonych operacji, takich jak: automatyczne streszczanie, tłumaczenie, ekstrakcja i analiza tekstu. Pracę umieszczam zatem dodatkowo w kontekście zaznaczonych przeze mnie kierunków rozwoju lingwistyki komputerowej w Polsce (Ogrodniczuk 2017: rozdział 3), które dadzą się streścić hasłem „składnia, semantyka, dyskurs". Relacje referencyjne należą do tej ostatniej, najtrudniejszej grupy.

1.3. Cele badawcze

Wymienione zagadnienia przełożyły się na kilka celów badawczych zrealizowanych w ramach opisywanych prac. Pierwszym i zasadniczym celem było stworzenie ogólnej, weryfikowalnej komputerowo typologii relacji referencyjnych. Zadanie to, podstawowe w przypadku każdego zjawiska naturalnego, jak się wydaje, nie było dotąd wykonane dla języka polskiego, dla innych języków zaś zostało zrealizowane fragmentarycznie. Zaproponowana typologia ma na celu zunifikowanie istniejących cząstkowych opisów relacji referencyjnych i uwzględnienie takich własności, jak: aspekt temporalny referencji, dysymilacja tożsamości obiektów, niejednoznaczność czy niedookreślenie.

Drugim celem, powiązanym z pierwszym, było przeprowadzenie weryfikacji powstałej typologii. W odróżnieniu od metod teoretycznych, wykorzystujących model kompetencji językowej idealnego użytkownika języka, do realizacji tego celu posłużyłem się metodologią korpusową, polegającą na analizie rzeczywistych danych językowych. Prace weryfikacyjne tego rodzaju były dotychczas prowadzone na

bazie korpusów małych (np. Poesio i in. 2004, Korzen i Buch-Kromann 2011), z liczbą i typami relacji ograniczonymi do szczególnych przypadków (np. Markert i in. 2003, Caselli i Prodanof 2006, Lassalle i Denis 2011) i ewaluacją dokonywaną niesystematycznie lub dającą mało obiecujące wyniki (np. Fraurud 1990, Riester i in. 2010). Na potrzeby prac opisywanych w niniejszej książce powstał obszerny (jeden z największych na świecie), zrównoważony i reprezentatywny zbiór tekstów anotowanych ręcznie relacjami referencyjnymi – korpus zależności referencyjnych, zawierający teksty wybrane z Narodowego Korpusu Języka Polskiego (Przepiórkowski i in. 2012). Dzięki powiązaniu z NKJP korpus ten może korzystać z wielopoziomowego opisu lingwistycznego dostępnego dla tekstów bazowych i stale rozszerzanego w badaniach niezależnych lingwistów.

Celem trzecim było stworzenie na bazie powstałego korpusu metod wykrywania relacji referencyjnych zgodnych z zaproponowaną typologią, implementacja wykorzystujących je narzędzi oraz ewaluacja tych narzędzi zgodnie ze stosowaną na świecie metodologią. Ten etap prac umożliwił przetestowanie różnych popularnych w nauce architektur rozwiązań oraz wypracowanie własnego zestawu cech lingwistycznych zapewniającego najlepsze wyniki narzędziowe. Ewaluacji ilościowej towarzyszyła próba oceny użytych algorytmów pod kątem popełnianych przez nie systemowych błędów.

1.4. Zakres badań

Najistotniejsze dla moich badań jest pojęcie koreferencji, do zdekodowania której niezbędne jest zarówno uwzględnienie referencji bez współodniesień (czyli fakt powiązania wzmianki tekstowej z jej desygnatem nawet w przypadku, gdy została przywołana w tekście tylko jeden raz), jak też większości przypadków anafory, której łańcuchy odpowiadają zwykle w pewnym stopniu klastrom koreferencyjnym. W opisie ograniczam się wyłącznie do koreferencji oraz asocjacji z komponentem nominalnym.

Podstawową jednostką badawczą jest dokument, co ogranicza moje działania do **koreferencji wewnątrzdokumentowej** (w odróżnieniu od **koreferencji międzydokumentowej**, czyli rozróżniania w całym zestawie dokumentów, które wzmianki odnoszące się na przykład do George'a Busha dotyczą ojca, a które syna). Przedmiotem badań są wszystkie dziedziny tematyczne i szeroki zestaw relacji (konfiguracja określana często w literaturze angielskim terminem *unrestricted*).

Interesuje mnie zarówno **tożsamość odwołania** (ang. *identity-of-reference*), jak i **tożsamość sensu** (ang. *identity-of-sense*; patrz definicje w rozdziale 3), a także przypadki referencji częściowej, w tym opisywane frazami kwantyfikowanymi, zaimkami upowszechniającymi, zaimkami wskazującymi z frazą podrzędną inną niż względna czy nawiązaniami eliptycznymi (liczne przykłady wyrażeń tego typu zawiera rozdział 3.2). Opisuję także przypadki rozmycia konceptualnego[1] w rozumieniu Fauconniera (1985), gdy jedna ze wzmianek wyróżnia pewną własność drugiej lub następuje pozorne sklejenie referentów w jeden metaobiekt. Badam także pseudoreferencyjne łańcuchy odwołań do obiektów mentalnych wprowadzanych do tekstu za pośrednictwem zaimków nieokreślonych i przeczących oraz wpływ różnorodnych zjawisk lingwistycznych na referencję.

Jak wynika z powziętych deklaracji, przedmiotem badań jest zatem tekst zastany – świadomie rezygnuję z analizy kognitywnych podstaw referencji, jej aspektu poznawczego czy logicznego; nie zamierzam także prowadzić rozważań psycho- ani socjolingwistycznych. Lingwistom pozostawiam opis wpływu referencji na inne zjawiska językowe z dziedziny struktury tekstu, badania nad jego spójnością czy stylistyką. Są to tematy na tyle rozległe, że każdy z nich wymagałby osobnej ścieżki badań.

Do kwestii analizy i anotacji metatekstowej nawiązuję jednak w kontekście prac informatyczno-lingwistycznych rozpoczętych w ramach innych projektów (patrz rozdziały 7.2 i 7.3). Dotychczasowym badaniom teoretycznym przyglądam się w rozdziale 2, ograniczając się do przywołania tych prac językoznawczych, które znalazły odzwierciedlenie w końcowych wersjach opisanych dalej algorytmów. Znacznie obszerniejszy wybór odwołań do tekstów interesujących z punktu widzenia polskich studiów nad zjawiskami referencyjnymi zawiera rozdział 2 monografii angielskojęzycznej (Ogrodniczuk i in. 2015).

1.5. Metodologia

Do analizy relacji referencyjnych została wykorzystana metoda korpusowa. Głównym założeniem tej metody jest próbkowanie rzeczywistych tekstów językowych z reprezentatywnego zbioru w celu uogólnienia otrzymanych wyników. Zalet użycia korpusu jest wiele: rozszerzenie intuicji językowej pojedynczego badacza na szerszą zbiorowość, zapewnienie obiektywnej weryfikacji materiału czy oczywista

[1] Określanego zwykle po angielsku jako *quasi-identity* lub *near-identity*; por. rozdział 3.4.5.

już dziś możliwość wykorzystania technik komputerowych do testowania hipotez naukowych na dużym zbiorze danych. Powstanie korpusu otwiera też wiele możliwości jego wykorzystania jeszcze długo po zakończeniu anotacji, czasem nawet do celów nieuświadamianych sobie przez jego autorów i przy użyciu narzędzi tworzonych za pomocą coraz to nowych metod.

Korpus zależności referencyjnych powstał na bazie tekstów Narodowego Korpusu Języka Polskiego – zasobu wzorcowego współczesnej polszczyzny, za pomocą dobierania próbek metodą losowania w sposób zapewniający zrównoważenie zbioru wynikowego. Do ręcznego oznaczenia tak powstałego korpusu relacjami referencyjnymi zostali zaangażowani eksperci–poloniści. Jednorodność opisu zapewniło opracowanie taksonomii i instrukcji anotacji, czyli dodawania informacji interpretacyjnej do danych tekstowych. Liczbę błędów w tym procesie ograniczono za pomocą porównywania wyników pracy wielu osób, działających niezależnie od siebie. Stabilność uzyskiwanej anotacji przeanalizowano metodą obliczania współczynnika zgodności anotatorów, eliminującego wpływ przypadku, końcową postać danych uzyskano zaś wypracowując optymalną strategię superanotacji.

Po zakończeniu fazy opracowania korpusu powstały narzędzia do automatycznego wykrywania relacji referencyjnych kilkoma różnymi metodami. Algorytmy opracowano metodą analizy – ręcznej i automatycznej – wydzielonego podkorpusu treningowego. Jakość powstałych rozwiązań oceniono metodą 10-krotnej walidacji krzyżowej na pozostałej części korpusu z wykorzystaniem standardowych, uznanych w środowisku miar efektywności wykrywania wzmianek, koreferencji i relacji pośrednich.

2.

Od ujęć teoretycznych do dekodowania relacji referencyjnych

Zjawisko referencji jest przedmiotem badań wielu dziedzin nauki: filozofii, semantyki logicznej, językoznawstwa. W niniejszej pracy szczególnie interesuje mnie aspekt lingwistyczno-komputerowy, zatem przedstawiony dalej przegląd jest syntezą tych elementów teoretycznych i praktycznych, które wydają się istotne z punktu widzenia komputerowego przetwarzania zależności referencyjnych. Przechodząc od rozważań teoretycznych nad relacjami referencyjnymi do opisu prób zastosowania wypracowanych teorii w pracach informatycznych, porównuję także prace prowadzone na świecie z badaniami języka polskiego. Poruszane tu problemy w dużej części były już opisywane wcześniej (Ogrodniczuk i in. 2015), większość zagadnień przedstawiam zatem skrótowo. W szczególności omówienie bogatej polskiej literatury teoretycznej ograniczam do aspektów wykorzystanych w późniejszych pracach informatycznych.

2.1. Pojęcie i zakres referencji

Pojęcie **referencji** – odniesienia fragmentu wypowiedzi do pozajęzykowego bytu, o którym ten fragment mówi – analizowano w semantyce logicznej już od ponad 150 lat, przeciwstawiając je pojęciu **sensu**, czyli wewnątrzjęzykowego odwołania danego wyrażenia do innych elementów systemu językowego. U Milla (1843) były one nazywane **konotacją** i **denotacją**, u Fregego (1892) **nominatem** i **sensem**, u Russella (1905) **denotacją** i **znaczeniem**, u Carnapa (1947) **ekstensją** i **intensją**, u Blacka (1949) **referencją** i **sensem**. Z biegiem lat zmieniała się jednak interpretacja tych terminów, zarówno w zakresie przypisywania (lub odmawiania) referencyjności określonym typom wyrażeń, jak i interpretacji stopnia pełności odwołania.

Badania nad zależnościami referencyjnymi w polszczyźnie sięgają okresu międzywojennego. W pracy Klemensiewicza (1937)[1] *wskaźniki nawiązania* są elementem szeroko zakrojonego modelu składniowego opisu języka, a analiza ich użycia stanowi pierwszą polską systematyczną próbę badania związków międzyzdaniowych na poziomie semantycznym za pomocą *relacji odniesienia zewnętrznego*. Klemensiewicz był też autorem podstawowej polskiej terminologii w dziedzinie referencji: wprowadził pojęcia *podstawa nawiązania* i *człon nawiązujący*, których realizacją były fragmenty tekstu odpowiadające nie tylko pojedynczym frazom, ale także całym zdaniom lub nawet akapitom.

Referencja jest właściwością użyć leksemów w tekście, a nie częścią ich znaczenia, „dotyczy nie wyrazów i wyrażeń języka, lecz tylko ich użyć w tekście – wypowiedzi i jej składników" (Paduczewa 1992: s. 12). Oznacza to, że istnieją różnokształtne wyrażenia tekstowe odwołujące się do tego samego referenta, czyli posiadające to samo znaczenie (jak *Gwiazda Poranna* i *Gwiazda Wieczorna* na określenie Wenus, patrz Kripke 2001: s. 44 i dalsze) oraz równokształtne i równoznaczne wyrażenia o różnej referencji (jak w zdaniu *Są matki i matki*). W przeciwieństwie do Searle'a (1975) czy Lyonsa (1977) Paduczewa zauważa także, że nieistotne jest ograniczenie referencji wyłącznie do świata rzeczywistego. Rozumienie to przyjmuje także Langacker (2008: s. 353), kwestionując w ogóle istnienie grup nominalnych niereferencyjnych. Myśl tę rozwija Kunz (2010), według której podczas przetwarzania tekstu przez odbiorcę powstaje i podlega interpretacji **mentalny świat tekstu** (ang. *mental textual world*), w którym odwołanie znajdują zarówno obiekty rzeczywiste, jak i wyobrażone, a także przywołane w wypowiedzi fakty hipotetyczne czy idee i pojęcia abstrakcyjne.

Kluczowe dla interpretacji charakterystyki referencyjnej obiektów jest wprowadzone przez Topolińską (1976: s. 60–62) pojęcie **wyznaczania sytuacyjnego**, wiążące obiektywny układ faktów językowych i pozajęzykowych ze świadomą intencją komunikatywną nadawcy tekstu oraz stopniem wiedzy o świecie oczekiwanej od odbiorcy. Poprawne dekodowanie referencji wymaga zatem określonej wiedzy ogólnej przekraczającej granice kompetencji językowej, znajomości poprzedzającego tekstu, świadomości konkretnej umowy społecznej wiążącej nazwy własne z ich referentami oraz znajomości realiów dotyczących referentów wyznaczonych w kontekście konkretnego aktu komunikacji. W pracy z roku 1984 Topolińska dodatkowo wskazuje na znaczenie idiolektycznej charakterystyki wrażenia różnicy semantycznej między wyrażeniami oraz wystarczający dla konstrukcji relacji

[1] Por. też kolejne wydania rozszerzone i poprawione (Klemensiewicz 1948, 1950, 1982).

warunek „bliskości znaczeniowej” (a nie tylko zastępowalności synonimicznej czy hiponimicznej), nawet w sytuacji różnic interpretacyjnych pomiędzy nadawcą a odbiorcą, jak w przykładach: *Pani A: Pokazały się ostatnio śliczne fajansowe kubki do mleka*[2]. *Pani B: Ach, takie filiżanki w kwiatki?* oraz *Pani A: Włożę dziś tę szarą płócienną sukienkę. Pani B: Ach, tę zieloną?*

Osobny problem stanowi kwestia interpretacji tożsamości referentów (czyli ich indywidualności, odrębności lub identyczności) w sytuacji częściowej zmiany ich własności. Często przyjmowana charakterystyka identyczności jako uniwersalnej, podstawowej i nierozkładalnej jednostki semantycznej (por. np. Wierzbicka 2010: s. 61) bywa kwestionowana w ujęciach uwzględniających takie czynniki zaburzające postrzeganie relacji identyczności, jak zmiana w czasie czy strukturze. Na przykład Fauconnier i Turner (2002) rozważają istnienie przestrzeni mentalnych (ang. *mental spaces*) – powstających podczas myślenia i mówienia modeli kognitywnych, które odbiorca komunikatu na bieżąco analizuje i syntezuje, decydując które obiekty należy utożsamić, a które rozdzielić. Propozycja Fauconniera i Turnera wpłynęła m.in. na rozumienie identyczności u Recasens i in. (2011), którzy wprowadzają stopniowalność tożsamości obiektu oraz pojęcie **częściowej identyczności** (ang. *near-identity*) na opisanie sytuacji, w której relacja identyczności między referentami nie zachodzi w sposób pełny. Wśród przykładów tego rodzaju autorzy wymieniają m.in. sytuację referencji pomiędzy postacią a jej przedstawieniem na obrazie i podają dość szczegółową taksonomię częściowej identyczności, w skład której wchodzą takie relacje, jak: metonimia, meronimia czy przesunięcie temporalno-przestrzenne.

2.2. Klasyfikacje typów wzmianek i relacji referencyjnych

Istniejące klasyfikacje rodzajów wzmianek i relacji referencyjnych uwzględniają różne aspekty relacji referencji i z tego względu nie są bezpośrednio porównywalne. Dalej przytaczam wraz z oryginalnymi przykładami te z nich, które wywarły największy wpływ na proponowaną w kolejnym rozdziale definicję wzmianki i taksonomię relacji referencyjnych.

[2] Podkreśleniem linią ciągłą oznaczam w przykładach w dalszej części tekstu wyrażenia koreferencyjne, linią przerywaną – zależne, ale o rozłącznej referencji. Symbolu Ø używam w miejscu wystąpienia wyrażenia eliptycznego.

2.2.1. Klemensiewicz

Klemensiewicz (1937) analizuje relacje składniowe między podstawą nawiązania (PN) a członem nawiązującym (CN), przyjmując za podstawę klasyfikacji relacji referencyjnych hierarchię wskaźników nawiązania:

1. wskaźniki gramatyczne:
   (a) spójniki: PN: *Prawdą żywą staje się tylko przeżycie, pozadoświadczalne wyczucie, które się w samym fakcie życia objawia.* CN: *Prawda zatem jest nieskończoną i objawiającą się, jak nieskończonym i objawiającym się jest życie.*,
   (b) zaimki anaforyczne: PN: *Zadawał pytania starszy z oficerów, porucznik.* CN: *Jego ciemna twarz sportowca o rysach twardych i nieregularnych wyrażała chłód i pogardę.*,
   (c) konstrukcje werbalne, odnoszące się do podstawy nawiązania: PN: *Dziewczyna zaśpiewała.* CN: *Podobało się.*,
   (d) części zdania (atrybuty, dopełnienia): PN: *Z seminarium duchownego idą klerycy.* CN: *Na spacer.* CN: *Po obiedzie.*,
   (e) zaimki pytajne: PN: *Kto przyszedł?* CN: *Piotr.*;
2. wskaźniki leksykalne:
   (a) wyrażenia niepełnoznaczne: PN: *Na wszystkie pytania leśniczy rudawickich lasów odpowiadał jednakowo.* CN: *Broń, którą nieopodal...* (z intencją odniesienia się do aktu odpowiadania),
   (b) synonimy,
   (c) wyrażenia analogiczne (*po pierwsze – po wtóre*, *naprzód – potem – w końcu*),
   (d) zaimki pytajne: PN: *Kiedy wyjeżdżasz?* CN: *Jutro.*;
3. wskaźniki tematyczne: PN: *Pójdziesz na koncert?* CN: *Nie wiem.*

Klemensiewicz wskazuje też dwie funkcje wyrażeń nawiązujących: *powiązanie*, zachodzące między parą niezależnych wyrażeń, z których drugie w jakiś sposób rozszerza pierwsze, ale z zachowaniem jego odrębności, oraz *włączenie*, gdy zrozumienie członu nawiązującego wymaga obecności podstawy nawiązania.

Z punktu widzenia badań nad spójnością tekstu koncepcja Klemensiewicza wydaje się niezwykle pojemna i odpowiada teorii nazywanej obecnie „gramatyką tekstu" lub „strukturą dyskursu" (tworzonego pomiędzy częściami wypowiedzi powyżej poziomu zdaniowego), do której autor odwołuje się w klasycznej już książce na temat polskiej składni (Klemensiewicz 1953). Niektóre elementy hierarchii wskaźników nawiązania wydają się jednak pochodzić z poziomów analizy innych niż referencyjny – relacje międzyzdaniowe dadzą się niekiedy analizować środkami czysto składniowymi, „tematyczne" wskaźniki odniesienia dotyczą zaś warstwy pragmatycznej (ten akurat poziom został zresztą usunięty przez autora w kolejnych wydaniach publikacji).

2.2.2. Topolińska

Topolińska (1984: s. 303–324) zajmuje się charakterystyką grup imiennych (nazw przedmiotów materialnych), wyróżniając grupy:

1. z referentem jednostkowym:
   (a) deskrypcje określone językowo zupełne (o jednoznacznej referencji), np. *stolica Polski za Jagiellonów*, *autor Pana Tadeusza*,
   (b) deskrypcje określone językowo niezupełne (których sama formalizacja językowa nie zapewnia jednoznacznej referencji lub kiedy referencja zmienia się wraz z sytuacją mówienia):
      i. wyznaczające jednoznacznie w określonej sytuacji, np. *Swędzi mnie ręka.*,
      ii. skorelowane z gestem jednoznacznego odniesienia, np. *Daj mi ten nóż!*,
   (c) grupy imienne w funkcji wyrażeń argumentowych nieidentyfikujących, np. *Coś mi wpadło do oka.*;
2. z referentem zbiorowym:
   (a) konstytuujące zbiorowość (pojmowaną dystrybutywnie lub kolektywnie), np. *Te kamienie są mokre.*, *Te kamienie ważą pół tony.*,
   (b) różnicujące element zbiorowości (wyróżniając wszystkie, część lub jeden z elementów), np. *Ostatniej nocy w Tel Awiwie Żydzi zaatakowali restaurację zatrudniającą Palestyńczyków.*

Klasyfikacja Topolińskiej wskazuje frazy nominalne jako jedyne jednostki, którym przysługuje charakterystyka referencyjna, tj. odniesienie do przedmiotu, który nazywają. Decyzja ta miała kluczowe znaczenie dla niniejszej pracy, stanowiąc podstawę ograniczenia opisu relacji do szeroko rozumianych grup nominalnych z podrzędnikami.

2.2.3. Paduczewa

Paduczewa (1992: s. 118–126) dzieli referencyjne grupy imienne na:

1. określone, np. *Ernest Hemingway urodził się w 1899 r.*, *Wszyscy moi studenci zaliczyli kolokwium.*;
2. słabo określone, np. *Mam ci coś do powiedzenia.*;
3. nieokreślone dla mówiącego, np. *Ktoś zjadł mój jogurt.*;
4. ze zneutralizowaną kategorią określoności, np. *Zatrzymał mnie policjant*. (z braku rodzajnika nie jesteśmy w stanie określić, czy nadawca ma na myśli *jakiegoś nieokreślonego policjanta* czy *konkretnego policjanta*).

Grupy niereferencyjne, czyli nieoznaczające żadnych wyróżnionych obiektów, dzielą się z kolei na:

1. egzystencjalne, które odnoszą się do klas obiektów, ale nie wyróżniają żadnego z nich:
   (a) dystrybutywne grupy imienne oznaczające uczestników rozdzielonych w pewnym zbiorze zdarzeń jednego typu (Paduczewa 1992: s. 127), np. *Czasami ktoś z nas go odwiedza.*, *Do każdego wychowanka przyjechali jego krewni.*;
   (b) niekonkretne grupy imienne, występujące w kontekście stłumionej asercji (tzn. z czasownikami *może*, *chce*, *powinien*, *należy*, z formami trybu rozkazującego, w pytaniach, negacji, z czasownikami performatywnymi itp.), np. *Jan chce się ożenić z jakąkolwiek cudzoziemką.*;
   (c) ogólnoegzystencjalne grupy imienne, odnoszące się do obiektów w sposób ogólny, bez wyróżniania konkretnego okazu, np. *Niektórzy ludzie mają alergię na gluten.*;

2. uniwersalne, odnoszące się do całej, abstrakcyjnej klasy obiektów, np. *Kto rano wstaje, temu Pan Bóg daje.*;

3. atrybutywne, które odnoszą się do jakiegoś jednostkowego bytu, ale nadawca nie ma na myśli jakiegoś konkretnego obiektu, np. *Najsilniejszy człowiek na świecie nie podniósłby 500 kg.*, *Ten, kto wygra, otrzyma nagrodę.*;

4. oznaczające rodzaj lub gatunek, np. *On postąpił jak mężczyzna.*, *Jaguary wymierają.*

Klasyfikacja Paduczewej została do opisywanych prac zaadaptowana w sposób krytyczny na mocy spostrzeżenia, że grupy uznane za niereferencyjne mogą tworzyć w tekście łańcuchy przypominające klastry koreferencyjne.

2.2.4. Clark i inne klasyfikacje zagraniczne

O ile opis typów wzmianek na potrzeby badań nad ich referencyjnością był przedmiotem badań lingwistów polskich, relacje referencyjne nie były dotąd przez nich opisywane wystarczająco systematycznie, warto zatem przyjrzeć sie najpopularniejszej w środowisku anglojęzycznym klasyfikacji relacji tego rodzaju przedstawionej przez Clarka (1977):

1. referencja bezpośrednia (ang. *direct reference*):

   (a) identyczność (ang. *identity*): *Spotkałem wczoraj człowieka. Ten człowiek opowiedział mi swoją historię.*

   (b) pronominalizacja (ang. *pronominalization*): *Spotkałem wczoraj człowieka, a on opowiedział mi swoją historię.*

   (c) epitet (ang. *epithet*): *Spotkałem wczoraj człowieka. Ten łajdak ukradł mi wszystkie pieniądze!*

   (d) przynależność do zbioru (ang. *set membership*): *Spotkałem wczoraj dwoje ludzi. Kobieta opowiedziała mi swoją historię.*;

2. referencja pośrednia (ang. *indirect reference by association*):

   (a) część konieczna (ang. *necessary part*): *Zajrzałem do pokoju. Sufit był bardzo wysoki.*

(b) część prawdopodobna (ang. *probable part*): *Wszedłem do pokoju. Okna wychodziły na zatokę.*

(c) część dedukowalna (ang. *inducible part*): *Wszedłem do pokoju. Żyrandole zaświeciły jasno.*;

3. charakterystyka:

(a) rola wymagana (ang. *necessary role*): *Jan został wczoraj zamordowany. Mordercy udało się uciec.*

(b) rola opcjonalna (ang. *optional role*): *Jan został wczoraj zamordowany. Nóż leżał obok ciała.*;

4. związki przyczynowo-skutkowe, konsekwencja i równoczesność:

(a) uzasadnienie (ang. *reason*): *Jan upadł. Chciał przestraszyć Marię.*

(b) przyczyna (ang. *cause*): *Jan upadł. Potknął się o kamień.*

(c) konsekwencja (ang. *consequence*): *Jan upadł. Złamał sobie rękę.*

(d) równoczesność (ang. *concurrence*): *Jan jest republikaninem. Maria jest równie głupia.*

Dużą zaletą taksonomii Clarka jest jej uniwersalność dzięki połączeniu w jeden system relacji bezpośrednich i pośrednich, słabością wydaje się natomiast wspólne traktowanie relacji metatekstowych i anaforycznych środków wyrazu, będących pojęciami z dwóch różnych płaszczyzn interpretacyjnych.

Klasyfikacja Clarka stała się podstawą kilku kolejnych, zwłaszcza w kontekście aplikacyjnym i zakresie odnoszącym się do relacji pośrednich. Ich podsumowanie zawiera artykuł Gardent i in. (2003), wymieniający trzynaście najczęściej stosowanych kategorii relacji pośrednich: *zbiór – podzbiór*, *zbiór – element*, *zdarzenie – argument*, *osoba – funkcja*, *osoba – atrybut*, *całość – część integralna*, *całość – część wyodrębnialna*, *całość – część tymczasowa*, *osoba – przedmiot*, *kolekcja – element*, *miejsce – obszar*, *miejsce – obiekt* i *czas – obiekt*. Proponują też własną klasyfikację użytą w anotacji korpusu PAROLE, ograniczoną do *relacji włączającej* (przynależność do zbioru), *relacji tematycznej* (agens, patiens, adresat, instrument itp. – czyli w zasadzie funkcji semantycznej), *relacji definicyjnej* (atrybut, meronim itp.), *relacji współuczestnictwa* i *relacji nieleksykalnej* (definiowanej strukturą wypowiedzi lub na podstawie wiedzy ogólnej).

W wielu przypadkach anotacji korpusowej klasyfikacje te okazały się zbyt szczegółowe: np. schemat anotacji anaforycznej korpusu ARRAU (Poesio i Artstein 2008), używany wcześniej z powodzeniem w anotacji korpusów GNOME (Poesio 2000) i VENEX (Poesio i in. 2004) ogranicza się do relacji *całość – część*, *przynależności do zbioru* oraz *konwersji*. Tego rodzaju ograniczona lista relacji była też stosowana w anotacji korpusu CESS-ECE (Recasens i in. 2007), wyróżniającego trzy relacje podstawowe oraz relację dodatkową na oznaczenie pozostałych rodzajów relacji. Popularny w środowisku badaczy języków słowiańskich Praski Korpus Zależnościowy (Prague Dependency Treebank, PDT) w jego obecnej wersji 3.0 (Zikánová i in. 2015: rozdział 4) używa sześciu typów relacji: *całość – część*, *zbiór – podzbiór/element*, *obiekt – funkcja*, *kontrast* (do łączenia przeciwieństw w przypadkach, gdy ich użycie wpływa na spójność tekstu), *jawna niekoreferencyjność* oraz *pozostałe* (niekategoryzowana ściślej grupa relacji pośrednich, których przykłady to *miejsce – mieszkaniec*, *autor – dzieło*, *zdarzenie – argument* czy *jednostka – narzędzie*).

2.3. Cechy relacji referencyjnych

Naturalną inspiracją dla badań korpusowych, których przedmiotem jest analiza cech tekstu polskiego ekstrahowalnych metodami komputerowymi, jest bogata polska literatura lingwistyczna. Praktycznym ograniczeniem w zastosowaniu zaproponowanych w niej metod analizy mechanizmów referencji jest dostępność narzędzi i zasobów dla języka polskiego, uwzględniających obecnie jedynie część własności językowych analizowanych przez badaczy-teoretyków. Efektywność metod opartych o sztuczne sieci neuronowe pozwala jednak sądzić, że niektóre z własności trudno reprezentowalnych w prostych algorytmach regułowych są jednak niejawnie uwzględniane w wektorowych reprezentacjach dystrybucyjnych tworzonych na bazie dużych korpusów. Dalej podsumowuję najważniejsze prace, które były inspiracją do stworzenia implementowanych algorytmów i posłużyły jako źródło cech istotnych w procesie dekodowania koreferencji.

Na definicję wzmianki wpłynęły przede wszystkim wspomniane już prace Topolińskiej (1984) i Paduczewej (1992), uzupełnione o wnioski z wcześniejszych prac Bellert (1971), wprowadzającej pojęcie *indeksu językowego* jako nazwy łącznika tekstowego realizowanego jako rzeczownik pospolity, grupa nominalna, nazwa własna, zaimek osobowy, względny lub zwrotny służący jako wyrażenie referencyjne,

oraz Grochowskiego (1976), badającego funkcje elipsy w strukturze linearnej tekstu.

Algorytmy dekodowania koreferencji wykorzystujące cechy leksykalne nawiązują m.in. do prac Pisarkowej (1969), badającej rozkład zaimków w wypowiedzeniach polskich i ich funkcje wewnątrzzdaniowe. Pisarkowa wskazuje m.in. interesującą różnicę dystrybucyjną w użyciu zaimków i ich odpowiedników nominalnych: te ostatnie pojawiają się, gdy mogłyby zawieść tradycyjne środki ujednoznaczniania. Co więcej, powtórzenia nominalne wymagają wówczas użycia zaimków wskazujących (*ten*, *ta*, *to*) lub synonimu świadczącego o tym, że dany desygnat powinien być znany z kontekstu.

Źródłem cech leksykalnych są także m.in. prace Fontańskiego (1986) i Grzegorczykowej (1996). Fontański bada warunki występowania w tekstach przymiotnikowych zaimków anaforycznych w odniesieniu do dwóch określonych wariantów tekstu: ewokacyjnego i nieewokacyjnego (narracyjnego). W wyrażeniach ewokacyjnych przymiotnikowe zaimki wskazujące takie jak *ten*, *tamten*, *ów* są używane znacznie rzadziej niż w bezpośredniej narracji. Grzegorczykowa wskazuje natomiast na znaczenie w procesie dekodowania odwołań leksemów o wbudowanej informacji anaforycznej, czyli takich, które wymagają obecności określonej informacji lub zakładają ją domyślnie. Przykładami takich leksemów są cząstki nawiązujące do wcześniej znanych faktów lub zdarzeń (*wreszcie*, *dopiero* itp.), przymiotniki określające stopień podobieństwa (*podobny*, *inny*), liczebniki *oba*, *obie* czy przymiotniki pochodne (*obopólny*, *obustronny*). Podobną funkcję pełnią niektóre czasowniki (jak *przeprosić*, którego użycie pozwala sądzić, że ktoś zrobił komuś innemu coś złego itp.) czy rzeczowniki (*sąsiad*, *kolega*, *przyjaciel*, które wymagają odwołania do obiektu innego niż podmiot – w przeciwieństwie do zaimka *swój*).

Kwestia wpływu szyku wyrazów w języku polskim na procesy anaforyzacyjne, analizowana m.in. przez Szwedek (1975) i Duszak (1986), znalazła odzwierciedlenie w zestawie cech dotyczących pozycji wzmianki w zdaniu. Badaczki zwracają uwagę na powiązanie koreferencyjności z akcentem zdaniowym i wykazują, że rzeczowniki o interpretacji nieokreślonej mają tendencję do pojawiania się w końcowej części zdania, podczas gdy rzeczowniki określone – na pozycjach wcześniejszych. Cechy takie jak odległość wzmianek w zdaniach czy ich obecność w tym samym akapicie nawiązują także do badań Honowskiej (1984), wskazującej na różnice między koreferencją zaimkową wewnątrz- i międzyzdaniową, kontrastując za-

imek zwrotny *się* i anaforyczny *go*, z których tylko ten ostatni może tworzyć linki[3] przekraczające granice zdań.

Uwagi Topolińskiej (1984), a za nią Grzegorczykowej (1990), że główne czysto formalne techniki anaforyzacyjne (pronominalizacja i powtórzenie) muszą zostać wzbogacone o zabiegi semantyczne oraz wiedzę ogólną zdecydowały o włączeniu do algorytmów cech je symulujących, opartych na sieciach semantycznych (Piasecki i in. 2009, Vetulani 2014) czy zasobach polskiej Wikipedii.

Wiele ciekawych własności anafory, które mogły zostać uwzględnione niejawnie, podaje Pasek (1991), argumentując, że do jej poprawnego dekodowania wymagana jest:

– wiedza semantyczna (o kategoriach obiektów mogących stać się argumentami predykatów określonych typów), jak w zdaniu *Położyłem ołówek na stole, ale Ø był pochyły i Ø się zsunął.* (to stoły są pochyłe, a ołówki mogą się z nich zsuwać),

– wiedza psychologiczna (temat zdania), jak w zdaniu *Jan powiedział Piotrowi, że Ø jest łobuzem.* (pejoratywne określanie samego siebie zachodzi stosunkowo rzadko),

– świadomość ogólnie akceptowanych norm, rozumienie sytuacji i ludzkiego zachowania, jak w zdaniach *Maria pokonała Annę, ponieważ Ø lepiej grała.* i *Maria zbeształa Annę, ponieważ Ø postąpiła lekkomyślnie.* (pokonanie kogoś oznacza lepszą grę; powodem besztania może być złe zachowanie).

Bezpośrednio implementowalny obszerny zestaw klas czynników wpływających na interpretację anafory wymienia natomiast Fall (1994). Są nimi: zgoda fleksyjna, ograniczenia składniowe i semantyczne czy istotność elementu w zdaniu. Wszystkie wymienione cechy zostały wprost użyte przez narzędzia powstałe w ramach pracy (patrz rozdział 5.2.2).

Analiza zrozumiałości tekstów tworzonego korpusu oraz badania nad zależnością między obecnością linków anaforycznych a zrozumiałością tekstu zostały zainspirowane pracą Marciszewskiego (1983)[4] kontrastującą *integralność składniową*

[3] W niniejszej pracy używam terminu *link* wymiennie z terminem *relacja* ze względu na jego zwięzłość i mniejszą wieloznaczność. Decyzja ta znajduje dodatkowe uzasadnienie w kontekście prowadzonych prac anotacyjnych – relacje tekstowe oznaczane są w dokumentach elektronicznych właśnie za pomocą linków.

[4] Por. też badania nad spójnością referencyjną dyskursu naukowego Stroińskiej (1992), Szkudlarek-Śmiechowskiej (2003) czy Trofimiec (2007).

tekstu (spójność) z jego *integralnością semantyczną* (koherencją) i wykazującą, że mnogość linków anaforycznych w tekście oraz ciągłość tematyczna są wciąż niewystarczające do określenia tekstu jako spójnego.

W porównaniach gęstości relacji wewnątrzdokumentowych z podziałem na analizowane typy tekstów znalazły odzwierciedlenie badania Gajdy (1982, 1990) i Pisarek (2012) wykazujące, że gęstość wyrażeń referencyjnych zależy od gatunku tekstu, ze znacznie większym udziałem odniesień w publikacjach naukowych w stosunku do tekstów literackich, co jest wynikiem większej nominalizacji tekstów naukowych (stosunek liczby rzeczowników do czasowników wynosi 4,2 dla publikacji naukowych, 3,3 dla tekstów artystycznych, 1,1 dla beletrystyki i tylko 0,8 dla tekstów mówionych – patrz Gajda 1982). Najpowszechniejszym typem linku anaforycznego jest powtórzenie leksykalne, co badacz wyjaśnia wysokim stopniem wiązania oraz precyzją, jakimi charakteryzują się tego rodzaju konstrukcje.

W rozpoczętych badaniach nad rolą koreferencji w strukturze metatekstu pobrzmiewa echo prac Wajszczuk (1978), która uznaje wiązanie za element szerszej teorii spójności tekstu i bada powiązania między kolejnymi wypowiedziami w procesie tworzenia spójnej wypowiedzi. Jej spostrzeżenia co do analogii wiązania fragmentów tekstu relacjami anaforycznymi do łączenia zdań składowych zdania złożonego za pomocą spójników są do dziś aktualne i znajdują potwierdzenie w pracach nad strukturą metatekstową we współczesnych zagranicznych korpusach dyskursu, takich jak Penn Discourse Treebank (patrz rozdział 7.2).

2.4. Projekty korpusowe

Chociaż rozkwit badań teoretycznych nad zależnościami referencyjnymi w tekście przypadł na lata 80. i 90. minionego wieku, dopiero podejście korpusowe przyniosło znaczący przełom, także ze względu na rozwój wielkoskalowych metod ewaluacyjnych. Biorąc pod uwagę duże projekty anotacyjne (w szczególności dla języków pro-drop) wśród najważniejszych korpusów relacji referencyjnych, powstałych w ciągu ostatnich dziesięciu lat należy wymienić:

– OntoNotes, wielojęzyczny system anotowanych wielopoziomowo korpusów angielskiego, hiszpańskiego, chińskiego i arabskiego (Pradhan i in. 2007) z anotacją relacji referencyjnych wykraczającą poza frazy nominalne, wyróżnieniem grup apozycyjnych oraz częściową anotacją wyrażeń ogólnych, niedospecyfikowanych lub abstrakcyjnych;

– NAIST, korpus języka japońskiego (Iida i in. 2007) z anotacją: koreferencji, podmiotów domyślnych, relacji tożsamości znaczeniowej;

– ARRAU, korpus języka angielskiego Poesio i Artsteina (2008), zawierający anotację: deiksy, niejednoznaczności referencyjnej oraz relacji asocjacyjnych;

– COREA, holenderski korpus prasowy (Hendrickx i in. 2008), anotowany: relacjami koreferencji, anafory związanej, relacjami asocjacyjnymi i predykatywnymi;

– AnCora-CO, korpus języka hiszpańskiego i katalońskiego (Recasens i in. 2010) z anotacją: konstrukcji eliptycznych, mowy zależnej, konstrukcji dzierżawczych, atrybutywnych i oznaczeniem fraz ogólnych, reprezentacją metonimii, deiksy, anafory związanej i relacji asocjacyjnych;

– Copenhagen Dependency Treebank (Korzen i Buch-Kromann 2011), korpus równoległy języka duńskiego z tłumaczeniami tekstów na angielski, niemiecki, włoski i hiszpański, anotacją relacji asocjacyjnych i typologią koreferencyjną;

– DIRNDL (Eckart i in. 2012, Björkelund i in. 2014), niemiecki korpus nagrań radiowych z anotacją informacji referencyjnej zgodną ze schematem RefLex (Baumann i Riester 2012);

– ISNotes (Hou i in. 2013), korpus języka angielskiego wzbogacający anotację podkorpusu Wall Street Journal z korpusu OntoNotes o opis relacji pośrednich;

– ANCOR, francuski korpus mowy spontanicznej (Muzerelle i in. 2013) z anotacją koreferencji fraz zagnieżdżonych oraz relacji asocjacyjnych (bliżej niekategoryzowanych);

– GECCo, German-English Contrasts in Cohesion (Lapshinova-Koltunski i Kunz 2014), wielomodalny korpus anotowany relacjami spójnościowymi, w tym koreferencyjnymi i pośrednimi;

– PCC, Potsdam Commentary Corpus (Stede i Neumann 2014), niemiecki korpus komentarzy prasowych z anotacją koreferencji nominalnej zgodną ze schematem PoCoS – Potsdam Coreference Scheme Krasavina i Chiarcos (2007) oraz bazową anotacją składniową;

– Prague Dependency Treebank (Zikánová i in. 2015: rozdziały 3 i 4), anotowany relacjami koreferencji gramatycznej (zachodzącymi w ramach zdania na bazie

stosunków składniowych) oraz tekstowej (wyrażanymi środkami pozagramatycznymi, mającymi charakter kontekstowy), z oznaczeniami: deiksy, apozycji, wyrażeń predykatywnych, konstrukcji skoordynowanych czy odpowiadających pojęciom konkretnym i ogólnym oraz relacjami asocjacyjnymi;

– PDC, Phrase Detectives Corpus (Chamberlain i in. 2016), anotowany metodą „pożytecznej zabawy" (ang. *game-with-a-purpose*, GWAP) relacjami zgodnymi ze schematem korpusu OntoNotes.

Systematyczna anotacja relacji referencyjnych w polskich zasobach ogranicza się do czterech pozycji[5]:

– NN, anaforycznego korpusu polszczyzny (Filak 2006), anotowanego relacjami pomiędzy zaimkiem a poprzednikami realizującymi relacje: anafory nominalnej, werbalnej, egzofory i wyrażeń idiomatycznych;

– korpusu LUNA (Marciniak 2010), zawierającego anotację nominalnych, pronominalnych oraz przysłówkowych relacji anaforycznych i kataforycznych, w których skład wchodziły pojęcia z predefiniowanej ontologii dotyczące transportu publicznego;

– Korpusu Języka Polskiego Politechniki Wrocławskiej (KPWr, patrz Broda i in. 2012b, Maziarz i in. 2016), ograniczającego analizę koreferencji do powiązań, w których skład wchodzą nazwy własne, z linkami łączącymi centra fraz oraz uwzględnieniem podmiotów domyślnych;

– opisywanego w niniejszej publikacji korpusu zależności referencyjnych (ang. *PCC – Polish Coreference Corpus*).

Ważnym czynnikiem w cytowanych wyżej pracach korpusowych jest to, że wykorzystują istniejące wcześniej korpusy bogato anotowane relacjami lingwistycznymi. Takim korpusem jest zdecydowanie Prague Dependency Treebank, którego drzewiasta anotacja gramatyczna oraz tektogramatyczna znakomicie uzupełnia anotację referencyjną i stanowi bogate źródło informacji dla narzędzi automatycznych budowanych na podstawie korpusu; podobne podejście stosuje KPWr, bazując na automatycznej anotacji morfoskładniowej.

[5] Świadomie pomijam zasoby eksperymentalne lub niedostępne, takie jak np. opisany w pracy Mitkova (1998) korpus tekstów stworzony na potrzeby narzędziowe na bazie internetowych podręczników i zawierający 180 ręcznie wyróżnionych zaimków.

Syntetyczne porównanie opisanych schematów (uwzględniające również schemat prezentowany w niniejszej książce) przedstawiają tabele 2.1 i 2.2.

2.5. Komputerowe implementacje modelu referencji

Na świecie pierwsze próby komputerowego przetwarzania relacji referencyjnych, początkowo ograniczonych do anafory, podjęto w latach 70. XX w.[6] Proponowane rozwiązania obejmowały algorytmy regułowe, w późniejszym okresie wykorzystujące mechanizmy wnioskowania, pierwsze reprezentacje wiedzy ogólnej oraz coraz bardziej wyrafinowane poziomy wiedzy lingwistycznej, od reguł składniowych (Hobbs 1976, 1978) po teorię centrowania opartą na założeniu, że tekst skupia się w danym momencie na pewnym zestawie tematów, które wzmacniają jego spójność (Grosz 1977, Sidner 1979, Brennan i in. 1987). Jako przeciwwaga pojawiły się propozycje rozwiązań o ograniczonym wykorzystaniu reprezentacji wiedzy (ang. *knowledge poor approaches*; patrz np. Mitkov i Styś 1997).

Proponowane algorytmy pozostawały często na papierze lub były ewaluowane na niewielką skalę, często ręcznie, w zakresie co najwyżej setek przykładów. Wraz ze wzrostem popularności metod korpusowych w połowie lat 90. rozpoczęto prace nad formalnymi metodami oceny jakości prezentowanych rozwiązań na podstawie większych zbiorów danych. Na konferencji MUC–6 (Message Understanding Conference; patrz też rozdział 2.6.1) na potrzeby zadania identyfikacji koreferencji jako procesu wspierającego zasadniczy temat ekstrakcji informacji opracowano instrukcję anotacyjną, stworzono na jej podstawie korpusy treningowy i testowy (po 30 dokumentów) oraz metrykę jakościową MUC (Vilain i in. 1995) oraz wykorzystujące ją narzędzie ewaluacyjne (ang. *scorer*).

Możliwość formalnej ewaluacji rozwiązań rozpoczęła okres intensywnych prac implementacyjnych oraz poszukiwanie optymalnego modelu reprezentacji relacji referencyjnych na potrzeby używanych algorytmów. Przez ok. 10 lat najpowszechniej były stosowane dwuetapowe algorytmy parujące (ang. *mention-pair* lub *pairwise*; patrz Aone i Bennett 1995), podejmujące decyzje klasyfikacyjne na podstawie cech par wzmianek, a następnie łączące je w klastry – klasy równoważności odpowiadające wspólnej referencji. Algorytm parowania został stopniowo wyparty

[6] Niektórzy badacze, jak np. Hirst (1981), uznają za pierwsze rozwiązanie tego typu jeszcze wcześniejszy system STUDENT (Bobrow 1964), zawierający pewną ograniczoną heurystykę dekodowania parafraz anaforycznych na potrzeby rozwiązywania zadań matematycznych; był to jednak jedynie dodatek do głównego celu pracy.

Tabela 2.1. Najważniejsze zagraniczne korpusy tekstów anotowane relacjami referencyjnymi – synteza

Korpus	**Język**	**Rozmiar**	**Anotacja składniowa**	**Centra fraz**	**Zakres i cechy charakterystyczne anotacji**	**Odnośnik bibliograficzny**
OntoNotes 5.0	angielski chiński arabski	1,6 mln słów 950 tys. słów 300 tys. słów	tak	skł.	podmioty domyślne, zagnieżdżenia fraz, apozycje	(Pradhan i in. 2007, 2012)
NAIST	japoński	38 tys. zdań	nie	nie	podmioty domyślne, struktura predykatów	(Iida i in. 2007)
ARRAU	angielski	94 tys. słów	nie	nie	konstrukcje atrybutywne, deiksa, struktura predykatów, relacje asocjacyjne	(Poesio i Artstein 2008)
COREA	holenderski	200 tys. słów	nie	skł.	anafora związana, relacje asocjacyjne, predykaty, deiksa, częściowa anotacja zagnieżdżeń i konstrukcji atrybutywnych	(Hendrickx i in. 2008)
AnCora-CO	hiszpański kataloński	400 tys. słów	tak	nie	predykaty, deiksa, podmioty domyślne, zagnieżdżenia i nieciągłości fraz	(Recasens 2010)
Copenhagen Dependency Treebank	duński angielski niemiecki włoski hiszpański	5 x 100 tys. słów	tak	nie	deiksa, relacje asocjacyjne	(Korzen i Buch-Kromann 2011)

Tabela 2.1. Najważniejsze zagraniczne korpusy tekstów anotowane relacjami referencyjnymi – synteza cd.

Korpus	Język	Rozmiar	Anotacja składniowa	Centra fraz	Zakres i cechy charakterystyczne anotacji	Odnośnik bibliograficzny
DIRNDL	niemiecki	3 tys. słów	nie	nie	koreferencja, relacje asocjacyjne	(Eckart i in. 2012, Björkelund i in. 2014)
ISNotes	angielski	11 tys. wzmianek	tak	skł.	stan informacyjny, relacje asocjacyjne	(Hou i in. 2013)
ANCOR	francuski	453 tys. słów	nie	nie	frazy zagnieżdżone i nieciągłe, relacje asocjacyjne	(Muzerelle i in. 2013)
GECCo	angielski niemiecki	365 tys. segmentów 372 tys. segmentów	tak	skł.	frazy zagnieżdżone i nieciągłe, predykaty, deiksa, relacje porównawcze, referencja do zdarzeń opisywanych dłuższymi fragmentami tekstu	(Lapshinova-Koltunski i Kunz 2014)
PCC 2.0	niemiecki	44 tys. słów	tak	nie	frazy nominalne zagnieżdżone i nieciągłe	(Stede i Neumann 2014)
PDT 3.0	czeski	58 tys. zdań	tak	skł./sem.	predykaty, deiksa, relacje asocjacyjne; anotacja na bazie struktury składniowej	(Zikánová i in. 2015)
PDC 1.0	angielski	20 tys. segmentów	nie	skł.	frazy zagnieżdżone, pleonastyczne ’it’, atrybuty	(Chamberlain i in. 2016)

Tabela 2.2. Polskie korpusy tekstów anotowane relacjami referencyjnymi – synteza

Korpus	**Rozmiar**	**Anotacja składniowa**	**Centra fraz**	**Zakres i cechy charakterystyczne anotacji**	**Odnośnik bibliograficzny**
NN	1103 relacje anaforyczne	nie	nie	anafora pronominalna, egzofora	(Filak 2006)
LUNA	2051 relacji	tak	sem.	linki anaforyczne do pojęć z ontologii komunikacji publicznej	(Marciniak 2010)
KPWr	45 tys. linków	tak	skł.	koreferencja z udziałem nazw własnych, także z uwzględnieniem podmiotów domyślnych	(Broda i in. 2012b)
PCC	21 865 klastrów	nie	sem.	podmioty domyślne, frazy zagnieżdżone i nieciągłe, relacje asocjacyjne, relacje pomocnicze	niniejsza publikacja

przez całościowy algorytm klastrowania (ang. *entity-based* lub *entity-mention*; patrz Luo i in. 2004) o znacznie większej sile wyrazu, porównujący cechy analizowanej wzmianki z cechami wcześniej utworzonych klastrów, co w dużym stopniu rozszerzyło zakres informacji dostępnej dla algorytmu klasyfikacyjnego.

Pod koniec lat 90. popularność zaczęły zyskiwać algorytmy maszynowego uczenia i także w dziedzinie koreferencji rozwiązania tego typu wygrywały konkurencję z narzędziami regułowymi (Connolly i in. 1994, McCarthy i Lehnert 1995, Kehler 1997, Soon i in. 1999, 2001, Ng i Cardie 2002, Rahman i Ng 2009), dominując przez kolejne 10 lat i stopniowo oferując coraz lepsze wyniki na dostępnym zbiorze testowym. Jednocześnie zaczęto zauważać braki dostępnej metody ewaluacyjnej, co doprowadziło do powstania dwóch nowych metryk: B^3 (Bagga i Baldwin 1998) oraz CEAF (Luo 2005), wykorzystanych w kolejnych konkursach ewaluacyjnych

CoNLL 2011 (z zadaniami dla języka angielskiego; patrz Pradhan i in. 2011) i 2012 (dodatkowo dla chińskiego i arabskiego; patrz Pradhan i in. 2012).

Równolegle do prac wykorzystujących techniki nadzorowanego uczenia maszynowego, na początku XXI w. niektórzy badacze zaczęli sugerować powrót do metod regułowych w celu poprawy jakości obsługi prostych przypadków zgodności wzmianek (Stuckardt 2001, Zhou i Su 2004, Mitkov i in. 2007). Intuicja ta została potwierdzona przez Haghighiego i Kleina (2009), których system do wykrywania koreferencji, oparty na regułach składniowo-semantycznych, okazał się lepszy od systemów statystycznych. Na podobnych zasadach skonstruowano system stanfordzki (Raghunathan i in. 2010, Lee i in. 2011), triumfujący w konkursie w 2011 r. i podejmujący decyzje klasyfikacyjne za pomocą uruchamianego sekwencyjnie zestawu modułów (tzw. sit, ang. *sieves*) wykorzystujących ręcznie tworzone reguły o zmniejszającej się precyzji.

Kolejnym naturalnym krokiem było stworzenie hybrydowego połączenia architektury sit z uczeniem maszynowym (Denis i Baldridge 2008, Chen i Ng 2012, Ratinov i Roth 2012) oraz jej adaptacja do innych języków, skutkująca znaczącą poprawą wyników (np. Krug i in. 2015). Obecnie najlepsze wyniki dla angielskiego osiągają algorytmy wykorzystujące sieci neuronowe (Lee i in. 2017, Zhang i in. 2018) – 69,20% miary F_1 na danych zadania CoNLL (Pradhan i in. 2012).

W dziedzinie dekodowania relacji asocjacyjnych pierwsze heurystyki zaproponowali Hahn i in. (1996) – reguły na konkretnym zbiorze gatunkowym – oraz Vieira i Teufel (1997), a także Poesio i in. (1997) – na bazie relacji WordNetu. Te ostatnie badania rozwinęli później Fan i in. (2005) oraz Roesiger i Teufel (2014), którzy zastosowali głębsze przeszukiwanie ścieżek, dzięki czemu uzyskali lepszą kompletność wyników.

Schulte im Walde (1998) zastosowała algorytm klastrowania słów w przestrzeni wielowymiarowej na danych korpusu BNC (British National Corpus), ale osiągnęła niewielką dokładność (22,7%). Poesio i in. (2002) oraz Markert i in. (2003), ograniczając się do relacji meronimicznych, zastosowali metodę poszukiwania wzorców składniowych – pierwsi w korpusie BNC (`the NP of NP`, `NP of NP`, `NP's NP`), drudzy w internecie (`floors of (the OR all) * apartments`). Wyszukiwarka internetowa była też używana w podobnym celu przez Bunescu (2003) oraz przez Poesio i in. (2004) do wykrywania nowych relacji meronimicznych. Na bazie tych metod powstał też system Lassalle'a i Denisa (2011) dla języka francuskiego, uzupełniony o relacje meronimiczne wyekstrahowane z tekstów metodą konstrukcji wzorców składniowych na bazie znanych wyrażeń. Jego dokładność, zmierzona

na danych korpusu DEDE (Gardent i in. 2005), dała wynik rzędu 23%. Wzorce składniowe tworzone na bazie leksykonu stosowali też Sasano i Kurohashi (2009).

Rahman i Ng (2012) potraktowali zadanie jako problem klasyfikacji stanu informacyjnego wzmianek. Po regułowym oznaczeniu dialogowego korpusu treningowego Nissim i in. (2004) użyli algorytmu SVM w zadaniu klasyfikacyjnym, dzięki czemu uzyskali wyniki rzędu 63,3–87,2% miary F dla czterech rodzajów relacji asocjacyjnych (całość – część, sytuacja, zdarzenie, zbiór – element). Cahill i Riester (2012) wykorzystali własności składniowe i semantyczne do wytrenowania modelu CRF na korpusie DIRNDL (Riester i in. 2010) dla kilku klas informacyjnych, w tym klasy asocjacyjnej.

Pierwszą w pełni automatyczną próbę identyfikacji nawiązań w języku polskim opisują Mitkov i Styś (1997) oraz Mitkov i in. (1998). Działanie algorytmu wykorzystującego wyłącznie informację gramatyczną (ręczne oznaczenia części mowy) testowane jest jednocześnie na trzech językach – oprócz polskiego także na angielskim i arabskim. Poprzedniki anaforyczne poszukiwane są w odległości co najwyżej dwóch zdań od wzmianki; wybór kandydata odbywa się przez ocenę wagi będącej kombinacją następujących czynników:

– określoność – sygnalizowana szykiem wyrazów, obecnością zaimków wskazujących lub powtórzeń;

– pierwszeństwo – czy fraza jest pierwszą w zdaniu;

– słownikowość – czy termin znajduje się w słowniku dziedzinowym;

– istotność czasownika – czy czasownik poprzedzający kandydata znajduje się na liście czasowników istotnych, takich jak *dyskutować*, *przedstawiać*, *ilustrować*;

– istotność frazy rzeczownikowej – czy fraza poprzedzająca kandydata znajduje się na liście znaczników strukturalnych (takich jak *rozdział*, *sekcja*, *tabela*);

– powtórzenie leksykalne – czy kandydat jest wymieniony więcej niż raz w danym akapicie;

– nagłówkowość – czy fraza nominalna jest częścią nagłówka sekcji;

– kolokacyjność – czy kandydat ma identyczną charakterystykę kolokacyjną co badany zaimek;

– odległość referencyjna – odległość (w zdaniach składowych) między kandydatem a badanym zaimkiem;

– nieprzyimkowość – preferencja dla kandydatów niebędących częścią frazy przyimkowej;

– bezpośrednia bliskość odniesienia – w zdaniach złożonych preferencja dla pierwszej frazy następującej po pierwszym czasowniku (dla zaimków występujących w kolejnym zdaniu także bezpośrednio po czasowniku).

W przypadku identycznych wyników obliczano je ponownie, podnosząc wagi niektórych parametrów, takich jak na przykład powtórzenie leksykalne, a następnie bliskość poprzednika.

Ewaluacja rozwiązania na korpusie 180 zaimków wykazała skuteczność na poziomie 93,3%, z następującymi wskaźnikami udziału poszczególnych parametrów: określoność – uwzględniana w 97,2% przypadków, odległość referencyjna – 94,4%, pierwszeństwo – 61,1%, nieprzyimkowość – 52,8%, istotność czasownika – 16,7%, powtórzenie leksykalne – 13,9%, bezpośrednia bliskość odniesienia – 2,8%.

Inne narzędzie do wykrywania pronominalnych relacji anaforycznych powstało w ramach prac nad parserem POLSYN (Kulików i in. 2004). Proces identyfikacji nawiązań opisany przez Ciurę i in. (2004) jest dwuetapowy: po zastąpieniu podmiotów domyślnych odpowiadającymi im zaimkami następuje poszukiwanie poprzedników wśród grup nominalnych o zgodnym rodzaju znajdujących się w odległości co najwyżej dwóch zdań od następnika – zaimka lub uprzednio wykrytych fraz (dzięki czemu w tekście mogą powstawać łańcuchy fraz o wspólnej referencji). W przypadku równoważności kandydatów brana jest pod uwagę ich funkcja w zdaniu, z preferencją podmiotu nad dopełnieniem.

Inną próbę częściowej identyfikacji nawiązań w polszczyźnie przeprowadzili Abramowicz i in. (2006) na potrzeby wykrywania obiektów w tekście. Algorytm bada podobieństwo napisów, używając miary Jaro-Winklera (Winkler 1999), i jest przez samych autorów określany jako nieobejmujący wszystkich przypadków.

Filak (2006) opisuje prace nad implementacją detektora anafory zaimkowej dla systemu GATE (Cunningham i in. 2002) stworzonego metodami uczenia maszynowego z wykorzystaniem drzewa decyzyjnego J48 z biblioteki WEKA (Hall i in. 2009) na bazie dostępnego minikorpusu. Algorytm wykorzystuje 17 dobranych ręcznie parametrów par wzmianek – od tradycyjnej zgodności powierzchniowej przez odległość porównywanych wzmianek w zdaniach i słowach, po zgodność rodzaju i liczby oraz uwzględnienie akcentowości zaimków[7]. Ewaluacja rozwiązania

[7]Wyjaśnienie kategorii gramatycznych użytych w bazowym dla algorytmu korpusie IPI PAN zawiera publikacja opisująca ten korpus (Przepiórkowski 2004: rozdziały 3.3–3.4).

wykazała jego precyzję na poziomie 50,7%, a kompletność na poziomie 53,5%. Niski wynik autorzy tłumaczą błędami w ręcznej anotacji przykładów, błędami ujednoznaczniania morfoskładniowego oraz niewielką ilością danych uczących.

IKAR (Broda i in. 2012b) jest narzędziem łączącym metody regułowe i uczenia maszynowego. Wzmianki (nazwy własne, zaimki, frazy rzeczownikowe ograniczone do głowy składniowej) łączone są wyłącznie z nazwami własnymi; do wykrywania powiązań między parami nazw własnych wykorzystywany jest algorytm uczenia maszynowego, w pozostałych przypadkach zaś – reguły heurystyczne. Klasyfikacja wykonywana jest sekwencyjnie, podobnie do algorytmu stanfordzkiego. Dla par nazw własnych stosowany jest klasyfikator drzewa decyzyjnego C4.5 trenowany na bazie cech uwzględniających część wspólną lematów słów tworzących wzmianki, różnicy w liczbie słów, zgodności rodzaj–liczba. Algorytmy regułowe zawierają ciekawą regułę SEMANTICLINK wykorzystującą podobieństwo semantyczne głów wzmianek na bazie Słowosieci (Piasecki i in. 2009).

Ogrodniczuk (2013) opisuje natomiast problem identyfikacji relacji koreferencyjnych metodą tłumaczenia i projekcji: tekst jest tłumaczony na język, dla którego istnieją dobre jakościowo narzędzia do wykrywania nawiązań (tu: angielski), uruchamiane jest narzędzie obcojęzyczne – Stanford CoreNLP (Lee i in. 2013), a następnie jego wyniki są przenoszone na wersję polską z wykorzystaniem dokonanego podczas tłumaczenia zrównoleglenia obu wersji na poziomie słów oraz dopasowaniem wzmianek, wykrytych osobno za pomocą polskiego detektora regułowego. Ewaluacja rozwiązania na bazie 260 tekstów (ok. 75 tys. słów) anotowanych ręcznie wykazała precyzję rozwiązania na poziomie 74,9%, a jego kompletność na poziomie 67,81%.

Obecnie w dwóch ośrodkach badawczych w Polsce prowadzone są prace nad szczegółowymi problemami z dziedziny wykrywania relacji referencyjnych; na Politechnice Wrocławskiej nad wykrywaniem podmiotów domyślnych powiązanych z nazwami własnymi (Kaczmarek i Marcińczuk 2015b, 2017), w Zespole Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN zaś nad wykorzystaniem relacji koreferencyjnych w streszczaniu tekstów prasowych (Kopeć 2018).

2.6. Metody ewaluacji

Ocena jakości działania systemu wykrywającego wzmianki i klastry koreferencyjne wymaga użycia formalnej metody obliczania miary dopasowania wyniku systemu

(nazywanego w skrócie **konfiguracją SYS** lub **odpowiedzią**, ang. *response*) do wyniku idealnego (nazywanego **konfiguracją GOLD** lub **kluczem**, ang. *key*) w możliwie wiarygodny i interpretowalny sposób. Strategii dokonywania takich obliczeń może być wiele: wykrywanie wzmianek może na przykład uwzględniać dopasowanie samych centrów semantycznych, dokładnych granic wzmianek albo stopnia tekstowego pokrycia klucza przez wynik systemu, wykrywanie klastrów zaś – różne sposoby oceny systemów ze względu na charakter błędów popełnianych przez proces klastrowania (czy na przykład wykrycie singletonu premiowane jest w taki sam sposób jak klastra wieloelementowego). Zgodnie z klasyfikacyjnym charakterem zadania wszystkie metody podają natomiast wartości **kompletności** (ang. *recall*, oznaczanej dalej jako *R*), oceniającej, jaki stopień wszystkich poprawnych wyników został wykryty przez system i **precyzji** (ang. *precision*, dalej *P*), informującej, ile z wykrytych wyników było poprawnych; ostateczną ocenę systemu stanowi ich średnia harmoniczna nazywana **miarą $\mathbf{F_1}$** (ang. *F-measure*, *F-score*).

W związku z tym, że metody nagradzające częściowe dopasowanie wzmianek użyte m.in. w zadaniu *Anaphora Resolution Exercise*[8] (Orăsan i in. 2008) czy *Coreference Resolution in Multiple Languages* na warsztacie *SemEval 2010*[9] (Recasens i Hovy 2011, Màrquez i in. 2012) nie są już stosowane w praktyce, dalej ograniczam się do zaprezentowania najważniejszych metod ewaluacji jakości systemów wykrywających klastry koreferencyjne, wypracowanych przez międzynarodowe środowisko naukowe i użytych do oceny jakości systemów prezentowanych w ramach niniejszej pracy.

Przez lata powstało kilkanaście miar ewaluacji procesu klastrowania koreferencyjnego, z których w praktyce obecnie używane są cztery: MUC (Vilain i in. 1995), B^3 (Bagga i Baldwin 1998) i CEAFE (Luo 2005) oraz ich średnia arytmetyczna, zastosowana po raz pierwszy na konferencji *Computational Natural Language Learning* i z tego względu nazywana miarą CoNLL[10]. Od wielu lat w środowisku podnoszone są zarzuty w stosunku do używanych obecnie metryk, na przykład Holen (2013) porównując wyniki ewaluacji przez człowieka z wynikami uzyskiwanymi automatycznie, uzależnia sukces procedury identyfikacyjnej od poprawnego klastrowania wzmianek o znaczącej wadze informacyjnej, Moosavi i Strube (2016) zauważają zaś brak korelacji między poszczególnymi miarami „standardowymi". Mimo to popularność miary CoNLL tłumaczy się jej charakterystyką równoważącą

[8] `http://rgcl.wlv.ac.uk/events/ARE/`.

[9] `http://stel.ub.edu/semeval2010-coref/`.

[10] Niekiedy używana jest też nazwa MELA od ang. *mentions, entities, links* (np. Pradhan i in. 2012).

najważniejsze cechy koreferencji, gdyż składowa miara B^3 koncentruje się głównie na własnościach wzmianek, MUC – linków, a CEAFE – klastrów. Proponowane nowe metryki ewaluacyjne: BLANC (BiLateral Assessment of Noun-phrase Coreference; patrz Recasens i Hovy 2011, Luo i in. 2014) czy LEA (Link-based Entity Aware evaluation metric; patrz Moosavi i Strube 2016) nie są jednak szeroko stosowane i zostały wyparte przez miarę CoNLL. Dalej opisuję krótko zasady obliczania trzech najważniejszych metryk oraz ich główne własności; ich dokładną prezentację wraz z przykładami obliczeń zawiera rozdział 14 monografii angielskojęzycznej (Ogrodniczuk i in. 2015).

2.6.1. Miara MUC

Metryka MUC (Vilain i in. 1995) była pierwszą próbą ewaluacji narzędzi do wykrywania koreferencji użytą w zadaniach MUC–6 i MUC–7[11]. Zasadą jej działania jest ocena liczby poprawnie wykrytych linków definiowanych w sposób „anaforyczny" jako powiązanie danej wzmianki z jej linearnym poprzednikiem z tego samego klastra. Kompletność obliczana jest jako stosunek poprawnie wykrytych linków do wszystkich poprawnych linków. Formalnie obliczana jest następująco:

$$R = \frac{\sum_{i=1}^{|GOLD|} (|GOLD_i| - |p(GOLD_i)|)}{\sum_{i=1}^{|GOLD|} (|GOLD_i| - 1)}$$

gdzie:

- $|GOLD|$ – jest liczbą klastrów koreferencyjnych w kluczu,
- $|GOLD_i|$ – jest liczbą wzmianek w klastrze $GOLD_i$ (a zatem $|GOLD_i| - 1$ odpowiada „rozpiętości klastra", czyli minimalnej liczbie linków pomiędzy wzmiankami zapewniających jego pełne pokrycie),
- $p(GOLD_i)$ – jest zbiorem wszystkich klastrów ze zbioru SYS, które zawierają co najmniej jeden element z $GOLD_i$ (czyli liczbą klastrów, na jakie „rozpadł się" klaster ze zbioru GOLD).

[11] Message Understanding Conference z lat 1995 (`http://www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html`) i 1997 (`http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`).

Liczba poprawnie wykrytych linków odpowiada zatem (dla każdego klastra) różnicy wielkości klastra (który stanowi ciąg wzmianek połączonych linkami) i liczby klastrów, na jakie „rozpadł się" on w zbiorze SYS. Wszystkie poprawne linki zlicza się, sumując „rozpiętości" wszystkich klastrów w zbiorze GOLD.

Precyzja zdefiniowana jest natomiast w naturalny sposób jako stosunek liczby poprawnie wykrytych linków do liczby wszystkich wykrytych linków, z wartościami zmiennych zdefiniowanymi odpowiednio:

$$P = \frac{\sum_{i=1}^{|SYS|} (|SYS_i| - |p(SYS_i)|)}{\sum_{i=1}^{|SYS|} (|SYS_i| - 1)}$$

Zaletą miary MUC jest jej prostota; podstawowy problem stanowi natomiast koncentracja miary wyłącznie na linkach tworzących klastry – bez uwzględniania singletonów. Konsekwencją jest zaburzona wartość wyniku w sytuacji, gdy zbiór wzmianek, na których działa system, jest inny niż w testowym zbiorze GOLD. Oznacza to nieprzydatność miary w zastosowaniach praktycznych z automatyczną ekstrakcją wzmianek – system wykrywający nadmierną liczbę wzmianek osiągnie taki sam wynik jak system wykrywający zbiór właściwy. Drugi zarzut dotyczy „książkowego" podejścia do precyzji i dokładności, przez co zwiększona liczba błędnie wykrytych linków (czyli scalenia przez system klastrów w istocie niekoreferencyjnych, co jest poważnym błędem) zupełnie nie wpływa na dokładność – w szczególności połączenie wszystkich wzmianek w jeden klaster da w wyniku stuprocentową kompletność przy niezerowej precyzji. Podobnie nie ma różnicy między połączeniem dużego klastra z mniejszym w stosunku do połączenia go z większym – oba nadmiarowe linki są oceniane tak samo, a przecież druga sytuacja wydaje się bardziej szkodliwa, bo błędnie oznacza jako koreferencyjne większą liczbę wzmianek.

2.6.2. Miara B^3

Miara B^3 (Bagga i Baldwin 1998) definiuje kompletność i precyzję, uśredniając wyniki dla pojedynczych wzmianek:

$$R = \sum_{i=1}^{N} \frac{1}{N} * \frac{|SYS(i) \cap GOLD(i)|}{|GOLD(i)|}$$

$$P = \sum_{i=1}^{N} \frac{1}{N} * \frac{|SYS(i) \cap GOLD(i)|}{|SYS(i)|}$$

gdzie:

- N – jest liczbą wzmianek w zbiorach GOLD i SYS[12],
- $GOLD(i)$ – jest klastrem (rozumianym jako zbiór wzmianek) w zbiorze GOLD zawierającym wzmiankę i,
- $SYS(i)$ – jest klastrem w zbiorze SYS zawierającym wzmiankę i.

Miara B^3 uwzględnia wielkość łączonych klastrów i singletony, natomiast jej działanie jest nieintuicyjne: definicja kompletności sprawia, że system łączący wszystkie wzmianki w jeden klaster da w wyniku stuprocentową kompletność, wzrost liczby singletonów powoduje zaś niebezpieczny wzrost precyzji, co utrudnia porównanie systemów w zastosowaniach praktycznych, gdyż liczba singletonów w rzeczywistych tekstach jest duża[13]. Co więcej, miara B^3 zakłada, że ewaluowany system działa na wzmiankach ze zbioru GOLD. Warianty miary dostosowujące ją do sytuacji wzmianek wykrywanych systemowo zaproponowali m.in. Bengtson i Roth (2008), Stoyanov i in. (2009), Rahman i Ng (2009) oraz Cai i Strube (2010).

2.6.3. Miara CEAF

Zasadą działania miary CEAF (Luo 2005) jest mapowanie klastrów ze zbiorów GOLD i SYS z wykorzystaniem pewnej miary podobieństwa, zdefiniowanej w dwóch wariantach: dla wzmianek (ang. *mention-based*) i klastrów (ang. *entity-based*), czego konsekwencją jest istnienie dwóch wariantów miary, odpowiednio CEAF-M i CEAF-E, którym odpowiadają następujące funkcje podobieństwa:

$$\phi_M(GOLD_i, SYS_j) = |GOLD_i \cap SYS_j|$$

$$\phi_E(GOLD_i, SYS_j) = \frac{2 * |GOLD_i \cap SYS_j|}{|GOLD_i| + |SYS_j|}$$

[12]Miara zakłada działanie algorytmu klastrującego na zbiorze wzmianek z klucza; jej wariant dla wzmianek wykrywanych automatycznie przedstawił Stoyanov (2009).

[13]86% wszystkich klastrów w korpusie AnCora-CO, 61% w korpusie ACE (Recasens 2010: rozdział 5.3.1) i, uprzedzając wywód, niespełna 83% w naszym korpusie (patrz tabela 4.11).

gdzie:

– $GOLD_i$ – jest klastrem ze zbioru GOLD,

– SYS_j – jest klastrem ze zbioru SYS.

Kompletność i precyzja zdefiniowane są jako:

$$R = \frac{\Phi(h^*)}{\sum_{i=1}^{|GOLD|} \phi(GOLD_i, GOLD_i)}$$

$$P = \frac{\Phi(h^*)}{\sum_{i=1}^{|SYS|} \phi(SYS_i, SYS_i)}$$

gdzie: $\Phi(h)$ – jest mapowaniem o najlepszym podobieństwie.

Wadą miary CEAF jest nadmierny udział singletonów w wyniku końcowym, co wypacza wynik ze względu na dużą liczbę singletonów w danych rzeczywistych, oraz równe traktowanie klastrów bez względu na ich rozmiar, co sprawia, że błędne połączenie dwóch małych klastrów ma identyczną wagę co przyłączenie małego klastra do dużego. Problemem jest też intensywność obliczeniowa rzędu $O(m^3 \log m)$, gdzie m jest liczbą wzmianek. Teksty Rahmana i Nga (2009) oraz Cai i Strubego (2010) zawierają propozycje wariantów miary CEAF.

2.6.4. Miara BLANC

Miara BLANC (Recasens 2010) bierze pod uwagę koreferencyjność wszystkich par wzmianek i oblicza kompletność oraz precyzję osobno dla par poprawnie zaklasyfikowanych jako koreferencyjne i niekoreferencyjne, a ostatecznym wynikiem jest średnia arytmetyczna tych wartości.

Zatem:

$$P_c = \frac{rc}{rc + wc} \qquad P_n = \frac{rn}{rn + wn} \qquad P = \frac{P_c + P_n}{2}$$

$$R_c = \frac{rc}{rc + wn} \qquad R_n = \frac{rn}{rn + wc} \qquad R = \frac{R_c + R_n}{2}$$

$$F_c = \frac{2P_cR_c}{P_c + R_c} \qquad F_n = \frac{2P_nR_n}{P_n + R_n} \qquad F = \frac{F_c + F_n}{2}$$

gdzie:

- rc (ang. *rightly coreferent*) – jest liczbą par wzmianek oznaczonych jako koreferencyjne zarówno w SYS, jak i w GOLD,
- wc (ang. *wrongly coreferent*) – jest liczbą par wzmianek oznaczonych jako koreferencyjne w SYS, ale niekoreferencyjne w GOLD,
- wn (ang. *wrongly non-coreferent*) – jest liczbą par wzmianek oznaczonych jako niekoreferencyjne w SYS, ale koreferencyjne w GOLD,
- rn (ang. *rightly non-coreferent*) – jest liczbą par wzmianek oznaczonych jako niekoreferencyjne zarówno w SYS, jak i w GOLD.

Gdy w trakcie obliczeń mogłoby zajść dzielenie przez zero (może się to zdarzyć, na przykład gdy system zwróci w wyniku wyłącznie singletony, co skutkuje zerową wartością wyrażenia $rc+wc$), wynik takiego obliczenia zostaje arbitralnie ustalony na zero.

Zaletą miary jest jej dużo większa zgodność z intuicją, gdy mamy do czynienia z dużą liczbą singletonów. Zasadniczą wadą miary jest kwadratowy wzrost liczby linków w stosunku do liczby wystąpień, co oznacza zależność wyników od długości tekstu. W konsekwencji system popełniający regularne błędy (np. sklejający czwórki, a nie pary wystąpień, co odpowiada zaburzeniu na poziomie jakiejś własności lingwistycznej) uzyskuje różne wyniki kompletności na dokumentach o różnej długości (w przeciwieństwie do starszych miar – MUC i B^3).

3.

Model relacji referencyjnych

Podstawowe rozróżnienia terminologiczne zjawisk referencji, koreferencji, anafory i asocjacji oraz zasady ich interpretacji przyjęte w opracowanym modelu zostały już w skrócie opisane w rozdziale wprowadzającym; dalej omawiam szczegółowo propozycję całościowej typologii relacji referencyjnych, obejmującej zarówno referencję właściwą, jak i pośrednią, ich aspekty wpływające na odbiór relacji oraz zależności dodatkowe, wspomagające dekodowanie wyrażeń referencyjnych. Proponowane podejście łączy elementy zaprezentowanych wcześniej ujęć teoretycznych, stanowiąc jednocześnie autorski sposób opisu relacji referencyjnych – uproszczony w stosunku do konkretnych modeli językoznawczych, mający jednak na celu skuteczne przetwarzanie informacji referencyjnej metodami komputerowymi. Prace te wymagają określenia, jakim wyrażeniom przypisuje się własność posiadania referencji, w jaki sposób wyznaczać ich tekstowe granice oraz jakie zjawiska lingwistyczne biorą udział w procesie dekodowania referencji i wpływają na jej interpretację (np. w zakresie konkretności i określoności odwołań).

3.1. Świat tekstu i własność referencji

W niniejszej pracy przyjmuję stanowisko Langackera (2008) i Kunz (2010: rozdział 2.1) odnośnie istnienia mentalnego świata tekstu, który nie musi odpowiadać światu rzeczywistemu oraz może zawierać fakty hipotetyczne czy idee i pojęcia abstrakcyjne. Podobnie do Vatera (2009) twierdzę, że referencję posiadają także: nazwy stanów, sytuacji, miejsc i określenia czasu, reprezentowane jako jednostki tekstowe. Takie rozumienie referencji jest wynikiem potraktowania jako punktu wyjścia dla moich rozważań wypowiedzi językowej, w której odniesienie przysługuje szerokiej klasie obiektów będących manifestacjami metafory konceptualnej w rozumieniu Lakoffa i Johnsona (1988). Za Topolińską (1976) uzależniam interpretację relacji referencyjnych od intencji komunikatywnej nadawcy wyrażonej w tekście.

Zgodnie z poczynionymi założeniami przypisuję własność referencji szerokiej klasie wzmianek, niezależnie od ich ewentualnych uwikłań frazeologicznych. Za

referencyjne uznaję na przykład części frazeologizmów, które w ogólnym przypadku mogą wchodzić w relację koreferencji, jak w przykładzie *Nie wahał się włożyć kija w mrowisko. Mrowisko to, czyli cały senat uniwersytecki, pozostawało zwykle niewzruszone.* Podobnie jako referencyjne traktuję wzmianki reprezentowane zaimkami nieokreślonymi (*ktoś*, *ktokolwiek*), przeczącymi (*nic*, *nikt*, *żaden*) i uniwersalnymi (*wszystko*, *wszyscy*) ze względu na często tworzone przez nie związki koreferencyjne, także z frazami określonymi, np. *Żaden z braci nie uczy się, żaden nie pracuje.*, *Pewien chłopiec oblał egzamin maturalny. Zdecydował on, że zrezygnuje z dalszych studiów.* (Bellert 1971). Jako niereferencyjne traktuję natomiast konstrukcje czysto składniowe, takie jak np. zaimki pytajne i względne (*Kto był pierwszym królem Polski? Bolesław Chrobry.*).

Ponieważ uznaję, że analizowany tekst wyznacza granice świata, który opisuje, nie wyróżniam w żaden specjalny sposób relacji egzoforycznych, czyli budujących odwołania za pomocą środków pozalingwistycznych (np. gdy uczestnik dialogu wskazuje ręką *tamto krzesło*). Podobnie postępuję w przypadku relacji homoforycznych, których podstawą jest fraza ogólna nabywająca znaczenia w danym kontekście sytuacyjnym (np. odwołanie do „królowej" może mieć różne znaczenie w zależności od kraju, w którym jest wypowiadane). Zakładam, że sam tekst określa ów kontekst w sposób wystarczający (tzn. jeśli użyte jest wyłącznie ogólne określenie „królowej", tekst traktuje o pewnej niedoprecyzowanej królowej ze świata tekstu).

3.2. Typy i granice wzmianek

Zgodnie z przyjętym ograniczeniem interesują mnie relacje referencyjne z komponentem nominalnym, których podstawowym nośnikiem tekstowym są uogólnione grupy nominalne. Ograniczenie to przekłada się w prosty sposób na konstrukcje lingwistyczne, które mogą zostać użyte do reprezentacji i dekodowania relacji referencyjnych.

Oprócz rzeczownika z podrzędnikami wzmiankę może zatem stanowić: zaimek osobowy, forma zaimka *siebie*, ew. zmodyfikowana formą leksemu *sam* (np. *mnie samego*, *siebie samą*), zaimek wskazujący wprowadzający zdanie podrzędne inne niż względne (*Mówiono o tym, że grzmi i pada.*), zaimek dzierżawczy (*Anna powiedziała, że ten płaszcz jest jej.*), podmiot domyślny, a także gerundium, wykazujące semantyczne podobieństwo do frazy zdaniowej (*Pojawienie się niedźwiedzia było*

niesamowite, ale bardzo go to wystraszyło.) Wzmianki nominalne mogą być także realizowane za pomocą elipsy (*Czytałeś książki Lema? Czytałem Ø.*, *Przeszył ją głęboki lęk. Znów ten głuchy, prawdziwy Ø.*) W związku z tym, że zagadnienie elipsy jest jednak w niniejszym opisie traktowane jako podlegające jedynie wstępnemu rozpoznaniu, konstrukcje eliptyczne ograniczam do przypadków jawnie wskazujących na brak wzmianki rzeczownikowej.

W związku z tym, że komponent nominalny może uczestniczyć w rozważanych dalej relacjach tylko po jednej stronie, opisuję także wzmianki nienominalne, powiązane z nieeliptycznym wskaźnikiem nominalnym (np. *Adam ugotował obiad wczoraj, a Ewa zrobi to jutro.*, *Zapadł mrok. Powiększyło to nasze szanse na ucieczkę.*)

Semantyczny charakter reprezentacji obiektów podlegających referencji sprawia, że opisuję głęboką strukturę składniową wzmianek – oprócz frazy nadrzędnej w ich granicach umieszczam także wszystkie frazy podrzędne składniowo. Frazy o różnych centrach semantycznych (czyli w sposób jawny odnoszące się do różnych obiektów) odpowiadają osobnym wzmiankom, a zatem wewnątrz frazy *dyrektor departamentu firmy* wyróżniam trzy wzmianki odnoszące się kolejno do *dyrektora departamentu firmy*, *departamentu firmy* i samej *firmy*. Powyższa zasada odnosi się także do konstrukcji z koordynacją – osobną referencję ma fraza współrzędna oraz jej składniki: *Asia i Basia mnie lubią. One są naprawdę ładne, szczególnie Aśka.* W przypadku wyliczeń wieloelementowych, ze względów praktycznych, oznaczam jako osobne wzmianki wyłącznie pojedyncze elementy wyliczenia i całe wyliczenie (bez licznych podzbiorów, chyba że zostały jawnie przywołane w dalszej części tekstu). Dopuszczam także istnienie wzmianek nieciągłych, np. *Czerwoną sobie kupiłam sukienkę.*, *Tylko takie książki kupuję, które mają dużo obrazków.*

Podrzędnikiem składowym centrum wzmianki może być fraza przymiotnikowa uzgadniająca formę (przypadek, liczbę, rodzaj) z nadrzędnym rzeczownikiem (np. *kolorowe kwiaty*, *zapaleńcy prowadzący swoje wojenki*), fraza przymiotnikowa w dopełniaczu l. poj. rodzaju nijakiego (*coś fantastycznego*), fraza rzeczownikowa w apozycji z nadrzędnikiem[1] (*malarz pejzażysta*), fraza rzeczownikowa w dopełniaczu (*kolega brata*), fraza liczebnikowa (*zabójca pięciu kobiet*), fraza zdaniowa (*dziewczyna, o której rozmawiamy*) czy fraza przyimkowo-nominalna (*ustawa o podatku dochodowym*).

Za pojedynczą wzmiankę uznaję również frazę z centrum w postaci skrótu, którego rozwinięciem jest rzeczownik (np. *ks. kanonik*, *prof. Balcerowicz*), frazę, której

[1] Wzmianki mogą być zatem bardzo rozbudowane, np. *Jan Kowalski, syn Juliusza, ojciec pięciorga dzieci, mąż Zuzanny, który zaginął miesiąc temu*.

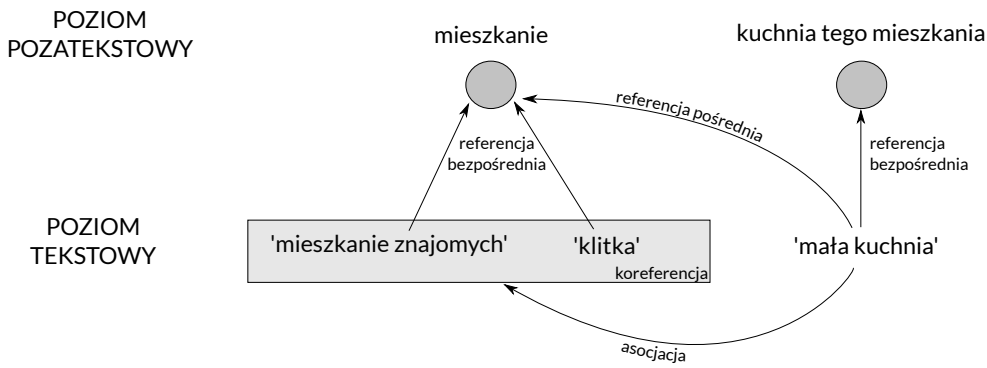
centrum składniowym jest liczebnik (*trzy rowery*), frazę z elipsą rzeczownikową, której centrum jest przymiotnik (*(zrób bukiet z tych czerwonych kwiatów i z) tych niebieskich*), frazę opisującą datę lub godzinę, której centrum jest rzeczownik w dopełniaczu lub fraza przymiotnikowa (*12 lipca br.*), frazę z korelatem (*coś, co może być opacznie zrozumiane*) oraz frazę skoordynowaną, także ze spójnikiem przecinkowym (*krzesło, stół i fotel*). W przypadkach wątpliwych dodatkowym kryterium wyznaczania granic wzmianki (oprócz składniowego) jest kryterium semantyczne: w treść wzmianki włączam wszystkie powiązane z nią frazy istotne dla jej interpretacji semantycznej.

3.3. Relacje tekstowe i pozatekstowe

Relacje referencyjne dzielę na potrzeby niniejszej pracy na dwa główne typy: **bezpośrednie** i **pośrednie**, przyjmując jako kryterium tego połączenia fakt wykraczania zjawiska referencji poza poziom tekstu. Istotą referencji bezpośredniej jest odwołanie do obiektu lub sytuacji pozatekstowej, realizowane za pomocą środków językowych – co oznacza, że określonym fragmentom tekstu przypisujemy pewne byty lub sytuacje ze świata tekstu. W przykładzie *Byliśmy wczoraj w mieszkaniu znajomych. Ale klitka. Mała kuchnia, mały pokoik, mała łazienka.* wzmianka *mieszkanie znajomych* przywołuje konkretny obiekt z mentalnego świata nadawcy. Istotą nawiązań pośrednich jest z kolei wskazanie referenta za pośrednictwem innego obiektu. W powyższym tekście wzmianka *mała kuchnia* odwołuje się do *mieszkania znajomych* za pośrednictwem mentalnej reprezentacji *kuchni mieszkania znajomych* (patrz rys. 3.1). Między wzmiankami o wspólnym odwołaniu zachodzi tekstowa relacja **koreferencji**; podobnie relacjom pośrednim przyporządkowuję tekstową relację **asocjacji**.

Domyślnie relacje referencyjne zachodzą w sposób pełny, czasem jednak tekst może zawierać informację o pewnym **aspekcie** relacji mogącym wpłynąć na zdekodowanie referencji. Może nim być na przykład przekonanie autora wypowiedzi o istnieniu zależności, która według innej interpretacji w ogóle nie zachodzi, czy fakt ograniczenia relacji w czasie.

Oprócz referencyjnych wyróżniam też relacje pomocnicze, mogące ułatwić proces rozstrzygania koreferencji. **Relacje wspierające** (przesłanki) wiążą wzmiankę z innym fragmentem tekstu wprowadzającym jakąś informację istotną do rozstrzygnięcia referencji, np. *Nutka szczeknęła i pobiegła w stronę pani. Ukochana suczka Ani wiedziała, że może liczyć na pieszczoty.*, *Ewa miała długie, czarne włosy, a Anka*



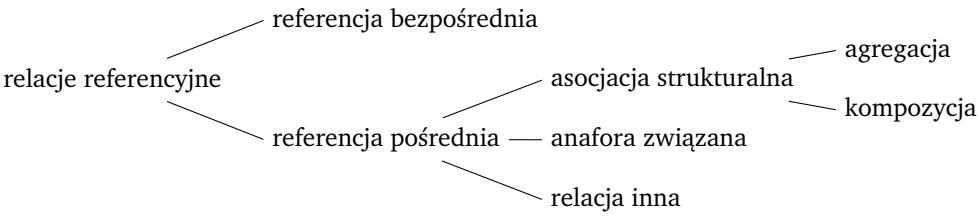
Rysunek 3.1. Zależności między relacjami referencyjnymi a tekstowymi

była blondynką. Kochałem obie, ale ożeniłem się z brunetką. **Relacje wykluczające** (negatywne) reprezentują natomiast informację wyłączającą wspólność referenta, np. *Janek nie jest ojcem Zuzi.*

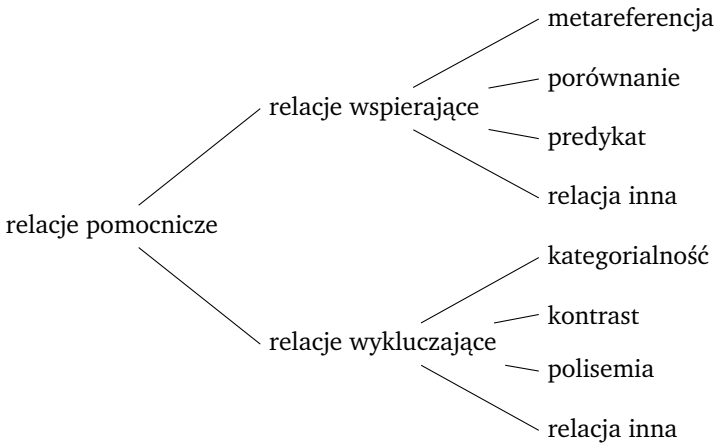
3.4. Typologia referencyjna

Łączny opis relacji referencyjnych i konstrukcji zależnych jest rozwinięciem koncepcji Kunz i in. (2016) oraz Lapshinovej-Koltunski i in. (2016), porównujących spójność tekstów wielu gatunków za pomocą analizy relacji identycznościowych i pośrednich. Rysunki 3.2–3.4 przedstawiają zaproponowaną przeze mnie wspólną klasyfikację relacji referencyjnych, pomocniczych i aspektów, użytą w dalszych pracach korpusowych i omówioną szczegółowo w dalszej części tego rozdziału. Ze względu na to, że niewystarczająca granularność relacji mogłaby sprawić, że odmienne zjawiska byłyby sklasyfikowane łącznie, ich zbytnie rozdrobnienie zaś – że przyporządkowanie do poszczególnych kategorii byłoby zbyt trudne, powstała taksonomia stara się znaleźć równowagę między zbytnią ogólnością a nadmierną szczegółowością opisu, zatem wyróżnia przede wszystkim relacje niekontrowersyjne i dobrze zdefiniowane.

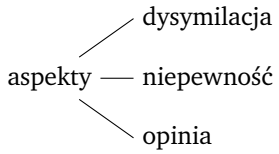
Największą różnicą w stosunku do licznych ujęć zagranicznych (np. Uryupina 2007, Poon i Domingos 2008, Haghighi i Klein 2009), na którą warto zwrócić osobną uwagę, jest potraktowanie relacji atrybutywnych jako niekoreferencyjnych z wyrażeniami, które określają. Same grupy nominalne użyte predykatywnie (niewskazujące na obiekt, ale na jego właściwości, np. *Tata Jana był architektem.*)



Rysunek 3.2. Szczegółowa typologia relacji referencyjnych



Rysunek 3.3. Relacje wspierające i wykluczające



Rysunek 3.4. Aspekty relacji referencyjnych

mają własności referencyjne, gdyż przedmiot ich odniesienia jest znany – jest nim pojęcie, stan lub właściwość (w tym przypadku fakt wykonywania zawodu architekta). Do sytuacji wyznaczonej w ten sposób można się zresztą wielokrotnie odwoływać, na przykład w dalszym zdaniu *Jan wie, że ten zawód jest bardzo*

wymagający. wzmianka *ten zawód* jest jawnie koreferencyjna ze wzmianką *architektem* ze zdania poprzedniego. Stan obiektu czy jego funkcja nie są natomiast tożsame z obiektem, toteż użycie atrybutywne (definicyjne) nie jest koreferencyjne z określeniem obiektu.

3.4.1. Koreferencja

Jak już wspominałem w rozdziale 1.1, relacja koreferencji może zostać zasygnalizowana w tekście na wiele sposobów. Niektóre z nich wykorzystują środki systemowe – leksykalne lub gramatyczne, inne wymagają natomiast wyjścia poza językowe mechanizmy interpretacji tekstu.

Systemowymi środkami językowymi, służącymi do ustanowienia koreferencji, są na przykład: powtórzenia tej samej frazy lub jej części (*Jan Kowalski – Kowalski*), elizja podmiotu, anafora czy katafora. Działanie pozasystemowe może natomiast wymagać zdekodowania relacji semantycznej w rodzaju synonimii (np. *W rozdziale drugim omawiany jest stan badań. Ta część książki przedstawia również plan całej pracy.*) czy hiperonimii – hiponimii (*Nasze owczarki dobrze pilnują domu. Te psy nigdy nie wpuszczą obcego za bramę.*), użycia wiedzy specjalistycznej lub encyklopedycznej (*Mecz Świtu Nowy Dwór Maz. z Polonią Warszawa zakończył się remisem 1:1. Pierwszą bramkę strzeliły Czarne Koszule.*), znajomości całego tekstu (*Kallina rozkazała mi cię karmić. Ligia rozkazała! Na to nie było odpowiedzi.*), zdekodowania metonimii tworzonej ad hoc (*– Który klient zamawiał kawę? – Kanapka z szynką.*[2]), analizy przeniesień słowotwórczych (*Do pokoju wszedł facet z głową jak balon. Balonowiec wyciągnął spluwę i wymierzył w szeryfa.*) czy zinterpretowania quasi-anafory (*Duszą towarzystwa był zięć Kowalskich. Młody prawnik właśnie wrócił ze Stanów.*). Działanie środków pozasystemowych zależy ściśle od kontekstu – np. tematu tekstu, dzięki czemu ten sam leksem o tym samym znaczeniu może mieć różną referencję (np. woda jako związek chemiczny lub napój).

3.4.2. Referencja pośrednia

Istotą referencji pośredniej jest wystąpienie powiązania pomiędzy obiektami przywoływanymi w tekście, które powoduje, że dana wzmianka odwołuje się nie tylko bezpośrednio do swojego referenta, ale także pośrednio do pewnego innego obiektu

[2] Por. (Fauconnier 1985: s. 144).

(w przykładzie z rys. 3.1 obiekty, do których odwołują się wzmianki *mała kuchnia* i *mieszkanie znajomych* pozostają ze sobą w relacji *całość – część*). Relację pośrednią rozumiem zatem jako zależność leksykalną „przeniesioną poza tekst", co pozwala na traktowanie jej jako relacji referencyjnej.

Zgodnie z typologiami zagranicznymi najmniej kontrowersyjną relacją pośrednią jest właśnie **asocjacja strukturalna** (relacja meronimiczna), realizowana jako agregacja lub kompozycja (powiązanie całości z częścią).

Relacja **agregacji** opisuje powiązanie zbioru z podzbiorem lub elementem (*Rząd pracuje nad zmianą przepisów podatkowych. Minister zapowiedział, że zmiany te będą obowiązywały od nowego roku.*, *Jeden z chłopców stłukł szybę, ale wszyscy zostali ukarani.*). Przyjmuję, że zbiory mogą być też definiowane za pomocą zaimków upowszechniających (*każdy*, *wszyscy*, *wszystko*, *zawsze*, *wszędzie*) albo frazy kwantyfikowanej (*niektórzy*, *większość*, np. *Każdy człowiek umrze, nawet ty.*, *Ogólnie rzecz biorąc ludzie są uczciwi, nawet jeśli niektórzy czasem mają z tym problem.*) Jako agregacja opisywane są także przypadki **rozdzielonego poprzednika** (ang. *split antecedent*), wyróżniające jakiś element ze zbioru opisywanego łącznie w innym miejscu tekstu (*Najpierw przyszedł Jan. Za jakiś czas Maria. Ale Ø wyszli razem.*)

Relacja **kompozycji** łączy całość z jej składową, a zatem opisuje relacje, takie jak *całość – część* czy *obiekt – substancja*, np. *Nie będę już więcej pić wina, nawet kropelki.*

Innym przykładem relacji tego typu jest **anafora związana** (ang. *bound anaphora*), odpowiadająca sytuacji, gdy zaimek tworzy zmienną odpowiadającą elementowi kwantyfikowanego zbioru (np. *Każdy uczestnik konferencji musi przedstawić swój artykuł.*) Ciekawy przypadek podobnego rodzaju stanowi także tzw. **anafora typu E** (Evans 1977), interpretowana często jako relacja bezpośrednia ze względu na tożsamość odwołania jednego z egzemplarzy wskazywanych przez definiowaną klasę i użytego zaimka (np. *Większość ludzi, którzy kupili osła, traktowała go dobrze.*).

Szczegółowy wykaz relacji pośrednich innych typów z literatury światowej, o którym wspominam w pracy (Ogrodniczuk i Zawisławska 2016), okazał się zbyt szczegółowy – jego weryfikacja wykazała niską reprezentatywność relacji innych typów i niewystarczającą zgodność anotatorów przy jednoczesnej dużej wariantywności klasyfikacji. W efekcie postanowiłem zachować szeroką kategorię relacji „innych" w celu reprezentacji przypadków jeszcze nieuwzględnionych w istniejących typologiach, a mogących ujawnić się podczas pracy korpusowej.

3.4.3. Relacje wspierające

Relacje wspierające (definicyjne) łączą wzmiankę z niekoreferencyjnym określeniem tekstowym, ułatwiającym jej interpretację referencyjną. Częstą relacją tego rodzaju test na przykład **metareferencja** wiążąca interpretowaną wzmiankę ze wzmianką odpowiadającą treści napisu, nazwie lub oznaczeniu modelu (*Stanęliśmy przed starym budynkiem. Na ścianie resztki napisu „KINO".*) Inne wyróżniane przez nas relacje tego typu to **porównanie** łączące wzmiankę z jej określeniem porównawczym (*Miał głowę jak jajo.*) czy anafora grupy nominalnej użytej predykatywnie (ang. *predicative anaphora*), która także nie jest relacją referencyjną, gdyż polega na częściowej wspólnocie treści lub prostej replice gramatycznej, a nie na wspólnym odniesieniu.

Podobne wsparcie dekodowania koreferencji (utrudnionej w poniższych przykładach obecnością równoważnych wzmianek o identycznej charakterystyce składniowej) może też nastąpić za pomocą wielu innych relacji semantycznych (których wskaźniki oznaczone są linią przerywaną), takich jak np.: relacja agens – patiens (*Opodatkowanie nie ma miejsca, jeżeli podatnik nabywa używany samochód. Wówczas nabywca nie płaci podatku od towarów i usług.*), relacja aspektowości (*Adam od godziny pisał list i popijał koniak. A kiedy już wszystko wypił i napisał całą odpowiedź, porwał kartkę na strzępy i wrzucił do kosza.*), synonimii nienominalnej (*Bruno uwielbiał Andrzeja, a Tomka nie znosił. Lubił go za jego łatwość wypowiadania się i pracowitość.*), transpozycji (*– To cud! Marcin żyje! Niestety, Grześkowi się nie udało. – Dzięki Bogu, Mój mąż wróci żywy! – krzyknęła Ala.*), przeciwstawności (*Paweł zapalił świeczkę i dał żonie wiązankę. – Zgaś to, nie mam ochoty na świętowanie – powiedziała Ewa.*) czy ogólnego powiązania czasownikowego (*Sebastian jadł gruszkę, a śliwkę zostawił na później. Wgryzał się w soczysty owoc, myśląc o lecie.*) Ze względu na spodziewaną niską reprezentatywność tego typu relacji w tekstach zdecydowałem się nie tworzyć ich szczegółowej typologii przed rozpoczęciem prac korpusowych, jednocześnie polecając anotatorom oznaczanie relacji wykraczających poza podane wcześniej kategorie jako „innych relacji wspierających", co miało pomóc w ich późniejszej analizie.

3.4.4. Relacje wykluczające

Relacje wykluczające (negatywne) służą do reprezentacji informacji o jawnej niekoreferencyjności obiektów, do których odwołuje się oznaczona w ten sposób para wzmianek. Ich celem jest wskazanie nietypowych sytuacji, w których takie

wykluczenie występuje, nie zaś oznaczanie wszystkich par niekoreferencyjnych wzmianek.

Anaforyczna relacja **kategorialności**, łączy obiekty należące do tej samej kategorii (będące okazami tej samej kategorii, „tego samego rodzaju"), ale o różnej referencji (np. *Kupię dziś kwiaty żonie. A Ty kupisz je swojej?*, *Człowiek, który dał czek swojej żonie był mądrzejszy niż człowiek, który dał go kochance.*) W języku angielskim przypadki tego rodzaju opisywane są często jako **anafora leniwa** (ang. *lazy anaphora*; patrz Karttunen 1976) albo **identyczność sensu** (ang. *identity-of-sense*) na zasadzie kontrastu z relacją **identyczności odniesienia** (ang. *identity-of-reference*). Zaimek funkcjonuje w nich jedynie jako zastępnik eliminujący powtórzenie tego samego wyrażenia.

Relacja **kontrastu** łączy przeciwstawione w tekście wzmianki o różnej referencji, z jawnie zaznaczonym zerwaniem linku między obiektami, co do których powiązania mogłaby zaistnieć wątpliwość, np. *Losy Czech wydają się bardziej związane z Niemcami, a losy Słowacji – z Rosją.*, *Nie wiem, jak ktoś mógłby pomyśleć, że Ø jestem ojcem Zosi.*. Nie łączy się relacją kontrastu kohiponimów, chyba że może istnieć wątpliwość co do ich niekoreferencyjności. Jako wykluczające się oznacza się także wzmianki **polisemiczne**, np. *Misja rozpoczęła się 8 kwietnia. W skład misji weszły 24 osoby.* Podobnie jak w przypadku relacji wspierających również dla relacji wykluczających przewidziano kategorię „relacji innej".

3.4.5. Aspekty

Aspekty mogą być przypisywane relacjom referencyjnym (zarówno koreferencji, jak i relacjom pośrednim, np. *To ich mieszkanie – owszem, ładne, ale ich dawna kuchnia była jednak większa.*) w celu zasygnalizowania ich szczególnych własności wpływających na odbiór referencji w danym kontekście. Wprowadzenie pojęcia aspektu pozwala na reprezentację opozycji pomiędzy własnościami tekstu traktowanymi jako domyślne (obiektywizm, pewność czy bezstronność nadawcy) a nacechowaniem tekstu (subiektywizm, niewiedza, uprzedzenie nadawcy). Pojęcie aspektu jest zbliżone do pojęcia *near-identity* (Recasens i in. 2011), *quasi-identity* (Ogrodniczuk i in. 2015: s. 62) czy intencji modelowania aktów mowy i przekonań jako cech relacji metatekstowych w powstającej właśnie trzeciej edycji korpusu PDTB (Webber i in. 2016). W proponowanej typologii wyróżniam następujące rodzaje aspektów.

Dysymilacja (inaczej: rozpodobnienie) sygnalizuje istnienie pewnego dodatkowego poziomu mentalnego w interpretacji referenta, co może powodować trudności w zdekodowaniu wartości referencyjnej wzmianek; zależnie od interpretacji efektem tego rozmycia jest traktowanie dwóch wzmianek jako odwołujących się do tego samego obiektu (czyli koreferencji) lub do dwóch różnych obiektów (czyli braku koreferencji). Dysymilacja może przyjmować postać rozmycia[3] pewnej cechy rozłącznych przedmiotów referencji w taki sposób, że oba odwołania wskazują na swoisty meta-obiekt (jak w zdaniu *Nie widziała „Przeminęło z wiatrem", ale czytała je.*, przywołującym obiekt mentalny odpowiadający zarówno książce, jak i filmowi posiadającym wspólną treść) albo wyodrębnienia pewnej cechy obiektu w sposób pozornie go rozdzielający, co szczególnie wyraźnie objawia się w przypadku cechy zmienności w czasie: *Warszawa jest pięknym miastem, ale przedwojenna Warszawa była jeszcze piękniejsza.* W przeciwieństwie do pojęcia *near-identity* relacji dysymilacji nie przypisuję własności rozmycia tożsamości obiektu, a jedynie zasygnalizowania trudności interpretacyjnej w dekodowaniu referenta w sytuacji niewerbalnego powoływania do życia pomocniczych obiektów o odrębnej tożsamości.

Niepewność jest aspektem służącym oznaczeniu wyrażonej przez nadawcę niejasności, niedookreślenia związku między parą obiektów: *Jest prezydentem, ale nie wiem, czy prezydentem Warszawy czy Ø Krakowa.*, *– Janek jest mężem Basi. – Nie, Janek jest mężem Emmy!*

Opinia (atrybucja) określa z kolei subiektywną naturę powiązania, wyrażoną przez podmiot tekstu, np. *Ewa znów się spóźniła, jak ja mam dość tej idiotki.* lub jego bohatera: *To jest Michał, niektórzy mówią, że przyszły prezydent Polski.*

[3] Por. pojęcia kompresji i dekompresji Fauconniera i Turnera (2002).

4.

Korpus zależności referencyjnych

Na potrzeby weryfikacji typologii relacji referencyjnych zaproponowanej w poprzednim rozdziale oraz w celu implementacji komputerowych narzędzi do wykrywania relacji referencyjnych został zebrany korpus tekstów reprezentatywny dla polszczyzny, a następnie przeprowadzono jego anotację zgodnie z tą typologią, tj. „proces nadawania interpretacji lingwistycznej danym językowym zgromadzonym w korpusie" (Leech 1997: s. 2). Podejście tego rodzaju jest obecnie typowe dla zadań inżynierii lingwistycznej, przede wszystkim ze względu na sukcesy osiągane przez systemy uczenia maszynowego z nadzorem (ang. *supervised machine learning*), dla których dane anotowane ręcznie są podstawowym surowcem.

Oprócz sformalizowania metod reprezentacji cech relacji referencyjnych w tekście istotną kwestią był także wybór źródeł do anotacji zapewniających dobrą reprezentację języka ogólnego. Nie sposób także nie wspomnieć o decyzjach technicznych, dotyczących zarówno organizacji procesu anotacji, jak i wyboru narzędzi służących do identyfikacji opisywanych cech tekstu.

Wynikiem prac anotacyjnych jest korpus zależności referencyjnych[1] – liczący ponad pół miliona słów, anotowany ręcznie zasób opisany relacjami referencyjnymi zgodnie z zaproponowaną wcześniej metodologią. W momencie publikacji książki jest to jeden z największych tego typu korpusów na świecie.

4.1. Wybór tekstów

Rozmiar korpusu został arbitralnie ograniczony do ok. 500 tys. słów ze względów praktycznych, wynikających z możliwości opłacenia ręcznej anotacji określonej liczby próbek ze środków grantu badawczego. Rozmiar pojedynczej próbki został natomiast ustalony na podstawie wstępnej analizy gatunkowej na ok. 300 słów ze względu na potrzebę zbadania zależności wykraczających poza poziom pojedynczego akapitu przy jednoczesnej niemożności udostępnienia całych tekstów chronionych prawem autorskim. Już po rozpoczęciu anotacji decyzja ta okazała się

[1] Dostępny na stronie `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`.

zgodna z nowymi badaniami nad zrozumiałością tekstu (Maziarz i in. 2012: s. 17), oceniającymi średnią długość akapitu tekstu prasowego[2] na 40–80 słów oraz z rozmiarem próbki przyjętym w niezależnym projekcie anotacji zależności anaforycznych na tekstach Korpusu Politechniki Wrocławskiej (Broda i in. 2012b), w którym liczba 300 słów jest opisywana jako wystarczająco duża do śledzenia relacji tego rodzaju i jednocześnie na tyle mała, by było możliwe uzyskanie zgody na wykorzystanie źródła.

Zarówno same próbki tekstów, jak i zasada konstrukcji korpusu, zapewniająca jego zrównoważenie i reprezentatywność – warunki konieczne do uznania korpusu za wiarygodny model języka – zostały przejęte z Narodowego Korpusu Języka Polskiego, podczas budowy którego uwzględniono m.in. takie czynniki, jak strukturę czytelnictwa w Polsce, różnorodność typologiczną i tematyczną języka, zrównoważenie stylistyczne oraz czasowe (Górski i Łaziński 2012). Czynniki te znacznie uprościły proces konstrukcji tego zasobu. Korpus zależności referencyjnych wykorzystuje zatem teksty 300-milionowego podkorpusu NKJP[3], wylosowane tak, by odzwierciedlać proporcje poszczególnych gatunków w korpusie wzorcowym (Przepiórkowski i in. 2012: tabela 5.1, s. 53).

Ekstrakcja danych polegała na losowym wyborze tekstu w ramach żądanego typu, a następnie losowym wyborze w tym tekście ciągłego zestawu pełnych akapitów o długości od 250 do 350 segmentów wyrazowych – jednostek odpowiadających wewnętrznej budowie słów polskich i wchodzących w samodzielne związki składniowe (zgodnie z tą zasadą słowo *potrzebował|że|by|ś* składa się z czterech segmentów)[4]. Z NKJP przejmowany był oryginalny podział na akapity, uzupełniany w przypadku tekstów mówionych o informację o mówcy, dostępną w metadanych. Wyniki losowania były dodatkowo przeglądane w celu odrzucenia próbek zawierających dane nieistotne dla badań nad zależnościami referencyjnymi – tekst nieciągły (np. spisy treści, tabele) czy złożony (np. z wielu krótkich notek prasowych zapisanych w NKJP jako pojedyncza próbka).

[2]Najczęstszego gatunku w korpusie zrównoważonym – por. dyskusje na temat budowy korpusu wzorcowego w książce NKJP (Przepiórkowski i in. 2012: rozdział 3, s. 25–36).

[3]Anotowany ręcznie podkorpus milionowy zapewniający jeszcze lepszą jakość tekstów nie mógł zostać wykorzystany ze względu na wielkość próbek ograniczoną do ok. 50 wyrazów.

[4]Osobnymi segmentami są też znaki interpunkcyjne oprócz wyjątków w rodzaju traktowanych jako pojedyncze segmenty form z apostrofem (*Chomsky'ego*) czy łącznikiem sygnalizującym odmianę (*PRL-u*); pełną listę kategorii tego rodzaju przypadków zawiera rozdział 6.2.2 książki opisującej korpus narodowy (Przepiórkowski i in. 2012).

Dodatkowo w celu umożliwienia zbadania zależności relacji koreferencyjnych od długości tekstu do korpusu zostało włączonych 21 kompletnych tekstów (po trzy z siedmiu działów: publicystyka/opinie, prawo, ekonomia, sport, nauka i technika, kultura, wiadomości krajowe) pochodzących z tzw. Korpusu Rzeczpospolitej (Presspublica 2002)[5]. Łącznie rozmiar tego podkorpusu „tekstów długich" (w odróżnieniu od podstawowego podkorpusu „tekstów krótkich") wyniósł ok. 36 tys. segmentów. W dalszej części wywodu posługuję się tymi nazwami, podając informacje statystyczne dla każdego podkorpusu osobno. Tabela 4.1 przedstawia strukturę korpusu z podziałem na kategorie tekstów oraz z informacją o procentowym udziale poszczególnych stylów i źródeł w każdym z podkorpusów.

4.2. Wybór strategii anotacyjnej

Sposób organizacji anotacji na potrzeby przetwarzania języka naturalnego jest tematem niezwykle obszernym, który doczekał się wielu szczegółowych opracowań (por. np. Pustejovsky i Stubbs 2012, Fort 2016). Jego najważniejszymi elementami są: wybór strategii anotacyjnej, narzędzi, powołanie zespołu współpracowników i ocena jakości prowadzonych prac. Jakkolwiek nie istnieją definitywne reguły dla tego procesu, zakończone wcześniej projekty anotacji korpusowej dostarczają cennych wskazówek w zakresie jego przeprowadzenia.

4.2.1. Liczba i profil anotatorów

Już sam problem podziału pracy wymaga kilku merytorycznie uzasadnionych ustaleń organizacyjnych. Zapewnienie opisu lingwistycznego wysokiej jakości wymaga z całą pewnością weryfikacji decyzji podejmowanych w procesie anotacyjnym, musi zatem istnieć pewien mechanizm kontroli działania pojedynczego anotatora. Mimo że wśród badaczy nie ma zgodności co do sposobu przeprowadzenia anotacji, który skutkowałby najlepszymi wynikami, najpopularniejszą metodą, zastosowaną również w projekcie NKJP (Przepiórkowski i in. 2012: rozdział 5.1), jest metoda *anotacji równoległej*, gdy wielu anotatorów–lingwistów (zwykle dwóch ze względów oszczędnościowych) wykonuje to samo zadanie dla danej próbki tekstu, a następnie różnice między ich decyzjami weryfikuje tzw. superanotator –

[5] Korpus ten został wybrany ze względu na jego dostępność i długą tradycję wykorzystania w polskiej lingwistyce komputerowej przy jednoczesnym braku możliwości swobodnego użycia pełnych tekstów NKJP.

Tabela 4.1. Struktura korpusu zależności referencyjnych

Źródło	Kategoria tekstu	Liczba tekstów	Liczba segmentów	Udział w podkorpusie
Narodowy Korpus Języka Polskiego	Dzienniki	442	122 917	24,79 %
	Pozostałe periodyki	402	116 407	23,48 %
	Literatura piękna	286	79 703	16,07 %
	Typ informacyjno-poradnikowy	98	27 212	5,49 %
	Literatura faktu	95	27 357	5,52 %
	Mówione konwersacyjne	83	25 370	5,12 %
	Internetowe interaktywne (blogi, fora, usenet)	63	17 755	3,58 %
	Internetowe nieinteraktywne (strony statyczne, wikipedia)	62	17 478	3,52 %
	Inne teksty pisane	55	15 142	3,05 %
	Mówione medialne	44	12 811	2,58 %
	Quasi-mówione (protokoły sesji parlamentu)	43	12 791	2,58 %
	Typ naukowo-dydaktyczny	35	10 259	2,07 %
	Książki publicystyczne	19	5 497	1,11 %
	Książki niebeletrystyczne nieklasyfikowane	18	5 167	1,04 %
	Łącznie	1 745	495 866	100,00 %
Korpus Rzeczpospolitej	Publicystyka, opinie	3	7 081	19,51 %
	Prawo	3	5 920	16,31 %
	Ekonomia	3	5 882	16,20 %
	Kraj	3	5 174	14,25 %
	Sport	3	4 326	11,92 %
	Kultura	3	4 125	11,36 %
	Nauka i technika	3	3 792	10,45 %
	Łącznie	21	36 300	100,00 %

ekspert w zakresie danego tematu. Na pytanie, czy praca superanotatora jest w ogóle potrzebna, należy odpowiedzieć twierdząco – istnienie „złotego standardu" (w porównaniu z pozostawieniem w korpusie wszystkich anotacji wariantywnych i dostosowaniu do nich metod ewaluacji) wydaje się pożyteczne ze względu na możliwość jego użycia przez standardowe metody oceny jakości wykorzystujące pojedynczą anotację „optymalną", a przy tym nie wyklucza udostępnienia anotacji cząstkowych. Beigman Klebanov i Beigman (2009) potwierdzili, że większa liczba anotatorów pozwala na wyeliminowanie wpływu przypadku na wyniki anotacji, Bayerl i Paul (2011) zaś zasugerowali, że większość zadań powinna zostać przeprowadzona przy udziale trzech lub czterech anotatorów, a najtrudniejszych – nawet pięciu. Uwzględniając to założenie oraz jednocześnie dążąc do optymalizacji kosztów, w naszych pracach zaplanowałem wstępnie udział trzech anotatorów i superanotatora.

Kolejną kwestią do rozważenia na etapie planowania prac jest profil potencjalnego anotatora – zakres jego wiedzy lingwistycznej i doświadczenie. Niezależnie od popularności metodologii crowdsourcingowych (wykorzystujących do specjalistycznych zadań dużą liczbę ochotników[6] zazwyczaj bez specjalistycznej wiedzy, po czym uśredniających ich decyzje) zostało jednak potwierdzone (Bhardwaj i in. 2010), że jakość anotacji doświadczonych anotatorów–specjalistów jest wyższa niż w przypadku anotatorów „przypadkowych" (ang. *crowd of non-experts*). W związku z tym, że zadanie anotacji relacji referencyjnych było mimo wszystko dla języka polskiego zadaniem eksperymentalnym, zakładającym podejmowanie decyzji strategicznych (takich jak np. doprecyzowywanie typologii relacji) w trakcie anotacji, z podejścia crowdsourcingowego postanowiłem w ogóle zrezygnować i polegać na pracy ekspertów–językoznawców, doświadczonych w zadaniach anotacji lingwistycznej różnych warstw opisu NKJP.

4.2.2. Anotacja szeregowa a anotacja równoległa

Podczas pierwszych eksperymentów anotacyjnych z kilkoma anotatorami i superanotatorem okazało się, że praca superanotacyjna jest wyjątkowo uciążliwa ze względu na konieczność jednoczesnej korekty potencjalnie dużej liczby niezgodności między wieloma wariantami struktur pochodzących z różnych warstw anotacji (różnice w granicach wzmianek, przypisaniu wzmianek do klastrów, wyborze

[6] Środowisko Phrase Detectives (Poesio i in. 2015) gromadzi wg jego twórców średnio 20–30 wariantów anotacji opisywanej jednostki.

centrów semantycznych, wyrażenia dominującego dla klastra, linków nieidentyczności owych). Dodatkowy problem stanowiła sama natura relacji referencyjnych, potencjalnie rozpiętych między odległymi fragmentami tekstu, w wyniku czego weryfikacja różnic anotacyjnych ograniczonych do lokalnego kontekstu powodowała trudności natury praktycznej: konieczność stałej konsultacji każdego przypadku z całym tekstem.

Zasugerowane przez superanotatora rozwiązanie tego problemu metodą kolejnych przybliżeń (najpierw korekta granic wzmianek, następnie korekta zawartości klastrów itd.) okazało się dużo łatwiejsze w trybie weryfikacji i korekty wyników pracy na tekście zawierającym oznaczenia pojedynczego anotatora, a nie porównywania wyników pracy kilku anotatorów. Dokonano zatem sprawdzenia, czy tego rodzaju metoda, nazwana *anotacją szeregową*, daje wyniki porównywalne z metodą anotacji prowadzonej równolegle przy zachowaniu tego samego poziomu kosztów (czyli liczby wariantów anotacji danego tekstu).

W eksperymencie przeprowadzonym na próbce 30 tekstów wykonano ich anotację jednocześnie metodą szeregową (z udziałem trzech anotatorów, z których każdy kolejny weryfikował i poprawiał pracę poprzedniego) i równoległą (w ramach której superanotator scalał wyniki prac dwóch anotatorów działających jednocześnie). Następnie porównano wyniki obu anotacji ze zbiorem stworzonym metodą konsylium, czyli posiedzeń grupy ekspertów (rozłącznej z zespołem anotatorów) mających za zadanie wypracowanie wspólnej „wzorcowej" wersji anotacji każdego z tekstów. Okazało się, że zgodność wyników uzyskanych metodą szeregową z wersją konsylium jest wyższa niż w przypadku metody równoległej w niemal wszystkich rodzajach badanych wskaźników (zgodność wzmianek, ich centrów semantycznych, klastrów) – poza zgodnością wyrażeń dominujących, mniej istotnych z punktu widzenia anotacji relacji referencyjnych (por. Ogrodniczuk i in. 2015: tabele 6.3–6.5, s. 89–90). W dalszych pracach nad korpusem przyjęto zatem metodologię anotacji szeregowej.

4.2.3. Preanotacja

Ważną kwestią jest także decyzja co do stosowania preanotacji, czyli wstępnego przetworzenia anotowanych tekstów narzędziami automatycznymi w celu wykrycia części anotowanych zjawisk (zwykle tych prostszych), przez co praca anotacyjna polega na poprawianiu i uzupełnianiu już istniejących oznaczeń, a nie tworzeniu ich od początku. Mimo niebezpieczeństwa pominięcia bardziej szczegółowych zjawisk

przez anotatorów zbyt ufających decyzjom automatu, na podstawie wcześniejszych studiów (np. Fort i Sagot 2010) podjęto decyzję o zastosowaniu preanotacji ze względu na związane z nią ogromne oszczędności (autorzy wspominają m.in. o dwukrotnym przyspieszeniu właściwej anotacji dzięki zastosowaniu automatycznej fazy wstępnej nawet przy użyciu narzędzia o niskiej jakości oraz o zaletach preanotacji w procesie wykrywania problemów w instrukcji anotacyjnej już na najwcześniejszym etapie prac).

Decyzja ta wydaje się uzasadniona zwłaszcza w kontekście anotacji wzmianek, z których większość złożona jest z pojedynczych wyrazów (rzeczowników, zaimków), co pozwala na ich proste i wiarygodne wykrycie istniejącymi metodami analizy morfoskładniowej. Mechanizm preanotacyjny został zatem włączony w proces anotacyjny już w jego wczesnym stadium (patrz rozdział 4.3.1).

4.2.4. Superanotacja automatyczna

W związku z dużym stopniem komplikacji zadania superanotacyjnego, wymagającego analizy wielu poziomów anotacji autorstwa wielu anotatorów, został przeprowadzony także eksperyment z automatycznym łączeniem anotacji cząstkowych, który miał odpowiedzieć na pytanie, czy istnieje możliwość całkowitego wyeliminowania superanotatora–człowieka z procesu anotacji i zastąpienia jego decyzji metodami maszynowymi. Mimo że prace zakładają udział anotatorów-ekspertów, weryfikacja możliwości automatycznego łączenia anotacji może mieć także znaczenie w obliczu szerokiego wykorzystania metod crowdsourcingowych w anotacji semantycznej, co daje łatwy dostęp do dużej liczby anotacji o nieznanej jakości.

W związku z tym, że nowe modele automatycznej poprawy błędów w rodzaju MACE (Hovy i in. 2013) nie zapewniają wyższej jakości niż proste techniki głosowania większościowego[7], przeprowadzono analizę jakościową i kosztową kilku prostych większościowych strategii superanotacyjnych dla zadania anotacji wzmianek. Przyjmuję oznaczenie x/y dla strategii włączania do wyniku wzmianki wskazanej przez x z łącznej liczby y anotatorów danej próbki. Jeśli liczba dostępnych anotacji jest większa niż y, wynik końcowy to średnia z wyników cząstkowych uzyskanych z analizy wszystkich relewantnych podzbiorów zestawów anotacji odpowiadających danej strategii.

[7] Za Rehbein i Ruppenhoferem (2017), na przykładzie kilku wariantów różnych zadań z dziedziny przetwarzania języka naturalnego.

Materiałem do oceny strategii są dane stu losowo wybranych tekstów korpusu anotowane przez pięciu niezależnych anotatorów, dla których w trybie konsylium stworzono superanotację wzorcową. Dodatkowo wykonano superanotację w trybie szeregowym (patrz rozdział 4.2.2) – przekazano superanotatorowi po 20 wybranych losowo tekstów z opisanego zestawu, zaanotowanych przez każdego z pięciu anotatorów. Tabela 4.2 przedstawia ocenę zgodności pojedynczych anotatorów oraz ich średnią (odpowiadającą wartości dla symulowanego pojedynczego anotatora), tabela 4.3 zaś – zgodność superanotacji i wyników działania omówionych strategii z anotacją wzorcową.

Tabela 4.2. Zgodność anotatorów z anotacją wzorcową

	Anotator 1	**Anotator 2**	**Anotator 3**	**Anotator 4**	**Anotator 5**	**Średnia**
P	89,31%	90,61%	88,22%	87,86%	90,79%	89,36%
R	84,98%	90,88%	90,13%	83,25%	89,59%	87,77%
$\mathbf{F_1}$	87,09%	90,75%	89,16%	85,49%	90,18%	88,54%
Koszt	n	n	n	n	n	—

Tabela 4.3. Zgodność superanotacji oraz strategii automatycznych z anotacją wzorcową

	2/3	**3/4**	**2/4**	**3/5**	**1/3**	**Superanotator**
P	91,27%	93,33%	89,41%	91,75%	82,74%	91,96%
R	87,80%	85,48%	90,11%	87,84%	94,30%	90,60%
$\mathbf{F_1}$	89,50%	89,23%	89,76%	89,75%	88,14%	**91,27%**
Koszt	$3n$	$4n$	$4n$	$5n$	$3n$	$2n$

Uzyskane wyniki wskazują m.in. na zgodną z intuicją najwyższą precyzję wyniku zapewnioną przez decyzję największej procentowo liczby anotatorów z badanego zestawu (dla wariantu 3/4) oraz największe pokrycie dla wariantu dopuszczającego decyzję co najmniej jednego anotatora (1/3). Mimo to największą średnią zgodność z wynikiem wzorcowym zapewnia anotacja szeregowa ze wsparciem superanotatora, dodatkowo wykazując najniższe koszty (dwie anotacje na tekst – bazowa i superanotacja), podczas gdy każda strategia łącząca ma koszt równy mianownikowi opisującego ją ułamka. Z tego też powodu podjęto decyzję o przeprowadzeniu pełnej superanotacji korpusu i rezygnacji ze stosowania metod superanotacji automatycznej.

4.3. Prace anotacyjne

Prace anotacyjne prowadzono w dwóch etapach, poprzedzonych fazą rozpoznawczą: w pierwszym powstał korpus koreferencji nominalnej i narzędzia do wykrywania relacji koreferencyjnych tego rodzaju, w drugim – jego poprawiona i rozszerzona wersja, wzbogacona o relacje referencyjne innych typów oraz nowe wersje narzędzi. Podejście to, wynikające z jednej strony ze względów czysto praktycznych (dwa osobne projekty anotacyjne), z drugiej zaś – z potrzeby weryfikacji anotacji powstałych zwłaszcza na pierwszym etapie prac, wydaje się pożyteczne i jest stosowane przez wielu badaczy na świecie. Przykładowo korpus praski (Prague Dependency TreeBank, PDT) w wersji 2.0 (Panevová i in. 2000) zawierał anotację koreferencji tekstowej i gramatycznej na szczątkowym poziomie; wersja poboczna o nazwie PDiT 1.0 (Prague Discourse Treebank, patrz Nedoluzhko i in. 2009) rozbudowująca relacje poziomu dyskursu rozszerzyła anotację koreferencji tekstowej o dodatkowe typy oraz wprowadziła anotację referencji pośredniej, a ostatecznie scalona wersja PDT 3.0 (Zikánová i in. 2015: rozdziały 3 i 4) zawiera anotację koreferencji zaimkowej w pierwszej i drugiej osobie oraz dodatkowe poprawki.

4.3.1. Faza rozpoznawcza

Właściwe prace anotacyjne poprzedził etap rozpoznawczy[8], którego celem było przetestowanie wstępnej instrukcji anotacyjnej w prostym środowisku, jeszcze przed wyborem docelowego narzędzia anotacyjnego. Eksperyment anotacyjny został przeprowadzony na 50 fragmentach tekstów o długości 20 zdań, wylosowanych z Narodowego Korpusu Języka Polskiego. Dane przekazane pojedynczemu anotatorowi–lingwiście zostały wstępnie przetworzone zestawem narzędzi analizy lingwistycznej dla języka polskiego w celu automatycznego wykrycia wzmianek – fraz nominalnych będących przedmiotem anotacji. W skład zestawu narzędzi weszły: analizator morfologiczny Morfeusz SGJP[9] (Woliński 2006), tager ujednoznaczniający Pantera[10] (Acedański 2010), parser powierzchniowy Spejd[11] (Prze-

[8] Jego szczegółowy opis zawiera rozdział 6.1 pracy (Ogrodniczuk i in. 2015: s. 89–94).

[9] Dostępny na stronie `http://sgjp.pl/morfeusz/`.

[10] Dostępny na stronie `http://zil.ipipan.waw.pl/PANTERA`.

[11] Dostępny na stronie `http://zil.ipipan.waw.pl/Spejd`.

piórkowski i Buczyński 2007) oraz narzędzie do rozpoznawania nazw własnych Nerf[12] (Waszczuk i in. 2013).

Treść próbek została zapisana w tekstowym, wielokolumnowym formacie SemEval/CoNLL (patrz rozdział 4.7.1), zawierającym w kolejnych liniach informacje o segmentach wyrazowych analizowanego tekstu, którym w poszczególnych kolumnach przypisano lemat, wartości kategorii gramatycznych oraz reprezentację wzmianek ze wskazaniem ich centrów semantycznych. Prostota formatu pozwala na jego rozbudowę o kolejne kolumny; w tym przypadku o reprezentację klastrów koreferencyjnych w formacie nawiasowym (ta sama liczba odpowiada temu samemu klastrowi).

Zadaniem anotatora było uzupełnienie opisu o brakujące wzmianki i klastry, usunięcie anotacji nadmiarowych oraz ewentualnie poprawienie oznaczeń centrów semantycznych. W wyniku eksperymentu udało się sformułować kilka założeń dla dalszych etapów procesu, przede wszystkim potrzebę szerokiej anotacji wzmianek zgodnie z ich potencją referencyjną, niezależnie od tworzonych przez nie związków anaforycznych w danym tekście oraz anotacji granic wzmianek zgodnie z ich pełną semantyką, bez ograniczania się jedynie do centrów wzmianek. Jest to podejście pozornie tylko różne od zastosowanego na przykład w korpusie czeskim, gdzie oznaczeniu podlegają centra wzmianek, ale uzupełnione o pełne, ręcznie ujednoznacznione poddrzewa analizy zależnościowej. Podejście takie nie było możliwe do uzyskania w ramach obecnych prac ze względu na duże koszty anotacji składniowej tego rodzaju i niedostateczną jakość dostępnych automatycznych parserów składniowych, stąd decyzja o realizacji dużo prostszego zadania tekstowego wyznaczania granic wzmianek odpowiadających drzewom rozbioru fraz.

4.3.2. Anotacja koreferencji nominalnej

Etap anotacji koreferencji nominalnej został przeprowadzony z udziałem siedmiu anotatorów i superanotatora wspomnianą wcześniej metodą anotacji szeregowej na podstawie przekazanych instrukcji (dot. merytorycznych i technicznych aspektów zadania). Teksty zostały poddane automatycznej preanotacji, praca anotatora polegała zatem na wykryciu i poprawieniu błędów anotacji automatycznej (niewykrycia lub nadmiarowego oznaczenia wzmianek, błędnych granic wzmianek,

[12] Dostępny na stronie `http://zil.ipipan.waw.pl/Nerf`.

nieoznaczenia lub błędnego oznaczenia centrów fraz, braku powiązań czy niewłaściwego powiązania między wzmiankami). Wśród wskazówek anotacyjnych znalazły się także zalecenia wstępnego przeczytania całego tekstu i zastosowania podejścia wieloetapowego (z podziałem na fazę wyróżniania wzmianek przy pierwszym czytaniu tekstu oraz oznaczania klastrów i relacji w kolejnych przebiegach), a także zachęta do konsultacji kwestii mogących budzić wątpliwości z innymi anotatorami za pomocą stworzonej do tego celu grupy dyskusyjnej.

Instrukcja anotacyjna zawierała liczne przykłady konkretnych konstrukcji składniowych, które miały podlegać anotacji wraz z wyjaśnieniami (np. odnośnie anotacji zaimków w mowie zależnej, fraz elektywnych czy wyliczeń) oraz wskazówki natury technicznej, np. dotyczących sposobu oznaczania podmiotów domyślnych poprzez tworzenie wzmianek z form osobowych czasowników związanych z danym podmiotem (np. *Zapytałam go, czy przyjdzie, ale mnie kompletnie zignorował.*)

W celu ułatwienia badań nad koreferencją w ramach opisu każdej z anotowanych wzmianek oznaczano **centrum semantyczne**, czyli „ten segment (lub wyraz składniowy) grupy, który najlepiej definiuje znaczenie danej grupy”[13]. Każdemu klastrowi koreferencyjnemu przypisywano ponadto **wyrażenie dominujące**, czyli „najpełniej go określające” – nazwę własną (imię, nazwisko, toponim, tytuł itp.), deskrypcję określoną lub inne wyrażenie o najbogatszej semantyce (czyli np. w łańcuchu nazw: *zwierzę* → *pies* → *jamnik* byłby to rzeczownik *jamnik*). Wyrażenie dominujące nie musiało pojawić się jako wzmianka w tekście, lecz mogło zostać stworzone z treści kilku wzmianek albo podane ręcznie.

Oprócz anotacji koreferencji nominalnej przeprowadzono także eksperyment z anotacją relacji quasi-identycznościowych w rozumieniu zbliżonym do Recasens i in. (2011). Anotatorzy zostali poproszeni o wskazanie między parami wzmianek relacji zbliżonych do bezpośrednich, lecz o pewnym stopniu wątpliwości interpretacyjnej (bez wskazywania konkretnej typologii relacji tego rodzaju). 5100 oznaczeń tego rodzaju stało się zalążkiem prac nad całościowym modelem relacji referencyjnych.

[13] „Na przykład w grupach: *wysoki student*, *najwyższy ze studentów* i *pięciu wysokich studentów*, centrami semantycznymi są formy rzeczownika *STUDENT*, choć grupy te różnią się centrami składniowymi: w pierwszym wypadku jest to segment *student* (*wysoki student* to grupa rzeczownikowa), w drugim – segment *najwyższy* (grupa przymiotnikowa, tzw. elektywna), a w trzecim – segment *pięciu* (grupa liczebnikowa)”. (Przepiórkowski 2008: rozdział 6.2.3, s. 110).

Anotacja została przeprowadzona z zastosowaniem narzędzia MMAX4CORE[14] opracowanego na bazie edytora MMAX2[15] (Müller i Strube 2006); do dystrybucji plików użyto programu DISTSYS[16] (patrz rozdział 4.4). Szczegółowy opis prac, które złożyły się na pierwszy etap anotacji, informacje na temat decyzji anotacyjnych, instrukcje obsługi narzędzi i obszerne wyjaśnienia metodologiczne z zakresu relacji bezpośrednich zawierają rozdziały 4–7 książki (Ogrodniczuk i in. 2015).

4.3.3. Anotacja ogólnych zależności referencyjnych

W drugim etapie anotacji została podjęta próba opisu brakujących konstrukcji składniowych koreferencyjnych ze wzmiankami nominalnymi oraz relacji pośrednich na plikach już istniejącego korpusu zgodnie z nowo powstałą instrukcją anotacyjną. Zadanie anotatorów polegało w szczególności na uzupełnieniu anotacji o nowe wzmianki nienominalne odpowiadające *per analogiam* konstrukcjom nominalnym (*Zapadał zmierzch. Bardzo go to przestraszyło.* – *„Zapadanie zmierzchu bardzo go przestraszyło.*), połączeniu ich w klastry koreferencyjne, dodaniu nowych relacji pośrednich odpowiednich typów oraz relacji wspierających i aspektów. Jednocześnie dokonano uzupełnienia (m.in. o konstrukcje eliptyczne) i weryfikacji opisu koreferencji nominalnej oraz korekty błędów w korpusie, traktując to zadanie jako kolejny etap anotacji szeregowej.

Anotacja relacji referencyjnych została przeprowadzona z zastosowaniem narzędzia MMAX4REF[17] (patrz rozdział 4.4), wersji narzędzia MMAX4CORE rozszerzonej o relacje z taksonomii zbiorczej opisanej w rozdziale 3. Przekazane instrukcje anotacyjne oprócz części merytorycznej zawierały także liczne wskazówki techniczne (np. odnośnie kierunku linków asocjacyjnych i wspierających). W związku z niedospecyfikowaniem relacji „innych”, podobnie jak w przypadku relacji quasi-identycznościowych z poprzedniego etapu prac, w instrukcji podano przykłady tego rodzaju relacji i poproszono anotatorów o dodawanie komentarzy dotyczących ich natury, które mogłyby posłużyć do ewentualnego uszczegółowienia taksonomii w systematyczny sposób.

[14] Dostępne na stronie `http://zil.ipipan.waw.pl/MMAX4CORE`.

[15] Aktualna wersja oryginalnego narzędzia anotacyjnego MMAX2 jest dostępna na stronie `http://mmax2.net/`; rozdział 7.1.2 w książce (Ogrodniczuk i in. 2015) zawiera szczegółowe informacje o zmianach wprowadzonych do oryginalnego narzędzia w ramach jego dostosowania do potrzeb projektu anotacji relacji bezpośrednich; dalsze rozszerzenia miały już charakter mniej radykalny i pomijam ich opis dla zachowania zwięzłości wywodu.

[16] Dostępny na stronie `http://zil.ipipan.waw.pl/DistSys`.

[17] Dostępne na stronie `http://zil.ipipan.waw.pl/MMAX4REF`.

4.4. Narzędzia anotacyjne

Ze względu na potrzebę umożliwienia anotatorom pracy bez aktywnego połączenia z internetem proces anotacji prowadzony był z wykorzystaniem narzędzia instalowanego w systemie operacyjnym komputera (w odróżnieniu od aplikacji działających w przeglądarce internetowej). Ten sposób pracy wymagał dystrybucji danych pomiędzy anotatorami i był realizowany za pomocą programu DISTSYS umożliwiającego pobranie transzy plików do anotacji lub superanotacji, wywołanie programu anotacyjnego i odesłanie wyników pracy do repozytorium. Serwer programu DISTSYS odpowiadał za losowe przydzielanie plików anotatorom w sposób zapewniający powstanie odpowiedniej liczby anotacji każdego pliku zgodnie z przyjętą strategią. Program klienta oprócz uruchamiania wymiany plików oraz narzędzia anotacyjnego umożliwiał także m.in. zapis na serwerze roboczych wersji anotowanych plików, prowadzenie anotacji z wielu komputerów czy odrzucanie przydzielonych tekstów zawierających błędy.

Po pobraniu na dysk anotatora tekst mógł zostać otwarty we właściwym narzędziu anotacyjnym – programie MMAX4CORE/MMAX4REF (w zależności od etapu anotacji) przygotowanym na bazie edytora MMAX2. Wybór programu MMAX2 jako platformy anotacyjnej (po jego porównaniu z innymi dostępnymi narzędziami[18]) był wynikiem spełnienia kilku warunków uznanych za kluczowe dla procesu anotacji semantycznej. Podstawowym była konieczność zapewnienia obsługi anotowanych konstrukcji składniowych: nieciągłości i zagnieżdżeń fraz, możliwości tworzenia klastrów wzmianek (a nie wyłącznie linków) i przypisywania klastrom określonych własności. Ze względu na konieczność dostosowania programu do potrzeb wybranej taksonomii relacji referencyjnych niezwykle istotna była także dostępność kodu na zasadach open source oraz jego implementacja w jednym z popularnych języków programowania, co ułatwiało rozbudowę i integrację narzędzia. Warunkiem koniecznym była możliwość użycia standardowego, otwartego formatu anotacji zewnętrznej (ang. *stand-off*). Nie bez znaczenia okazał się też prosty interfejs użytkownika, rozszerzalny i dostosowywalny do indywidualnych potrzeb (np. za pomocą skrótów klawiaturowych do najczęściej używanych poleceń), posiadający

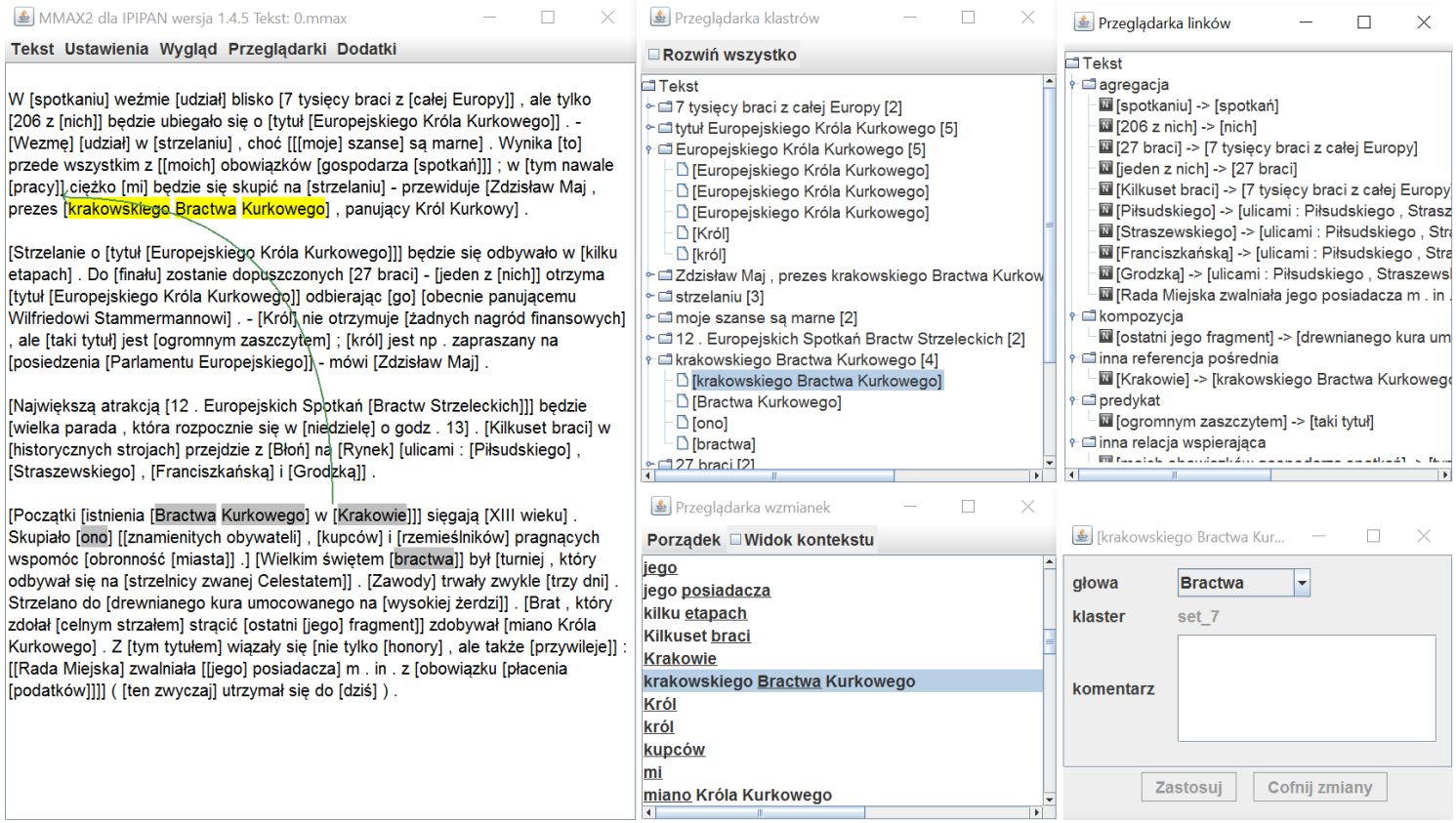
[18] DTTOOL (Aone i Bennett 1994), ALEMBIC (Day i in. 1998), MATE (McKelvie i in. 2001), MMAX (Müller i Strube 2001), COREFDRAW (Harabagiu i in. 2001), PALINKA (Orăsan 2003), WORDFREAK (Morton i LaCivita 2003), ACE TOOLS (Doddington i in. 2004), CALLISTO (Day i in. 2004), SERENGETI (Stührenberg i in. 2007), TRED (Pajas i Štěpánek 2008), UAMCORPUSTOOL (O'Donnell 2008), INFOREX (Marcińczuk i in. 2012), BRAT (Stenetorp i in. 2011, 2012) – ich porównanie zawiera rozdział 4.2 w książce (Ogrodniczuk i in. 2015).

tryb wspierający wyświetlanie całej próbki tekstu i najważniejszych składników anotacji na pełnym ekranie oraz zapewniający łatwą nawigację pomiędzy elementami okna programu. Wymaganiem postulowanym przez uczestników wcześniejszych projektów anotacyjnych była możliwość pracy w trybie offline (w odróżnieniu od interfejsu dostępnego wyłącznie z poziomu przeglądarki internetowej). W końcu liczyła się także stabilność i dojrzałość narzędzia, potwierdzona jego użyciem w zakończonych projektach anotacyjnych podobnego typu (np. Hodosh i in. 2010, Kunz 2010, Recasens i Martí 2010, Schäfer i in. 2012).

Program MMAX4REF składa się z kilku wzajemnie zależnych okien, z których najważniejsze prezentuje anotowany tekst i pozwala na oznaczanie jego fragmentów oraz przypisywanie im interpretacji lingwistycznych w zakresie zależności referencyjnych. Pozostałe okna prezentują szczegółowe informacje o relacjach zachodzących w tekście, aktualizowane na bieżąco i synchronizowane z oknem tekstu. Rysunek 4.1 przedstawia przykładowy wygląd okna anotacji. Z lewej strony ekranu wyświetlany jest tekst z oznaczonymi granicami wzmianek i ich ewentualnymi zagnieżdżeniami. Wskazanie wzmianki (tu: *krakowskiego Bractwa Kurkowego*) powoduje podświetlenie wzmianek z nią koreferencyjnych (*Bractwa Kurkowego*, *ono*, *bractwa*) zarówno w tekście, jak i w przeglądarce klastrów. Jeśli wzmiance przypisano jakieś relacje pośrednie, zostaną one oznaczone w tekście (strzałką łączącą ją z wzmianką zależną) i wyświetlone w przeglądarce linków. W podobny sposób wskazanie wzmianki w przeglądarce wzmianek, klastrów lub linków spowoduje wyróżnienie jej tekstowej reprezentacji i wyświetlenie związanych z nią zależności. Standardowy tryb anotacji umożliwia tworzenie wzmianek poprzez zaznaczanie fragmentów tekstu (także nieciągłych), wybór centrum semantycznego spośród słów tworzących wzmiankę, łączenie wzmianek w klastry z wyróżnieniem wyrażenia dominującego (patrz rozdział 4.3.2) i oznaczanie relacji pośrednich wraz z ich aspektami.

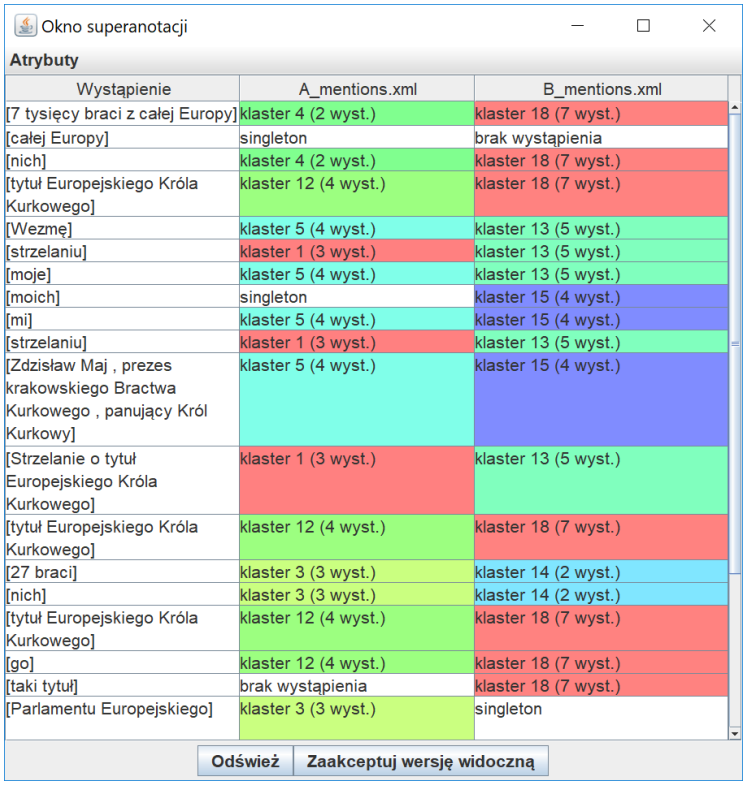
Tryb superanotacyjny (patrz rys. 4.2 i 4.3) umożliwia ponadto wyświetlenie różnic w wariantach anotacji danego tekstu, tworzonych przez kilku anotatorów (na poziomie wzmianek, ich centrów semantycznych, klastrów koreferencyjnych, linków relacji pośrednich i pomocniczych oraz aspektów), a następnie wybór właściwej anotacji kliknięciem. Jeżeli podczas ujednolicania anotacji żadna z wersji nie może być uznana za poprawną, superanotator może dokonać zmian w trybie edycji w głównym oknie programu.

Program anotacyjny zawiera także szereg dodatkowych ułatwień w pracy anotatora, takich jak: możliwość kopiowania fragmentów tekstu do schowka, cofania



Rysunek 4.1. Okna anotacji w programie MMAX4REF

Rysunek 4.2. Tryb superanotacji programu MMAX4REF – okno wzmianek (3 anotatorów)



Rysunek 4.3. Tryb superanotacji programu MMAX4REF – okno klastrów (2 anotatorów)

zmian, automatyczny zapis stanu pracy czy możliwość dostosowania interfejsu do indywidualnych potrzeb (zmiana wielkości czcionki, zapamiętywanie rozmiaru i położenia okien).

4.5. Zgodność anotatorów

Panuje dość powszechna opinia (np. Recasens 2010, Zikánová i in. 2015: rozdział 2.7), że mimo zapewnienia wyczerpującej i przetestowanej na wstępnym etapie prac instrukcji anotacyjnej decyzje anotatorów w zadaniach anotacji semantycznej i pragmatycznej cechuje duży poziom subiektywizmu, co przekłada się na niższe niż w przypadku prostszych zadań wyniki zgodności anotacji. Efekt złożoności zadania dodatkowo wzmacnia zależność pomiędzy jednostkowymi decyzjami, szczególnie widoczna w przypadku koreferencji, gdzie decyzja o włączeniu wzmianki do danego klastra ma konsekwencje dla zawartości pozostałych klastrów. Trudno zatem podać odpowiedni poziom pożądanej zgodności anotatorów, zwłaszcza w procesie dekodowania koreferencji, jednak uzyskany wynik wydaje się być dobrą wskazówką do oceny jakości implementowanych systemów, wyznaczając górną granicę możliwości narzędzia automatycznego.

Dalej podaję wartości zgodności dla różnych elementów procesu anotacyjnego obliczone dla wszystkich tekstów anotowanych niezależnie przez trzech anotatorów, czyli niemal pełnego zasobu korpusowego z wyłączeniem 100 tekstów anotowanych przez 5 anotatorów na potrzeby badań nad eksperymentalnymi strategiami superanotacji automatycznej (patrz rozdział 4.2.4).

4.5.1. Wzmianki

Ocena zgodności anotacji wzmianek w określonym zestawie danych wymaga podania co najmniej dwóch wartości liczbowych, odpowiadających zgodności zestawów wzmianek z dokładnością do centrum wzmianki oraz z uwzględnieniem pełnych granic wzmianek. W związku z trudnością oceny wpływu czynnika losowego na oznaczanie wzmianek jako takich, ograniczam się do podania wartości wskaźnika zaobserwowanej zgodności (liczby wzmianek wspólnych dla wszystkich anotatorów w stosunku do łącznej liczby wzmianek wskazanych przez dowolnego anotatora). Wartości te wynoszą w zbiorze tekstów krótkich odpowiednio 89,94% dla centrów i 81,59% dla dokładnych granic, w zbiorze tekstów długich zaś odpowiednio 89,84% i 81,64%.

W przypadku wzmianek o wspólnych granicach możemy dodatkowo obliczyć średnią zgodność wyboru centrów semantycznych w ramach odpowiadających sobie wzmianek, stosując wzór Benneta i in. (1954), uwzględniający już efekt przypadkowego wyboru:

$$S = \frac{p_{A_O} - p_{A_E}}{1 - p_{A_E}}$$

gdzie p_{A_O} oznacza zgodność zaobserwowaną (stosunek liczby wzmianek o tym samym centrum semantycznym do łącznej liczby wzmianek), p_{A_E} zaś – zgodność przypadkową (średnie prawdopodobieństwo wyboru danego segmentu jako centrum wzmianki, zależne od długości wzmianki i wynoszące 1 dla singletonów).

Zgodność S obliczona dla zbioru tekstów krótkich wynosi 99,73%, dla tekstów długich zaś – 99,98%.

4.5.2. Klastry koreferencyjne

Zgodność anotacji klastrów koreferencyjnych mierzona jest za pomocą wartości κ (Fleiss 1971) uwzględniającej czynnik przypadku. Wyniki tych obliczeń z podziałem na typy tekstów przedstawiono w tabeli 4.4.

Obserwowana zgodność anotacji klastrów, z uwzględnieniem singletonów (liczby klastrów wspólnych dla wszystkich anotatorów w stosunku do łącznej liczby klastrów wskazanych przez dowolnego anotatora), wynosi dla tekstów krótkich 78,52%, dla tekstów długich zaś – 76,72%. Zgodność anotacji wyrażeń dominujących obliczona dla niesingletonowych klastrów wspólnych dla wszystkich anotacji wynosi 87,95% dla tekstów krótkich i 84,23% dla tekstów długich (uwzględnienie czynnika losowego nie jest możliwe, gdyż anotator może podać dowolny tekst wyrażenia dominującego opisującego klaster, a nie wyłącznie dokonać wyboru wzmianki dominującej).

Uzyskane wartości odpowiadają znacznej zgodności, są porównywalne z podawanymi dla innych języków (np. Pradhan i in. 2012, Zikánová i in. 2015) i nie wykazują istotnych zależności od typu ani długości tekstu. Warto jedynie zwrócić uwagę na zaskakująco wysoką zgodność anotacji dla tekstów konwersacyjnych, internetowych interaktywnych i quasi-mówionych wynikającą z dostępności w tego rodzaju tekstach oznaczeń mówców, klastrowalnych w oczywisty sposób (na bazie pełnej zgodności form tekstowych wzmianek).

Tabela 4.4. Zgodność κ anotacji klastrów koreferencyjnych

Kategoria tekstu	κ
Teksty krótkie	0,8883
Literatura piękna	0,9343
Mówione konwersacyjne	0,9304
Internetowe interaktywne (blogi, fora, usenet)	0,9189
Inne teksty pisane	0,9142
Książka niebeletrystyczna nieklasyfikowana	0,9080
Mówione medialne	0,9024
Literatura faktu	0,8995
Quasi-mówione (protokoły sesji parlamentu)	0,8993
Internetowe nieinteraktywne (statyczne strony, Wikipedia)	0,8836
Dzienniki	0,8765
Pozostałe periodyki	0,8628
Typ informacyjno-poradnikowy	0,8551
Książki publicystyczne	0,8422
Typ naukowo-dydaktyczny	0,8402
Teksty długie	0,8570
Kraj	0,9438
Sport	0,9070
Kultura	0,8984
Nauka i technika	0,8561
Publicystyka/opinie	0,8369
Prawo	0,7916
Ekonomia	0,7651

4.5.3. Pozostałe relacje

Zgodność anotacji relacji pośrednich, pomocniczych i aspektów również mierzono, obliczając κ Fleissa dla relacji danego rodzaju, zliczając linki wskazane we wszystkich anotacjach, z uwzględnieniem ich kierunkowości i bez uwzględniania wartości aspektu. Biorąc pod uwagę sposób prowadzenia anotacji (polegający na łączeniu wzmianek i ekstrapolowaniu tej decyzji na klastry), wartości te podajemy w dwóch wariantach, dla linków między wzmiankami oraz relacji między klastrami. Wyniki obliczeń, tym razem wyłącznie dla tekstów krótkich ze względu na znacznie mniejszą liczbę linków tego rodzaju w tekstach długich, przedstawiono w tabeli 4.5.

Tabela 4.5. Zgodności κ anotacji relacji pośrednich i pomocniczych

Rodzaj relacji	**Liczba linków między wzmiankami**	κ	**Liczba relacji między klastrami**	κ
Relacja pośrednia				
Agregacja	14 744	0,2390	7 472	0,3336
Kompozycja	5 935	0,2808	3 906	0,4017
Anafora związana	595	0,7495	319	0,8545
Inna relacja pośrednia	5 626	0,1459	3 509	0,2666
Relacja wspierająca				
Metareferencja	422	0,8617	220	0,9271
Porównanie	408	0,8540	245	0,9079
Predykat	3 045	0,4179	1 634	0,5989
Inna relacja wspierająca	4 080	0,2954	2 523	0,4406
Relacja wykluczająca				
Kontrast	2 636	0,3832	1 808	0,4841
Kategorialność	1 349	0,5920	745	0,7201
Polisemia	531	0,8809	223	0,9675
Inna relacja wykluczająca	206	0,9014	130	0,9362

Stosunkowo niska zgodność relacji najmniej kontrowersyjnej kategorii asocjacji strukturalnej (agregacji i kompozycji) wynika z dużej liczby linków tego rodzaju w porównaniu z liczbą linków dla pozostałych relacji i przyjęcia założenia o pełnej zgodności tekstów w ogóle niezawierających danej relacji. Warto jednak zauważyć, że w przypadku relacji asocjacyjnych łączenie elementów tekstu jest praktycznie nieograniczone, na co wskazują także inni badacze (por. np. Zikánová i in. 2015: s. 237; patrz także rozdział 6.4.3), a co powoduje dużą wariancję obserwowanych wyników. Anotację aspektów cechuje bardzo niska zgodność (0,0202), co potwierdza trudności związane z interpretacją zjawisk w rodzaju rozmycia konceptualnego (patrz rozdział 2.1).

Podane wartości są dodatkowo trudno porównywalne z wynikami uzyskiwanymi dla innych języków, gdyż praktycznie każdy projekt anotacyjny stosuje własny schemat opisu relacji asocjacyjnych, a dodatkowo niezwykle rzadka jest możliwość porównania wyników pracy więcej niż 2 anotatorów. Można jednak przyjąć, że zgodność na poziomie 0,3–0,4 może być uznana za typową (por. np. wyniki dla

holenderskiego korpusu COREA, Hendrickx i in. 2011), co pozwala z optymizmem patrzeć na wyniki uzyskane w ramach niniejszej pracy.

4.6. Korekta błędów

Według Fort i Sagota (2010: s. 23) większość projektów anotacyjnych cierpi na brak środków do ręcznej całościowej korekty korpusu i stosuje automatyzację korekty na podstawie wskazówek zebranych w fazie anotacji; gdy błędy są systemowe, proces nie wymaga udziału eksperta. Nasz przypadek jest nieco inny – w związku z dwiema fazami anotacji relacji referencyjnych realizowanymi w dwóch następujących po sobie projektach oraz niezależnym zadaniem anotacji relacji dyskursywnych (patrz rozdział 7.3) na tym samym materiale tekstowym udało się wykryć część błędów w kolejnych przebiegach anotacji ręcznej; etap analizy automatycznej był jedynie jej uzupełnieniem.

Dzięki anotacji wieloprzebiegowej poprawione zostały błędy trudne do wykrycia metodami automatycznymi, takie jak np.: obecność śródtytułów umieszczonych na końcu wylosowanych fragmentów czy tekstów urwanych lub błędnie wyekstrahowanych i w ten sposób zaburzających strukturę próbki. W przypadku gdy wprowadzane poprawki wymagały podziału tekstu na części (np. z powodu przeoczenia włączenia do korpusu tekstu sklejonego z kilku krótkich notek prasowych, zaanotowanego w standardowy sposób, a jednocześnie ewidentnie niespójnego, z wyróżniającymi się osobnymi częściami), teksty podzielone nie były usuwane z korpusu, a jedynie dodatkowo oznaczane. W ten sposób powstał w pełni anotowany podkorpus 62 „mikrotekstów", udostępniany wraz z podstawową zawartością korpusu, natomiast nieuwzględniany w dalszych analizach statystycznych korpusu (oraz niewykazany w tabeli 4.1) ze względu na to, że teksty tego rodzaju nie spełniają przyjętego kryterium rozmiarowego. Dla pełności wywodu, w tabeli 4.6 podsumowano podstawowe własności tego zbioru.

W ramach pojedynczych próbek podjęto decyzję, żeby poprawiać jedynie ewidentne błędy konwersji/filtrowania tekstów wprowadzone na etapie zbierania materiału korpusowego oraz w anotacji wytworzonej w trakcie naszych prac. Tekstów mówionych nie poprawiano w ogóle.

Błędy wykryte automatycznie dotyczyły głównie strukturalnych własności korpusu, takich jak: obecność pustych linków, podwójnie oznaczonych granic wzmianek, klastrów zawierających pojedyncze wzmianki, linków prowadzących do nieistniejących wzmianek czy niezgodności liczby aspektów z liczbą odpowiadających

Tabela 4.6. Struktura podkorpusu „mikrotekstów”

Kategoria	**Liczba tekstów**	**Liczba segmentów**
Dzienniki	40	4 832
Pozostałe periodyki	11	964
Literatura piękna	4	637
Typ informacyjno-poradnikowy	2	131
Literatura faktu	2	350
Książka niebeletrystyczna	2	235
Internetowe nieinteraktywne	1	40
Łącznie	62	7 189

im relacji. Poprawki techniczne miały na celu wskazanie pominiętych grup dominujących, wybór nieustalonych centrów semantycznych, uspójnienie numeracji słów i wzmianek oraz korektę błędów w strukturach niepodlegających anotacji, wprowadzonych przez używane narzędzia automatyczne na wczesnym poziomie opracowania tekstów (takich jak znaki interpunkcyjne włączone w treść wzmianek). Nieliczne usterki (nadmiarowe łamanie wierszy w tekstach, nadmiarowe dywizy w treści słów, znaki spoza zestawu liter ASCII) okazały się wynikiem błędów konwersji i także zostały poprawione.

4.7. Udostępnienie korpusu

W związku z przyjęciem zasady samodzielności tworzonego korpusu, teksty źródłowe (w tym przypadku pochodzące z Narodowego Korpusu Języka Polskiego) zostały skopiowane, a nie jedynie powiązane z korpusem bazowym, z anotacją zewnętrzną (ang. *stand-off*) zastosowaną na poziomach reprezentacji lingwistycznej. W ten sposób korpus zależności referencyjnych może być używany niezależnie od NKJP.

Dane korpusowe są dostępne w kilku opisanych dalej standardowych formatach, z czego trzy mają charakter roboczy: formaty CoNLL/SemEval i MMAX były wykorzystywane na różnych etapach anotacji, format narzędzia BRAT[19] (Stenetorp i in. 2011, 2012) zaś został użyty do wizualizacji wyników. Podstawowym formatem wynikowym jest natomiast XML-owy format TEI P5. Dane we

[19]Dostępne na stronie `http://brat.nlplab.org`.

wszystkich wymienionych formatach dostępne są do pobrania na licencji Creative Commons – Uznanie autorstwa 4.0 pod adresem `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`. Korpus został ponadto udostępniony do przeglądania (patrz rozdział 4.7.4) oraz przeszukiwania (patrz rozdział 4.7.5).

4.7.1. Format SemEval/CoNLL

Format SemEval/CoNLL jest referencyjnym formatem reprezentacji danych wykorzystywanym podczas konkursów narzędzi do wykrywania koreferencji w serii warsztatów ewaluacji semantycznej (Recasens i Hovy 2011). Dane są zapisane w tekstowym formacie wielokolumnowym w sposób zgodny z zasadą przyjętą dla wcześniej organizowanych zadań z dziedziny wykrywania zależności składniowych i semantycznych w ramach konferencji CoNLL[20] (ang. *Conference on Computational Natural Language Learning*): każda linia dokumentu odpowiada jednemu segmentowi reprezentowanego tekstu, opisanemu lingwistycznie przez umieszczenie wartości wybranych kategorii w kolejnych kolumnach oddzielanych znakiem tabulacji – patrz rysunek 4.4.

Pierwsze sześć kolumn zawiera odpowiednio: formę tekstową segmentu, odpowiadający mu lemat, oznaczenie klasy gramatycznej oraz, wyłącznie dla segmentów nominalnych, wartości kategorii przypadka, rodzaju i liczby. Przedostatnia kolumna zawiera informację o centrum semantycznym wzmianek wielosegmentowych w postaci numeru klastra wzmianki umieszczonego w linii zawierającej segment centrum. Kolumna ostatnia zawiera informację o wzmiankach i klastrach w postaci nawiasowej: okrągły nawias otwierający z liczbą określa początek wzmianki, nawias zamykający z tą samą liczbą – koniec wzmianki. Jeśli wzmianka ma postać jednego segmentu, opis ogranicza się do pojedynczej linii. Informacja o początku lub końcu wielu wzmianek w danej linii oznaczona jest znakiem kreski pionowej oddzielającej oznaczenia nawiasowe. Wzmianki należące do tego samego klastra oznaczone są tą samą liczbą. Przykładowo fraza *sypialni emira* z rysunku 4.4 należy do klastra nr 2 (wraz ze wzmianką *ją*), a jej centrum semantycznym jest segment *sypialni*.

Format SemEval/CoNLL był używany wyłącznie podczas wstępnych prac anotacyjnych (patrz rozdział 4.3.1), jednak ze względu na jego prostotę i dostępność posługującego się właśnie nim oficjalnego narzędzia ewaluacyjnego dla zadania

[20] `http://www.clips.uantwerpen.be/conll2008/`.

Forma tekstowa segmentu	Lemat	Klasa grama-tyczna	Przy-padek	Liczba	Rodzaj	Centrum semantyczne wzmianki	Wzmianka/ klaster
Stare	stary	adj	nom	pl	f		(1
Amerykanki	Amerykanka	subst	nom	pl	f	head:1	1)
biegają	biegać	fin	-	pl	-		
po	po	prep	-	-	-		
sypialni	sypialnia	subst	loc	sg	f	head:2	(2
emira	emir	subst	gen	sg	m1		(3)|2)
,	,	interp	-	-	-		
[]	[]	ppron3	nom	pl	f		(1)
fotografują	fotografować	fin	-	pl	-		
ją	on	ppron3	acc	sg	f		(2)
,	,	interp	-	-	-		
[]	[]	ppron3	nom	pl	f		(1)
zaglądają	zaglądać	fin	-	pl	-		
do	do	prep	-	-	-		
starego	stary	adj	gen	sg	n		(4|(5
łóżka	łóżko	subst	gen	sg	n	head:4,5	5)
i	i	conj	-	-	-		
szaf	szafa	subst	gen	pl	f		(6)|4)
.	.	interp	-	-	-		
[]	[]	ppron3	nom	pl	f		(1)
Są	być	fin	-	pl	-		
przejęte	przejęty	adj	nom	pl	f		
.	.	interp	-	-	-		

Rysunek 4.4. Wyniki anotacji testowej w formacie SemEval/CoNLL

wykrywania koreferencji[21] (ang. *scorer*) planowane jest jego udostępnienie jako jednego z formatów korpusu oraz jako formatu wynikowego narzędzi do wykrywania koreferencji dostępnych w Multiserwisie[22] – aplikacji webowej grupującej serwisy sieciowe do przetwarzania języka polskiego.

4.7.2. Format MMAX

Kolejną dostępną reprezentacją korpusu jest format programu anotacyjnego MMAX4REF, rozszerzający XML-owy schemat narzędzia bazowego MMAX2[23] o dodatkowe atrybuty klastrów i relacji pośrednich. Próbka korpusowa składa się z trzech plików (n jest identyfikatorem próbki):

- n`.mmax` – plik główny, zawierający metadane tekstu;
- n`_words.xml` – poziom segmentacji tekstu i informacji morfoskładniowej;
- n`_mentions.xml` – poziom wzmianek i relacji referencyjnych.

Plik główny (patrz rys. 4.5) zawiera: link do poziomu segmentacyjnego, tytuł tekstu uzupełniony o informację na temat jego powiązania z plikiem źródłowym NKJP oraz kategorię tematyczną tekstu.

```
<mmax_project>
  <words>168_words.xml</words>
  <title val="Zapis świata. Traktat metafizyczny (akapity p-57,
              p-58, p-59, p-60 z tekstu IPIPAN_1301919980826
              w Narodowym Korpusie Języka Polskiego - podkorpus
              300-milionowy)"/>
  <catRef val="Dzienniki"/>
</mmax_project>
```

Rysunek 4.5. Reprezentacja metadanych próbki w formacie MMAX

Plik segmentacyjno-morfoskładniowy (patrz rys. 4.6) ma postać ciągu elementów `<word>`, w których treści znajdują się słowa (w tym wariancie już bez podziału na

[21] `http://conll.github.io/reference-coreference-scorers/`.
[22] Dostępna na stronie `http://multiservice.nlp.ipipan.waw.pl`.
[23] Opisany okładniej w dokumentacji dostępnej na stronie `http://mmax2.net`.

segmenty) i znaki interpunkcyjne, w atrybutach zaś – informacje o lemacie (`base`), klasie fleksyjnej (`ctag`) i wartościach kategorii morfoskładniowych (`msd`). Podział na akapity i zdania wyznaczają atrybuty `lastinpar` i `lastinsent` o wartości `true` odpowiednio dla ostatniego słowa w akapicie i zdaniu. Przejęty z NKJP atrybut `hasNps` o wartości `true` reprezentuje brak odstępu pomiędzy dwoma elementami `<word>` (słowem a znakiem interpunkcyjnym).

Plik zawierający opis wzmianek i referencji (patrz rys. 4.7) ma postać ciągu elementów `<markable>` odwołujących się do słów z poziomu segmentacji w atrybucie `span`. Wartością tego atrybutu może być:

- pojedynczy identyfikator (dla naszego przykładu np. `word_2`) oznaczający wzmiankę będącą pojedynczym słowem (*„spotkaniu”*),
- para identyfikatorów przedzielonych dwiema kropkami (np. `word_10..word_11`) oznaczająca wzmiankę złożoną z ciągu słów (*„całej Europy”*),
- ciąg oddzielanych przecinkami identyfikatorów lub ich par tworzących przedział (np. `word_44,word_48..word_51`) oznaczający wzmiankę nieciągłą.

W atrybutach zapisana jest również informacja o głowie semantycznej wzmianki (atrybut `mention_head`) oraz o relacjach referencyjnych. Wzmianki połączone relacjami koreferencji oznaczone są tą samą wartością atrybutu `mention_group` (`set_14` dla wzmianek *7 tysięcy braci z całej Europy* oraz *nich*). Każda wzmianka danego klastra oznaczona jest też (nadmiarowo) atrybutem `dominant` o wartości wyrażenia dominującego. Linki relacji pośrednich i pomocniczych reprezentowane są w elementach odpowiadających wzmiankom stanowiącym źródło linku poprzez podanie identyfikatora wzmianki będącej celem linku w wartości atrybutu *`nazwa_relacji`* (w tym przykładzie wzmianka *7 tysięcy braci z całej Europy* opisana jest jako pozostająca w relacji agregacji ze wzmianką *206 z nich*)[24]. Atrybuty aspektów relacji mają nazwy postaci *`nazwa_relacji`*`_facet`, ich wartością jest zaś lista rozdzielonych średnikami angielskich nazw aspektów (*dissimilation*, *uncertainty*, *opinion*).

[24]Pełną listę nazw atrybutów odpowiadających typom z taksonomii relacji referencyjnych zawiera instrukcja do wyszukiwarki korpusowej (`http://pcc.nlp.ipipan.waw.pl/manual`).

```
<words>
  <word id="word_1" base="w" ctag="prep"
        msd="loc:nwok">W</word>
  <word id="word_2" base="spotkanie" ctag="subst"
        msd="sg:loc:n">spotkaniu</word>
  <word id="word_3" base="wziąć" ctag="fin"
        msd="sg:ter:perf">weźmie</word>
  <word id="word_4" base="udział" ctag="subst"
        msd="sg:acc:m3">udział</word>
  <word id="word_5" base="blisko" ctag="adv"
        msd="pos">blisko</word>
  <word id="word_6" base="7" ctag="ign"
        msd="">7</word>
  <word id="word_7" base="tysiąc" ctag="subst"
        msd="pl:gen:m3">tysięcy</word>
  <word id="word_8" base="brat" ctag="subst"
        msd="pl:gen:m1">braci</word>
  <word id="word_9" base="z" ctag="prep"
        msd="gen:nwok">z</word>
  <word id="word_10" base="cały" ctag="adj"
        msd="sg:gen:f:pos">całej</word>
  <word id="word_11" base="Europa" ctag="subst"
        msd="sg:gen:f">Europy</word>
  <word id="word_12" base="," ctag="interp"
        msd="" hasNps="true">,</word>
  <word id="word_13" base="ale" ctag="conj"
        msd="">ale</word>
  <word id="word_14" base="tylko" ctag="qub"
        msd="">tylko</word>
  <word id="word_15" base="206" ctag="ign"
        msd="">206</word>
  <word id="word_16" base="z" ctag="prep"
        msd="gen:nwok">z</word>
  <word id="word_17" base="on" ctag="ppron3"
        msd="pl:gen:m2:ter:akc:praep">nich</word>
  ...
</words>
```

Rysunek 4.6. Reprezentacja informacji segmentacyjno-morfoskładniowych w formacie MMAX

```
<markables>
  <markable id="markable_1" span="word_2"
            mention_head="spotkaniu"/>
  <markable id="markable_2" span="word_4"
            mention_head="udział"/>
  <markable id="markable_4" span="word_5..word_11"
            mention_head="braci" mention_group="set_14"
            dominant="7 tysięcy braci z całej Europy"/>
  <markable id="markable_5" span="word_10..word_11"
            mention_head="Europy"/>
  <markable id="markable_6" span="word_15..word_17"
            mention_head="nich" indirect_aggregation="markable_4" />
  <markable id="markable_7" span="word_17"
            mention_head="nich" mention_group="set_14"
            dominant="7 tysięcy braci z całej Europy"/>
  ...
</markables>
```

Rysunek 4.7. Uproszczona reprezentacja relacji referencyjnych w formacie MMAX

4.7.3. Format TEI

Format TEI (ang. *Text Encoding Intiative*; Burnard i Bauman 2007) jest popularnym standardem reprezentacji danych w dziedzinie nauk humanistycznych, jego dostosowanie do potrzeb Narodowego Korpusu Języka Polskiego (Bański i Przepiórkowski 2009) umożliwia zaś zapis analiz lingwistycznych na wielu poziomach. Podczas budowy obecnego korpusu format ten został dostosowany oraz rozszerzony o zapis informacji o relacjach referencyjnych.

Zastosowanie anotacji zewnętrznej wymaga podjęcia decyzji, czy oprócz całego korpusu również pliki pojedynczych próbek (a może nawet indywidualnych poziomów anotacji) powinny być dostępne samodzielnie. W NKJP przyjęto zasadę częściowej niezależności poszczególnych poziomów anotacji, co umożliwia korzystanie z korpusu w sposób selektywny, poprzez wyłączanie warstw wyższych, ale powoduje też kilka niedogodności. Mimo użycia elementu `<teiCorpus>` już na poziomie pojedynczej próbki nie istnieje ani plik opisujący tę próbkę, ani plik opisujący cały korpus. Brak opisu próbki wymaga przeciążania opisu nagłówkowego i używania go we wszystkich plikach warstw opisu lingwistycznego, a jedyną cechą łączącą

zestaw plików w próbkę jest ich zapis w tym samym katalogu systemu plików. Mimo wszystko zrozumiała chęć zachowania zgodności z NKJP oraz dostępność API programistycznego do przetwarzania korpusu w tym formacie sprawiły, że podjęto decyzję, żeby nie ingerować w schemat opisu NKJP. Próbka korpusu jest zatem zestawem następujących plików (niektóre występują w postaci spakowanej):

- `header.xml` – nagłówek z metadanymi próbki: tytułem tekstu, tekstową informacją o powiązaniu z tekstem źródłowym z NKJP oraz klasyfikatorem tematyczno-gatunkowym tekstu;
- `text.xml` – tekst właściwy, z podziałem na akapity;
- `ann_segmentation.xml` – podział tekstu na zdania i segmenty;
- `ann_morphosyntax.xml` – interpretacje morfoskładniowe poszczególnych segmentów;
- `ann_mentions.xml` – informacja o wzmiankach i ich własnościach;
- `ann_coreference.xml` – informacja o zależnościach referencyjnych.

Każdy z tych plików zawiera deklarację przestrzeni nazw TEI (`http://www.tei-c.org/ns/1.0`) i NKJP (`http://www.nkjp.pl/ns/1.0`), z których pochodzą wszystkie elementy użyte w dokumencie. Każdy może też funkcjonować samodzielnie jako zgodny z TEI. Dokładne informacje na temat elementów nagłówka korpusu oraz szczegółów opisu anotacji lingwistycznej w zakresie segmentacji i morfoskładni zawiera książka projektowa NKJP (Przepiórkowski i in. 2012: rozdział 10), dalej opisuję jedynie warstwy wprowadzone na potrzeby reprezentacji własności referencyjnych[25]. Zapis ten zasadniczo nie jest przeznaczony do czytania przez człowieka, jednak w celach kontrolnych elementy odpowiadające reprezentowanym konstrukcjom zostały dodatkowo opatrzone komentarzami XML-owymi przytaczającymi tekstową postać tych struktur (interpretowanych segmentów, wzmianek, klastrów i linków).

Wzmianki

Ze względu na możliwość wystąpienia rozległych wzmianek wielozdaniowych (i potencjalnie również wieloakapitowych) zbiór wzmianek (patrz rys. 4.8) inaczej niż w opisie pozostałych warstw w NKJP nie powiela ogólnej struktury pliku

[25] Patrz także rozdział 8.2.1 książki (Ogrodniczuk i in. 2015).

źródłowego z podziałem na akapity i zdania, lecz zawiera się w pojedynczym elemencie blokowym dodanym ze względu na konieczność zachowania zgodności z TEI. Wzmianki segmentami (`<seg>`) odwołującymi się (za pomocą znaczników `<ptr>`) do odpowiadających im elementów z warstwy segmentacji (wskazywanych metodą linku XML-owego w atrybucie `target`). Segment, będący głową semantyczną wzmianki, oznaczony jest atrybutem `type` o wartości `semh`.

Zależności referencyjne

Informacja o relacjach referencyjnych (patrz rys. 4.9) zapisana jest również w postaci segmentów odwołujących się w treści do wzmianek będących w danej relacji; każde takie odwołanie jest elementem `<ptr>` posiadającym w atrybucie `target` identyfikator XML-owy segmentu odpowiadającemu przywołanej wzmiance. Typ relacji podany jest jako wartość cechy `type` za pomocą mechanizmu struktur cech (elementy `<f>` i `<fs>`). Dla relacji bezpośrednich struktura zawiera dodatkową informację o wzmiance dominującej, odwołania do wzmianek wskazują zaś elementy klastra. Dla pozostałych relacji wskazany jest kierunek linku, natomiast liczba odwołań do wzmianek ograniczona jest do dwóch.

4.7.4. Format narzędzia BRAT i wersja online korpusu

BRAT jest środowiskiem anotacyjno-prezentacyjnym, które zostało zaadaptowane do wizualizacji korpusu w wersji do przeglądania w internecie[26]. Zmianie uległ m.in. sposób prezentacji klastrów koreferencyjnych poprawiający czytelność tekstu zawierającego długie łańcuchy powiązań[27]. Zmodyfikowana wersja narzędzia pod nazwą BRAT4REF jest dostępna pod adresem `http://zil.ipipan.waw.pl/brat4ref`; na rysunku 4.10 przedstawiono wizualizację przykładowego tekstu.

Użycie wersji online jest stosunkowo intuicyjne: po wyborze tekstu z listy zostaje on otwarty w głównym oknie przeglądarki. Wzmianki oznaczone są etykietami MENTION (lub MEN/M w przypadku wzmianek krótszych niż oznaczająca je etykieta). Wskazanie wzmianki kursorem myszy powoduje wyświetlenie jej granic, własności oraz pozostałych wzmianek należących do tego samego klastra (oznaczanych na zielono), a także innych relacji, w których funkcjonuje dana wzmianka (oznaczanych innymi kolorami). Dla wskazanej wzmianki prezentowane są także

[26] Dostępna na stronie `http://cothec.nlp.ipipan.waw.pl/`.

[27] Oryginalną wersję prezentacji linków koreferencyjnych można zobaczyć podczas prezentacji warstwy koreferencyjnej w Multiserwisie (`http://multiservice.nlp.ipipan.waw.pl/pl/`).

```
<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="PCC_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <p>

          <!-- spotkaniu -->
            <seg xml:id="mention_1">
              <fs type="mention">
                <f name="semh"
                   fVal="ann_morphosyntax.xml#morph_1.1.2-seg"/>
              </fs>
              <ptr target="ann_morphosyntax.xml#morph_1.1.2-seg"/>
            </seg>

            <!-- 7 tysięcy braci z całej Europy -->
            <seg xml:id="mention_3">
              <fs type="mention">
                <f name="semh"
                   fVal="ann_morphosyntax.xml#morph_1.1.8-seg"/>
              </fs>
              <ptr target="ann_morphosyntax.xml#morph_1.1.6-seg"/>
              <ptr target="ann_morphosyntax.xml#morph_1.1.7-seg"/>
              <ptr target="ann_morphosyntax.xml#morph_1.1.8-seg"/>
              <ptr target="ann_morphosyntax.xml#morph_1.1.9-seg"/>
              <ptr target="ann_morphosyntax.xml#morph_1.1.10-seg"/>
              <ptr target="ann_morphosyntax.xml#morph_1.1.11-seg"/>
            </seg>

            ...
        </p>
      </body>
    </text>
  </TEI>
</teiCorpus>
```

Rysunek 4.8. Reprezentacja wzmianek w formacie TEI

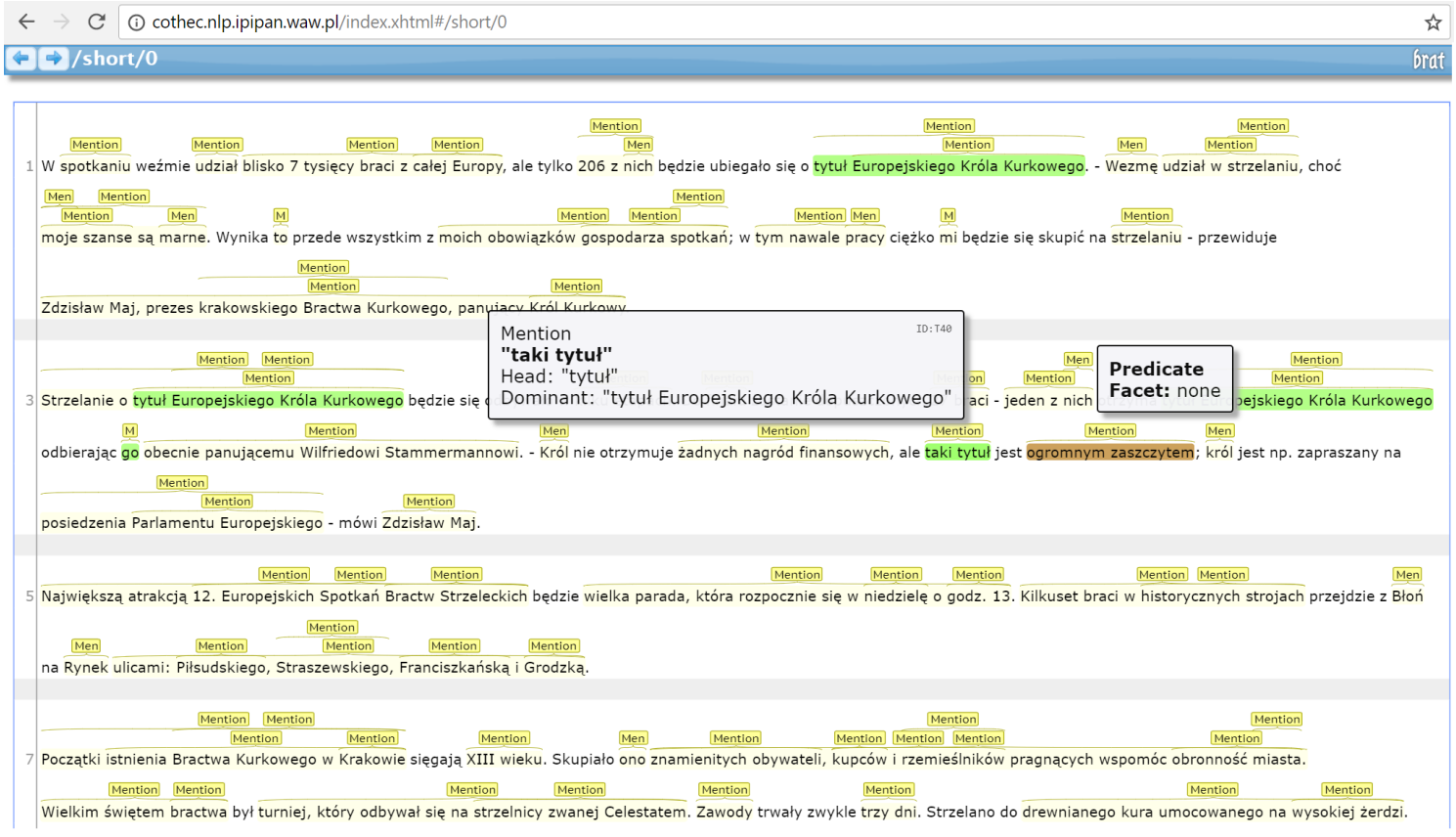
```
<?xml version="1.0" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
           xmlns:xi="http://www.w3.org/2001/XInclude"
           xmlns:nkjp="http://www.nkjp.pl/ns/1.0">
  <xi:include href="PCC_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <p>

        <!-- spotkaniu; spotkań -->
          <seg xml:id="coreference_1">
            <fs type="coreference">
              <f name="type" fVal="indirect_aggregation"/>
            </fs>
            <ptr type="source" target="ann_mentions.xml#mention_1"/>
            <ptr target="ann_mentions.xml#mention_19"/>
          </seg>

          <!-- 7 tysięcy braci z całej Europy; nich -->
          <seg xml:id="coreference_2">
            <fs type="coreference">
              <f name="type" fVal="ident"/>
              <f name="dominant"
                 fVal="7 tysięcy braci z całej Europy"/>
            </fs>
            <ptr target="ann_mentions.xml#mention_3"/>
            <ptr target="ann_mentions.xml#mention_6"/>
          </seg>

          ...
        </p>
      </body>
    </text>
  </TEI>
</teiCorpus>
```

Rysunek 4.9. Reprezentacja zależności referencyjnych w formacie TEI



Rysunek 4.10. Wizualizacja korpusu w środowisku graficznym BRAT

```
T0  Mention 2 11   spotkaniu
#0  empty T0       Head: "spotkaniu"
T1  Mention 19 25  udział
#1  empty T1       Head: "udział"
T2  Mention 26 63  7 tysięcy braci z całej Europy
#2  empty T2       Head: "braci"
#3  empty T2       Dominant: "7 tysięcy braci z całej Europy"
T4  Mention 75 85  206 z nich
...
R0  indirect_aggregation Arg1:T4 Arg2:T2 none
*   Coref T2 T5
R1  supporting_predicative Arg1:T46 Arg2:T49 opinion
*   Coref T24 T30 T32 T36 T6
*   Coref T31 T34 T38 T7
*   Coref T18 T8
...
```

Rysunek 4.11. Reprezentacja zależności referencyjnych w formacie BRAT

jej własności: centrum, wyrażenie dominujące klastra, typy i aspekty relacji wiążących wzmiankę z innymi fragmentami tekstu. Wersja online umożliwia także proste przeszukiwanie korpusu.

Powstała także wersja korpusu do pobrania, którą można w prosty sposób wykorzystać na przykład do samodzielnej wizualizacji fragmentu korpusu na własnym stanowisku roboczym. Format BRAT[28] składa się z dwóch plików: z czystym tekstem próbki (z rozszerzeniem `.txt`) oraz z informacją o relacjach zachodzących pomiędzy fragmentami tekstu (z rozszerzeniem `.ann`) definiowanymi za pomocą indeksów znaków. Opis relacji ma postać tekstową definiującą poszczególne wzmianki (oznaczone symbolem `T`), relacje typu identycznościowego (`* Coref`) – lub pośredniego (`R` z angielską nazwą relacji i oznaczeniem aspektu na końcu linii), centra semantyczne (`Head`) i wyrażenia dominujące (`Dominant`) – patrz rysunek 4.11.

[28] Opisany dokładniej na stronie narzędzia: `http://brat.nlplab.org/standoff.html`.

4.7.5. Wyszukiwarka korpusowa

MTAS (Brouwer i in. 2017) jest stabilną i szybką wyszukiwarką korpusową, która staje się powoli standardowym interfejsem wyszukiwawczym w korpusach tekstów polskich z wielowarstwowym opisem lingwistycznym[29]. Udostępnienie korpusu w tej konfiguracji[30] zapewnia bogate funkcje przeszukiwawcze, z możliwością użycia w treści zapytań wyrażeń regularnych, negacji, filtrowania wyników z wykorzystaniem metadanych czy łączenia warstw analitycznych. Rysunek 4.12 przedstawia wyniki wyszukiwania w korpusie wzmianek zawierających elipsę.

4.8. Statystyki korpusowe

Analiza własności statystycznych korpusu może pomóc w zrozumieniu mechanizmów lingwistycznych działających w zjawisku referencji, ale także w ogólniejszym badaniu spójności tekstu. Przedstawiona dalej charakterystyka zawiera analizę najważniejszych własności tekstu i jednostek referencyjnych, podane rodzaje statystyk należy zatem traktować jako przykładowe.

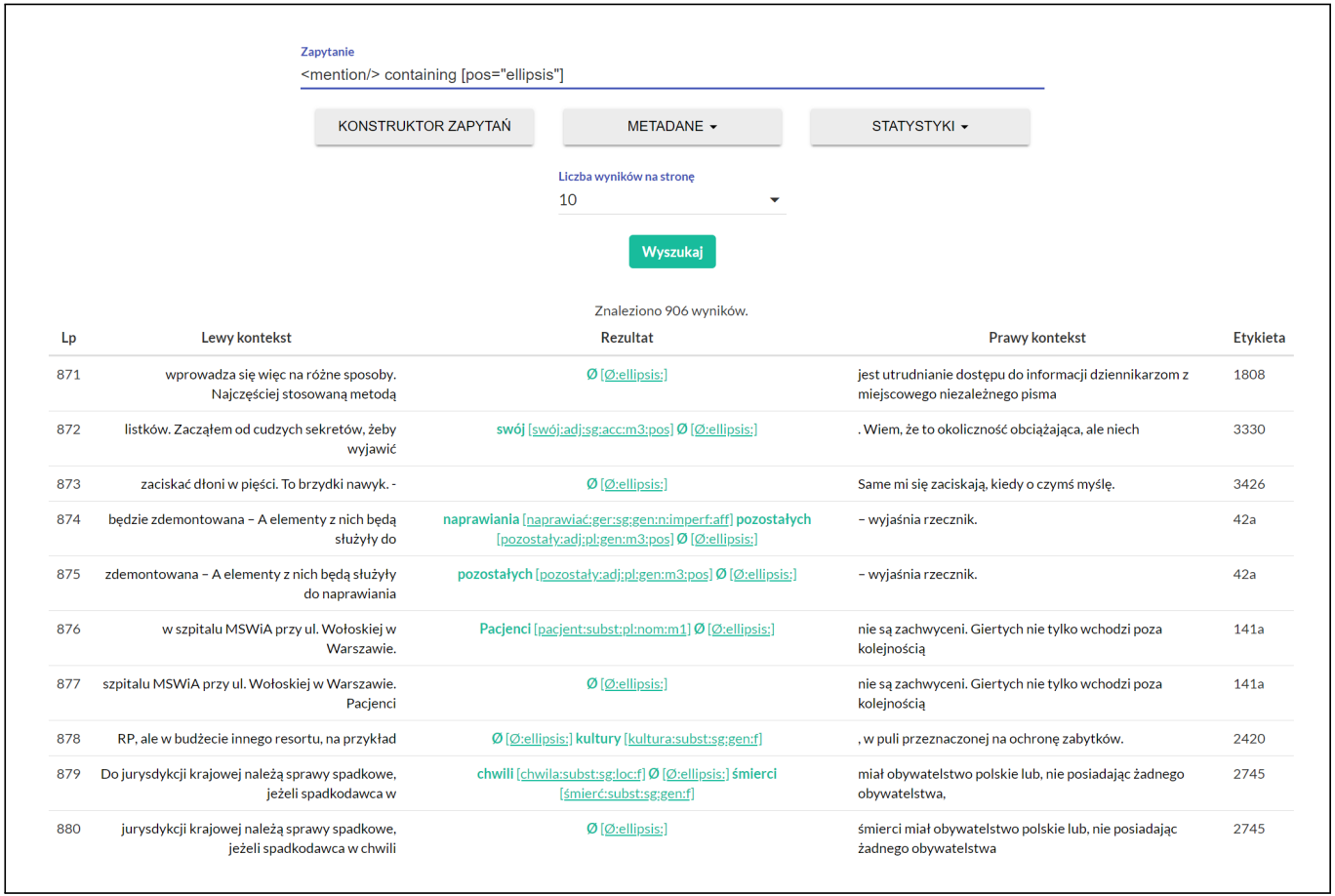
Zasadniczą część korpusu stanowi zróżnicowany gatunkowo podkorpus tekstów krótkich, podaję jednak także wartości dla podkorpusu tekstów długich z podziałem na kategorie tematyczne, zdając sobie sprawę, że wszystkie teksty z tego podkorpusu reprezentują styl prasowy, sam zbiór zaś jest naturalnie zbyt mały, by dostarczyć wiarygodnych danych. Mimo wszystko przedstawienie różnic dziedzinowych oraz próba porównania wpływu długości tekstu na charakter własności relacji anaforycznych wydają się interesujące.

4.8.1. Własności tekstów

Łączna liczba segmentów w korpusie wynosi 532 166, z czego na podkorpus tekstów krótkich przypada 495 866 segmentów, na podkorpus tekstów długich zaś 36 300 segmentów. W tabeli 4.7 przedstawiono statystykę średniej liczby akapitów, zdań i segmentów w tekstach korpusu z podziałem na kategorię tekstu. Średnia

[29] MTAS został użyty do udostępnienia m.in. milionowego podkorpusu NKJP (`http://nkjp.nlp.ipipan.waw.pl/`), korpusu tekstów polskich z XVII i XVIII w. (`http://korba.edu.pl/`), korpusu tekstów polskich z lat 1830–1918 (`http://korpus19.nlp.ipipan.waw.pl/query_corpus/`) czy Korpusu Dyskursu Parlamentarnego (`http://sejm.nlp.ipipan.waw.pl/`).

[30] Dostępny na stronie `http://pcc.nlp.ipipan.waw.pl/`.



Rysunek 4.12. Udostępnienie korpusu w środowisku wyszukiwarki MTAS

długość tekstu wyniosła 284 segmenty dla tekstów krótkich i 1729 segmentów dla tekstów długich. Dodatkowo tabela zawiera dane o średniej zrozumiałości tekstu (1 – tekst najprostszy, 7 – najmniej zrozumiały) uzyskane za pomocą Jasnopisu (Gruszczyński i Ogrodniczuk 2015)[31].

W naturalny sposób teksty konwersacyjne podzielone są na krótkie akapity i krótkie zdania; podobnie dzieje się w tekstach literatury pięknej ze względu na obecność dialogów. W ramach akapitu najkrótsze zdania występują w zbiorze „innych tekstów pisanych", który w NKJP obejmuje teksty urzędowo-kancelaryjne, perswazyjne (ogłoszenia, reklamy, teksty propagandowe), krótkie teksty instruktażowe i listy. Z najdłuższych zdań składają się teksty prawne i ekonomiczne. Podobnie dane potwierdzają intuicyjne przypuszczenie, że typ publicystyczny oraz teksty internetowe nieinteraktywne (czyli np. Wikipedia czy strony statyczne) w podkorpusie tekstów krótkich są pod względem podziału na akapity najbardziej zbliżone do tekstów prasowych w podkorpusie tekstów długich.

4.8.2. Własności wzmianek

Łączna liczba wzmianek w korpusie wynosi 185 802 (w podkorpusie tekstów krótkich – 172 731, tekstów długich – 13 071). Daje to średnio 3,5 wzmianki na 10 segmentów tekstu, co wydaje się wartością wysoką, ale w pełni uzasadnioną ze względu na sposób reprezentacji referencji (każda fraza zagnieżdżona o osobnym centrum semantycznym stanowi niezależną wzmiankę – patrz rozdział 3.2). Liczba faktycznych tekstowych odniesień do obiektów przywoływanych w dyskursie jest zatem dużo wyższa niż czytelnik uświadamia sobie podczas tworzenia wypowiedzi. Podobnie konsekwencją przyjętego sposobu anotacji (w szczególności włączania fraz względnych do granic wzmianki) jest zróżnicowana długość wzmianek[32], jednak aż 87% wzmianek w tekście ma długość nie większą niż 5 segmentów.

W tabeli 4.8 przedstawiono podstawowe różnice między kategoriami tekstów osobno w podkorpusach tekstów krótkich i długich. Średnia „gęstość" wzmianek w podziale na gatunki tekstów jest zbliżona, co wydaje się wynikać z naturalnej konstrukcji wypowiedzi, złożonej z opisu czynności, jej aktorów i podmiotów. Różnice międzygatunkowe są zgodne z wnioskami wcześniejszych badań (np. Gruszczyński i Ogrodniczuk 2015: rozdział 3, s. 67) wskazujących na większą „dynamikę" tekstów mówionych (w szczególności konwersacyjnych – mowy parlamentarne

[31] Dostępny na stronie `http://www.jasnopis.pl`.

[32] Najdłuższa wzmianka w korpusie liczy aż 147 segmentów.

Tabela 4.7. Charakterystyka tekstów w podziale na kategorie – wartości średnie

Kategoria tekstu	Liczba akapitów w próbce	Liczba zdań w akapicie	Liczba segmentów w zdaniu	Zrozumiałość tekstu
Teksty krótkie	7,32	2,61	14,89	3,56
Mówione konwersacyjne	21,93	1,52	9,16	1,18
Literatura piękna	12,32	2,12	10,68	2,32
Mówione medialne	7,98	2,56	14,25	2,73
Inne teksty pisane	7,96	1,77	19,49	5,73
Typ informacyjno-poradnikowy	7,81	2,24	15,85	4,27
Quasi-mówione	6,44	2,48	18,59	4,65
Dzienniki	6,19	2,95	15,23	3,70
Internetowe interaktywne	5,17	2,98	18,29	2,44
Książka niebeletrystyczna	5,06	3,07	18,52	3,83
Książki publicystyczne	4,68	3,39	18,20	3,84
Internetowe nieinteraktywne	4,44	3,29	19,29	5,15
Literatura faktu	4,40	3,87	16,93	3,61
Typ naukowo-dydaktyczny	3,91	3,84	19,50	4,74
Pozostałe periodyki	3,79	4,15	18,41	4,08
Teksty długie	30,33	3,33	17,13	4,29
Wiadomości z kraju	43,67	2,70	14,62	4,00
Publicystyka i opinie	36,00	4,37	15,00	3,33
Prawo	31,67	2,96	21,07	5,33
Ekonomia	27,67	3,31	21,39	5,00
Nauka i technika	25,67	2,79	17,64	4,66
Sport	24,33	4,22	14,05	3,33
Kultura	23,33	3,06	19,28	4,33

są zwykle przygotowywane do wygłoszenia w formie pisanej), mierzoną stosunkiem konstrukcji czasownikowych do rzeczownikowych. Te ostatnie odpowiadają wzmiankom; większe nasycenie tekstu wyrażeniami referencyjnymi oznacza zatem tekst bardziej formalny.

Własności gramatyczne wzmianek

W tabeli 4.9 przedstawiono charakterystykę wzmianek w całym korpusie pod względem ich własności gramatycznych. Wartości w tabeli nie sumują się do 100% przede wszystkim ze względu na charakter składniowy nazw własnych wyrażonych zwykle grupą nominalną oraz brak możliwości wiarygodnej kategoryzacji części wzmianek nierozpoznanych przez analizator morfologiczny.

Większość wzmianek zaimkowych występuje w klastrach singletonowych; wyjątkiem są wzmianki wyrażone zaimkami osobowymi oraz elipsami: tylko 4,97% zaimków osobowych w tekstch krótkich i 3,86% w tekstach długich oraz 8,04% elips w tekstach krótkich i 8,18% w tekstach długich nie tworzy klastrów.

Warto zwrócić uwagę na stosunkowo dużą liczbę wzmianek wyrażonych podmiotem domyślnym, co jest ważną wskazówką dla procesu implementacji systemów do wykrywania wzmianek i koreferencji. Liczba podmiotów domyślnych, biorących udział w procesach koreferencyjnych, jest średnio dwa razy wyższa niż konstrukcji zaimkowych, uważanych za podstawowy nośnik zjawiska anafory.

Budowa wzmianek

W tabeli 4.10 przedstawiono charakterystykę wzmianek w całym korpusie z uwzględnieniem ich struktury. Wzmianki nieciągłe, niezwykle trudne do przetwarzania automatycznego, stanowią zaniedbywalny procent (0,93% w tekstach krótkich i 0,61% w tekstach długich). Ciekawy przypadek stanowią wzmianki, z których jedna jest zagnieżdżona w drugiej, przy czym jednocześnie należą one do tego samego klastra (jak w przypadku całej frazy *Ewę Nowowiejską – siostrę Adama Nowowiejskiego, syna człowieka, który go wychował i skatował za amory do córki*, koreferencyjnej z jej podkreślonym fragmentem). Istnienie tego typu konstrukcji jest jednak wyłącznie konsekwencją wybranej reprezentacji wzmianek: włączania apozycji i podrzędników do treści wzmianki.

Tabela 4.8. Charakterystyka własności referencyjnych: wzmianki

Kategoria tekstu	Liczba wzmia--nek na 10 seg.	Średnia długość wzmianki w seg.	Procent wzmianek o długości 1	⩽2	⩽3	⩽4	⩽5
Teksty krótkie	3,48	2,91	49,89%	70,59%	79,62%	84,73%	88,23%
Inne teksty pisane	3,97	4,61	28,98%	51,20%	63,78%	71,07%	76,52%
Internetowe nieinteraktywne	3,93	3,73	35,85%	58,75%	70,60%	78,21%	82,90%
Typ informacyjno-poradnikowy	3,73	3,08	43,92%	65,69%	76,09%	82,37%	87,06%
Quasi-mówione	3,68	3,92	39,52%	62,45%	72,25%	77,82%	82,29%
Książka niebeletrystyczna	3,60	3,30	49,36%	68,86%	77,94%	83,15%	86,26%
Literatura faktu	3,51	2,99	49,20%	69,94%	79,12%	84,54%	87,86%
Dzienniki	3,51	2,91	45,16%	68,69%	78,49%	84,05%	88,02%
Typ naukowo-dydaktyczny	3,50	3,54	39,49%	62,05%	72,59%	79,25%	84,07%
Pozostałe periodyki	3,45	3,27	43,27%	65,59%	75,94%	81,93%	85,91%
Internetowe interaktywne	3,39	2,23	64,73%	80,35%	87,58%	91,22%	93,70%
Książki publicystyczne	3,34	3,30	47,44%	68,06%	77,42%	82,70%	85,80%
Literatura piękna	3,33	2,07	67,85%	83,92%	89,60%	92,30%	94,06%
Mówione konwersacyjne	3,21	1,37	82,23%	93,08%	96,35%	97,75%	98,59%
Mówione medialne	3,20	2,65	55,28%	75,31%	83,19%	87,19%	89,92%
Teksty długie	3,60	3,24	41,40%	64,68%	75,37%	81,74%	86,32%
Prawo	3,79	3,94	31,99%	55,02%	68,23%	76,08%	81,43%
Wiadomości z kraju	3,69	2,77	45,65%	69,60%	79,61%	85,43%	89,57%
Nauka i technika	3,65	3,50	37,28%	61,12%	71,24%	78,47%	82,95%
Publicystyka i opinie	3,58	2,82	52,86%	73,41%	82,25%	87,06%	90,33%
Kultura	3,57	3,48	38,38%	63,65%	73,91%	79,01%	84,44%
Sport	3,47	2,65	48,37%	72,45%	80,19%	86,46%	90,26%
Ekonomia	3,45	3,54	33,32%	57,24%	70,97%	78,55%	84,36%

Tabela 4.9. Gramatyczne typy wzmianek

Typ wzmianki	Teksty krótkie		Teksty długie	
	Liczba	Procent	Liczba	Procent
Grupa nominalna	127 092	73,58%	10 553	80,74%
Nazwa własna	19 087	11,05%	1 760	13,46%
Podmiot domyślny	17 939	10,39%	868	6,64%
Zaimek				
osobowy	8 719	5,05%	466	3,57%
dzierżawczy	5 207	3,01%	424	3,24%
wskazujący	3 645	2,11%	200	1,53%
upowszechniający	693	0,40%	50	0,38%
nieokreślony	542	0,31%	17	0,13%
przeczący	408	0,24%	22	0,17%
Elipsa	480	0,28%	8	0,06%

Tabela 4.10. Wewnętrzna budowa wzmianek

Typ wzmianki	Teksty krótkie		Teksty długie	
	Liczba	Procent	Liczba	Procent
Nieciągła	1 612	0,93%	80	0,61%
Zagnieżdżona	44 426	25,72%	3814	29,18%
poziomu 2	18 982	10,99%	1707	13,06%
poziomu 3	7 268	4,21%	697	5,33%
poziomu > 3	4 262	2,47%	416	3,18%
Z zagnieżdżeniem w tym samym klastrze	297	0,17%	13	0,10%
Z częścią wspólną	28 772	16,66%	2544	19,46%

4.8.3. Statystyka relacji referencyjnych

Klastry koreferencyjne

Łączna liczba klastrów w korpusie wynosi 21 865 (w podkorpusie tekstów krótkich – 20 408, tekstów długich – 1 457). W tabeli 4.11 zaprezentowano podstawowe własności klastrów koreferencyjnych z podziałem na typy tekstów.

Udział klastrów określonej wielkości jest zbliżony w obu podkorpusach. Najobszerniejszy klaster w podkorpusie tekstów krótkich zawiera 41 wzmianek (w dialogu mówionym; wzmianki są głównie zaimkami w pierwszej osobie i podmiotami

Tabela 4.11. Charakterystyka podstawowych własności klastrów koreferencyjnych

Kategoria tekstu	Liczba klastrów na 100 seg.	Wielkość klastra	Procent klastrów rozmiaru 1	⩽2	⩽3	⩽4	⩽5
Teksty krótkie	4,12	1,47	82,61%	92,41%	95,34%	96,76%	97,60%
Quasi-mówione	5,28	1,54	77,96%	90,01%	93,96%	96,05%	97,26%
Inne teksty pisane	5,12	1,43	81,64%	92,42%	95,31%	97,25%	98,13%
Internetowe interaktywne	4,76	1,66	76,69%	89,17%	93,31%	95,32%	96,42%
Typ informacyjno-poradnikowy	4,60	1,44	82,20%	92,46%	95,55%	97,04%	97,95%
Mówione medialne	4,52	1,64	76,77%	88,48%	92,38%	94,74%	96,51%
Dzienniki	4,34	1,40	82,70%	92,85%	95,91%	97,37%	98,15%
Książki publicystyczne	4,29	1,41	81,94%	93,04%	95,72%	97,17%	98,47%
Typ naukowo-dydaktyczny	4,05	1,34	84,49%	93,76%	96,34%	97,61%	98,43%
Internetowe nieinteraktywne	4,03	1,28	86,80%	94,89%	97,25%	98,30%	98,90%
Pozostałe periodyki	3,96	1,34	84,61%	93,92%	96,57%	97,67%	98,42%
Literatura faktu	3,95	1,42	83,99%	92,92%	95,79%	97,14%	97,90%
Książka niebeletrystyczna	3,85	1,45	84,47%	93,13%	95,24%	96,88%	97,74%
Mówione konwersacyjne	3,63	2,45	72,32%	83,84%	88,27%	90,16%	91,55%
Literatura piękna	3,48	1,73	81,96%	90,88%	93,40%	94,80%	95,61%
Teksty długie	4,01	1,54	82,78%	91,97%	95,04%	96,50%	97,57%
Nauka i technika	4,46	1,50	81,67%	91,97%	95,44%	96,42%	97,83%
Wiadomości z kraju	4,43	1,86	77,64%	87,99%	92,29%	93,95%	95,51%
Kultura	4,39	1,51	81,45%	91,80%	95,18%	97,13%	97,75%
Ekonomia	4,13	1,43	82,92%	92,06%	95,22%	96,77%	98,17%
Publicystyka i opinie	3,81	1,63	82,60%	91,88%	95,17%	96,59%	97,42%
Prawo	3,68	1,37	86,67%	93,28%	95,72%	97,13%	98,04%
Sport	3,40	1,61	84,19%	94,19%	95,81%	97,10%	97,85%

Tabela 4.12. Typy zgodności wzmianek w klastrach

Typ zgodności	Teksty krótkie		Teksty długie	
	Klastrów	Procent	Klastrów	Procent
Zgodność liczby	14 736	72,21%	1 059	72,68%
Zgodność rodzaju	11 065	54,22%	891	61,15%
Zgodność liczby i rodzaju	10 124	49,61%	791	54,29%
Zgodność lematu centrum	6 649	32,58%	587	40,29%
Identyczność tekstowa	3 583	17,56%	308	21,14%
Razem	20 408	100,00%	1 457	100,00%

domyślnymi), w podkorpusie tekstów długich zaś – 129 wzmianek (w biografii św. Wojciecha; wzmianki to odniesienia do bohatera oraz podmioty domyślne).

Zgodność wzmianek w klastrach

W tabeli 4.12 przedstawiono statystykę podstawowych własności uzgodnienia centrów semantycznych wzmianek w klastrach niesingletonowych. Podane wartości oznaczają liczbę klastrów, w których wszystkie centra wzmianek są zgodne na mocy podanej własności. Zgodność rodzaju w przypadku rodzaju męskiego zakłada utożsamienie rodzajów męskich `m1`–`m3`.

Koreferencja a spójność tekstu

W tabeli 4.13 zaprezentowano własności klastrów wpływające na spójność tekstu. Dla klastrów niesingletonowych podaję informację, ile z nich zawiera wzmianki należące do różnych zdań (pozostałe zawierają się w całości w jednym zdaniu) lub akapitów.

Wysoki współczynnik udziału singletonów w zbiorze klastrów koreferencyjnych w powiązaniu z bardzo wysokim współczynnikiem udziału klastrów międzyzdaniowych i stosunkowo wysokim międzyakapitowych potwierdza decydującą rolę koreferencji w zapewnieniu spójności tekstu oraz świadczy o silnym charakterze powiązań spójnościowych – stosunkowo niewielka liczba klastrów wieloelementowych pokrywa przeważającą część tekstu. Zwraca także uwagę wysoka spójność tekstów prawnych mierzona pokryciem akapitów linkami referencyjnymi. Nie należy natomiast wyciagać zbyt daleko idących wniosków z wysokiego udziału

Tabela 4.13. Charakterystyka klastrów w poszczególnych stylach

Kategoria tekstu	Procent klastrów międzyzdaniowych	Procent klastrów międzyakapitowych	Procent zdań nie pokrytych	Procent akapitów nie pokrytych
Teksty krótkie	80,94%	47,52%	13,71%	9,97%
Quasi-mówione	86,81%	55,70%	6,54%	3,25%
Inne teksty pisane	76,39%	60,90%	31,92%	15,30%
Internetowe interaktywne	72,46%	43,74%	12,87%	13,80%
Typ informacyjno-poradnikowy	78,42%	52,84%	15,20%	12,16%
Mówione medialne	75,30%	49,05%	13,57%	3,42%
Dzienniki	86,92%	51,02%	10,62%	10,92%
Książki publicystyczne	73,31%	28,39%	11,92%	7,87%
Typ naukowo-dydaktyczny	76,63%	36,87%	14,07%	10,95%
Internetowe nieinteraktywne	79,72%	39,72%	11,04%	16,00%
Pozostałe periodyki	77,53%	36,23%	11,17%	9,71%
Literatura faktu	77,06%	31,08%	12,13%	10,05%
Książka niebeletrystyczna	76,38%	40,20%	12,54%	12,09%
Mówione konwersacyjne	86,30%	73,59%	12,41%	3,41%
Literatura piękna	81,52%	55,76%	15,91%	10,70%
Teksty długie	87,85%	58,75%	10,49%	7,19%
Nauka i technika	91,72%	68,05%	12,99%	7,44%
Wiadomości z kraju	93,01%	66,38%	10,69%	9,32%
Kultura	82,32%	59,67%	7,14%	7,48%
Ekonomia	85,19%	46,09%	14,46%	7,27%
Publicystyka i opinie	87,41%	58,52%	7,41%	7,42%
Prawo	88,53%	59,63%	8,42%	3,91%
Sport	86,39%	55,10%	12,33%	7,47%

klastrów międzyakapitowych w danych konwersacyjnych i quasi-mówionych, gdyż wynika on zapewne z obecności sztucznych oznaczeń mówców.

Relacje pośrednie i pomocnicze

W tabeli 4.14 przedstawiono dystrybucję aspektów poszczególnych typów, w tabeli 4.15 zaś – podstawową charakterystykę własności relacji pośrednich i pomocniczych. Dane korpusowe wykazują, że udział linków aspektowych jest niewielki – tylko 3,44% relacji w przypadku tekstów krótkich i 3,7% relacji dla tekstów długich zostało oznaczonych aspektem.

Tabela 4.14. Statystyka linków aspektowych

Aspekt	**Teksty krótkie**		**Teksty długie**	
	Liczba	**Procent**	**Liczba**	**Procent**
Dysymilacja	295	1,88%	41	2,98%
Opinia	180	1,15%	8	0,58%
Niepewność	64	0,41%	2	0,15%
Brak aspektu	15 128	96,56%	1 327	96,30%

Tabela 4.15. Statystyka relacji pośrednich i pomocniczych

Typ relacji	Teksty krótkie			Teksty długie		
	Liczba linków	Odległość we wzmiankach	Odległość w słowach	Liczba linków	Odległość we wzmiankach	Odległość w słowach
Relacje pośrednie						
Agregacja	8 481	12,25	33,02	732	41,64	112,97
Kompozycja	1 908	7,72	20,67	100	40,10	112,64
Anafora związana	289	3,19	7,02	28	3,07	6,71
Relacja inna	868	17,42	47,86	106	65,03	184,36
Relacje wspierające						
Metareferencja	151	4,33	10,91	3	1,33	0,67
Porównanie	96	3,74	8,88	5	3,00	6,20
Predykat	2 100	3,01	5,94	250	2,28	3,57
Relacja inna	870	9,05	17,76	88	11,80	21,42
Relacje wykluczające						
Kategorialność	218	10,71	27,89	6	9,17	25,67
Kontrast	216	3,08	8,01	9	2,44	5,44
Polisemia	2	10,00	20,50	0	—	—
Relacja inna	5	2,80	6,80	0	—	—

5.

Implementacja

W trakcie prac implementacyjnych powstał zestaw narzędzi do dekodowania wzmianek i klastrów koreferencyjnych w tekstach polskich. Narzędzia te stworzono wykorzystując algorytmy popularne w ciągu kilku ostatnich lat. W niniejszym rozdziale przedstawiam je w skrócie, odnosząc się do wcześniejszych publikacji opisujących je szczegółowo i prezentując metody dochodzenia do najlepszego wyniku za pomocą różnych metod komputerowych. Narzędzie do dekodowania relacji pośrednich i pomocniczych traktuję jako eksperymentalne przede wszystkim ze względu na stosunkowo niewielką ilość dostępnych danych treningowych. Mimo to podejmuję próbę jego stworzenia dla trzech relacji o największej frekwencji.

5.1. Wykrywanie wzmianek

W związku z ograniczeniem zakresu badań do relacji referencyjnych z komponentem nominalnym skupiam się głównie na wykrywaniu uogólnionych grup nominalnych zgodnie z ich opisem z rozdziału 3.2. Narzędzi do wykrywania poszczególnych konstrukcji tego rodzaju w tekście polskim jest wiele – od analizatorów morfoskładniowych, takich jak np. Morfeusz 2 (Woliński 2014), oferujących analizy dla rzeczowników i zaimków, przez narzędzia do wykrywania nazw własnych, takie jak np. Nerf (Waszczuk i in. 2013), aż po chunkery i parsery powierzchniowe posługujące się gramatyką języka polskiego, np. Spejd (Przepiórkowski i Buczyński 2007) lub wytrenowane na danych NKJP, np. Iobber (Radziszewski 2012). Komponenty te mogą zostać zestawione w celu identyfikacji wzmianek oraz ich cech w różny sposób; w ramach niniejszej pracy powstały dwa tego rodzaju narzędzia, regułowe oraz wykorzystujące uczenie maszynowe.

5.1.1. System regułowy

W konfiguracji podstawowej wyniki analizy morfologicznej ograniczone do kategorii zaimków osobowych (`ppron12` i `ppron3`[1]), form rzeczownikowych (`subst`), rzeczowników deprecjatywnych (`depr`) oraz gerundiów (`ger`) ujednoznaczniano za pomocą tagera morfoskładniowego Pantera (Acedański 2010). Następnie wykrywano grupy nominalne i zaimkowe z użyciem parsera Spejd wykorzystującego gramatykę języka polskiego opracowaną na potrzeby NKJP (Głowińska 2012). Gramatyka ta uwzględniała m.in. konstrukcje skoordynowane (*Jan albo Maria*, *rządu i parlamentu*), grupy z modyfikatorem nominalnym tworzącym apozycję (*terroryści samobójcy*), grupy z modyfikatorem liczebnikowym (*matka trójki dzieci*, podrzędniki w postaci fraz względnych (*szpieg, który mnie kochał*), grupy liczebnikowe traktowane jako nominalne (*dwa rowery*) czy cytaty. Spejd odpowiadał również w przeważającej mierze za wybór centrów semantycznych fraz. Z poziomu nazw własnych pobrano wyrażenia nazewnicze zawierające co najmniej jeden element rzeczownikowy lub zaimkowy. W ostatnim kroku przetwarzania lista kandydatów na wzmianki została uzupełniona wynikami działania zaimplementowanego w ramach opisywanych prac niezależnego komponentu do wykrywania podmiotów domyślnych (Kopeć 2014) przez wstępne oznaczenie osobowych form czasownika w zdaniach niezawierających grupy rzeczownikowej ani zaimkowej w mianowniku.

W związku z inną definicją grupy nominalnej w NKJP niż przyjęta w opisywanym zadaniu (brak zagnieżdżeń dowolnego poziomu, frazy względne niebędące częścią fraz rzeczownikowych, osobno zdefiniowana grupa liczebnikowa) konieczna stała się rozbudowa gramatyki uwzględniającej wszystkie przyjęte założenia. Dokonano zatem reorganizacji gramatyki, uporządkowania i zmiany kolejności jej reguł[2], co pozwoliło na przykład na reprezentację zagnieżdżonych grup nominalnych. Praca ta wymagała podjęcia określonych decyzji w związku z możliwością wystąpienia niejednoznacznych reprezentacji, jak w przypadku wyrażenia *Matki Teresy*, które równie dobrze może zostać zinterpretowane jako apozycja, jak i zwykła sekwencja dwóch rzeczowników w dopełniaczu. W przypadku ciągów rzeczownikowych zdecydowano się tworzyć zagnieżdżenia wyłącznie w sytuacji, gdy grupa składa się z rzeczowników w dopełniaczu rozpoczynających się małą literą. Po zebraniu kandydatów usunięto wzmianki nadmiarowe, np. odpowiadające tym samym

[1] Oznaczenia kategorii morfoskładniowych podano w formalizmie NKJP, patrz np. `http://nkjp.pl/poliqarp/help/plse2.html`.

[2] Prace te zostały szczegółowo opisane w artykule (Ogrodniczuk i in. 2014).

ciągom zidentyfikowanym przez różne komponenty składowe albo krótsze z fraz o wspólnym centrum semantycznym.

W celu uzupełnienia detektora regułowego[3] wykorzystano także zaawansowane informacje leksykalne pochodzące ze słownika walencyjnego WALENTY (Hajnicz i in. 2015), polegające na sterowaniu doborem granic wzmianek wchodzących w skład wielopozycyjnych schematów składniowych. Konfiguracja ta wykorzystuje obserwację, że w przypadku kandydatów na wzmianki wyrażonych schematem rzeczownikowym wszystkie pozycje składniowe schematu powinny zostać połączone w pojedynczą wzmiankę, gdyż tworzą jej pełną semantycznie definicję, jak we frazie *[konflikt]*$_{\text{NOUN}}$ *[polskiego ambasadora]*$_{\text{NP(GEN)}}$ *[z polskim księdzem]*$_{\text{PREPNP(Z,INST)}}$. W przypadku schematów czasownikowych działa zasada odwrotna: frazy rzeczownikowe lub przyimkowe na różnych pozycjach składniowych schematu odpowiadają różnym rolom semantycznym i nie mogą zostać połączone w pojedynczą wzmiankę: *[gratuluję]*$_{\text{VERB}}$ *[Włochom]*$_{\text{NP(DAT)}}$ *[awansu]*$_{\text{NP(GEN)}}$.

Wyniki zostały dodatkowo poprawione z wykorzystaniem dostępnych list przyimków złożonych pochodzących ze słownika walencyjnego[4], *Uniwersalnego Słownika Języka Polskiego* (Dubisz 2006)[5] oraz gramatyki parsera Spejd. Rzeczowniki wchodzące w skład przyimków złożonych zostały usunięte z listy kandydatów na wzmianki na mocy obserwacji, że użyte w tej funkcji formy niezwykle rzadko wchodzą w związki referencyjne.

5.1.2. System statystyczny

System do wykrywania wzmianek wykorzystujący algorytmy uczenia maszynowego, konkurencyjny w stosunku do regułowego, działa na danych wejściowych preanotowanych za pomocą łańcucha analizy lingwistycznej, złożonego ze standardowych narzędzi (wykorzystywanych również w wariancie regułowym): tagera Pantera, parsera Spejd oraz narzędzia do rozpoznawania nazw własnych Nerf. Podmioty domyślne rozpoznawane są wspomnianym detektorem (Kopeć 2014), modele identyfikacji wzmianek ograniczonych do centrów semantycznych i wyszukiwania granic wzmianek trenowane są zaś niezależnie.

Do rozpoznawania wzmianek w wariancie ograniczonym do centrów używany jest algorytm JRIP wykorzystujący cechy segmentów pochodzące z kolejnych warstw

[3] Prace te zostały szczegółowo opisane w artykule (Ogrodniczuk i Nitoń 2017).

[4] Dostępny na stronie `http://walenty.ipipan.waw.pl/rozwiniecia_typow_fraz/`, patrz typ `comprepnp`.

[5] Wersja elektroniczna: `http://usjp.pwn.pl/`.

analizy lingwistycznej: tagera (klasa gramatyczna danego, poprzedniego i następnego segmentu, liczba gramatyczna), parsera powierzchniowego (własność bycia centrum grupy nominalnej, kategoria słowa składniowego, którego częścią jest segment), narzędzia do wykrywania nazw własnych (własność bycia częścią nazwy własnej, pierwszym segmentem w nazwie własnej, odległość w segmentach od kolejnej nazwy własnej), bezpośrednio z warstwy tekstowej (czy następny segment to przecinek lub kropka, czy segment rozpoczyna się wielką literą) oraz własności liczbowych (długość segmentu lub odpowiadającego mu lematu w znakach czy długość zdania w segmentach).

Rozpoznawanie granic wzmianek polega na analizie (algorytmem J48) cech wykrytego wcześniej centrum wzmianki oraz pozostałych segmentów tworzących zawierające je zdanie (zarówno pojedynczych segmentów, jak i cech pary centrum – kandydat) w wariancie zlematyzowanym i tekstowym. Użyte cechy kandydata to:

– klasa gramatyczna danego, poprzedniego i następnego segmentu;
– kategoria nadrzędnego słowa składniowego segmentów sąsiadujących;
– liczba, rodzaj i osoba gramatyczna segmentu;
– kategoria nadrzędnego słowa składniowego;
– własność bycia centrum wzmianki wykrytej w poprzednim etapie;
– obecność przecinka jako następnego segmentu;
– własność rozpoczęcia segmentu wielką literą.

Cechy centrum ograniczone zostały do obecności kropki jako następnego segmentu oraz własności rozpoczynania się wielką literą.

Cechy pary centrum – kandydat to:

– własność tworzenia przez oba segmenty wspólnego słowa składniowego, wspólnej nazwy własnej i wspólnej grupy nominalnej;
– ich odległość w segmentach, porządek w zdaniu, własność bycia tym samym segmentem;
– odległość do nazwy własnej, której częścią jest kandydat;

– przynależność do wspólnego rzeczownikowego schematu składniowego (p. Ogrodniczuk i Nitoń 2017);

– przynależność segmentu poprzedzającego dany do aktualnie rozpatrywanej wzmianki.

5.2. Wykrywanie koreferencji

Wykrywanie relacji koreferencyjnych jest zadaniem o ugruntowanej pozycji w światowej literaturze; w polszczyźnie prace podejmowano jednak sporadycznie i cząstkowo (patrz rozdział 2.5). Opisywane rozwiązanie można zatem uważać za pierwsze dostępne publicznie i na licencji otwartej[6] narzędzie tego rodzaju dla języka polskiego, dodatkowo realizowane za pomocą konkurencyjnych algorytmów i metod analitycznych. Efektem tych prac jest powstanie kilku w pełni funkcjonalnych narzędzi do wykrywania nawiązań w języku polskim w ramach teorii opracowanej w trakcie prac korpusowych oraz przy zastosowaniu różnorodnych modeli: regułowego, statystycznego, wykorzystującego zestawy klasyfikatorów („sita") i sieci neuronowe. Architektura wynikowa (model hybrydowy), osiągająca obecnie najlepsze wyniki, stanowi połączenie dwóch ostatnich metod.

5.2.1. System regułowy

System regułowy RULER[7], opisany w artykułach (Ogrodniczuk i Kopeć 2011a,b), został pomyślany jako punkt odniesienia (ang. *baseline*) dla dalszych prac nad wykrywaniem koreferencji. Został on oparty na modelu Haghighiego i Kleina (2007), zawierającym pewną liczbę ręcznie stworzonych reguł uwzględniających zarówno cechy par wzmianek (ang. *mention-pair*), jak i klastrów (ang. *entity-based*; patrz definicje z rozdziału 2.5). Reguły oparto głównie na własnościach składniowych wzmianek, takich jak np. zgodność składniowa centrów fraz.

Stopień zgodności par wzmianek obliczany jest od bazowej wartości 0,5 (odpowiadającej równej szansie zgodności i niezgodności) przy zastosowaniu następujących reguł:

[6] System IKAR (Broda i in. 2012b) nie doczekał się publicznej wersji, został natomiast odtworzony lokalnie w celu porównania narzędzi dla polszczyzny na potrzeby zadania ekstrakcji informacji (Kaczmarek i Marcińczuk 2015a).

[7] Dostępny na stronie `http://zil.ipipan.waw.pl/Ruler`.

1. wymuszającej zgodność liczby i rodzaju (co oznacza wartość zmniejszoną do zera dla wzmianek niezgodnych składniowo),
2. zapobiegającej łączeniu wzmianki zagnieżdżonej z nadrzędną,
3. promującej wzmianki nominalne o identycznej zlematyzowanej postaci tekstowej,
4. promującej wzmianki zaimkowe w trzeciej osobie zgodne z nominalnymi (gdyż występują one w tekstach często po wzmiankach nominalnych).

Działanie algorytmu klastrującego polega na obliczeniu zgodności badanej wzmianki z każdym z wcześniej stworzonych klastrów. Stopień zgodności wzmianki i klastra stanowi maksimum ze stopni zgodności danej wzmianki ze wszystkimi wzmiankami w klastrze. W przypadku przekroczenia ustalonej wartości progowej wzmianka dołączana jest do klastra o najlepszej zgodności lub – w przypadku równoważności klastrów – do klastra zawierającego wzmiankę tekstowo najbliższą.

5.2.2. System statystyczny

System BARTEK[8], nazywany „statystycznym" dla odróżnienia od modelu regułowego i usprawniający wersję opisaną w artykułach (Kopeć i Ogrodniczuk 2012, Nitoń 2016), został oparty na architekturze systemu BART[9] (Versley i in. 2008) umożliwiającego konfigurację modelu maszynowego uczenia wykorzystującego porównanie cech par wzmianek.

Przykłady treningowe zostały wybrane za pomocą następującego algorytmu[10]:

```
dla każdej wzmianki m:
  dla każdej wzmianki n poprzedzającej m (od tekstowo najbliższej):
    jeśli m i n są koreferencyjne:
      dla każdej wzmianki o pomiędzy n i m
      (włączając n, wyłączając m):
        jeśli o i m są koreferencyjne:
          stwórz z pary (o, m) przykład pozytywny
        w przeciwnym razie:
          stwórz z pary (o, m~) przykład negatywny
```

[8] Dostępny na stronie `http://zil.ipipan.waw.pl/Bartek`.
[9] Dostępny na stronie `http://bart-coref.org/`.
[10] Por. `http://www.sfs.uni-tuebingen.de/~versley/BART/BART-intro.pdf`.

Inicjalnie w skład cech uczących systemu, inspirowanych zarówno zestawem pochodzącym z systemu oryginalnego, jak i obszerną analizą Uryupiny (2007) oraz polskimi pracami teoretycznymi (patrz rozdział 2.3), weszło 147 cech z pięciu grup:

- cech powierzchniowych, wykorzystujących przekształcenia postaci tekstowej wzmianki (np. usunięcie interpunkcji, znaków diakrytycznych, utworzenie skrótu z pierwszych liter słów wchodzących w skład wzmianki, badanie podciągów itp.);
- cech składniowych, takich jak tradycyjna zgodność liczby/rodzaju/osoby;
- cech semantycznych, jak np. zgodność klasy semantycznej wzmianek, relacje synonimii, hiperonimii – na podstawie Słowosieci (Piasecki i in. 2009) oraz zgodność synonimiczna nazw własnych – na podstawie Wikipedii (m.in. w zakresie informacji na temat przekierowania stron, co w większości przypadków odpowiada identyczności pojęć);
- cech metatekstowych, dotyczących zależności istotności wzmianki w tekście od jej pozycji (stąd cechy badające numer akapitu i zdania zawierającego wzmiankę, liczone od początku tekstu czy odległość wzmianki w słowach od początku akapitu lub zdania) oraz wzajemnego położenia wzmianek (liczba akapitów, zdań, wzmianek oddzielających badaną parę wzmianek);
- cech anaforycznych, badających występowanie wzmianki o identycznym centrum semantycznym w poprzednim zdaniu czy odległość między wzmiankami o tym samym centrum.

Ostatecznie po wykonaniu eksperymentów ablacyjnych następujące cechy zostały uznane za znaczące i pozostawione w końcowej wersji systemu statystycznego:

- zgodność rodzaju, liczby gramatycznej, form tekstowych oraz lematów wzmianek w parze;
- część mowy centrum semantycznego poprzednika i następnika;
- trzy cechy obliczające odległość w zdaniach pomiędzy wzmiankami w parze;
- cechy sprawdzające, czy wzmianka jest w pierwszej lub drugiej osobie, osobno dla poprzednika i następnika;

– cecha mieszana weryfikująca, czy wzmianki należą do tego samego zdania, anafora jest zaimkowa, a poprzednik jest pierwszą wzmianką w akapicie;

– cecha mieszana sprawdzająca, czy wzmianki znajdują się w sąsiadujących zdaniach, występują bezpośrednio po sobie, anafora jest zaimkowa i zachowana jest zgodność osoby i liczby gramatycznej;

– dwie cechy badające obecność linku lub przekierowania na stronie Wikipedii odpowiadającej wzmiance.

Do trenowania modelu użyto drzewa decyzyjnego J48 z pakietu WEKA (Hall i in. 2009), do klastrowania przykładów testowych zaś następującego algorytmu:

```
dla każdej wzmianki m:
  dla każdej wzmianki n poprzedzającej m (od tekstowo najbliższej):
    jeśli m i n są koreferencyjne:
        połącz klastry m i n
        przerwij działanie
```

5.2.3. System sitowy

Kolejnym zaimplementowanym wariantem systemu koreferencyjnego był system klasyfikatorów o zmniejszającej się precyzji, tzw. **sit** (ang. *sieves*), zainspirowany systemem Lee i in. (2011) oraz jego późniejszymi wersjami (patrz rozdział 2.5), zwycięskim w zadaniu wykrywania koreferencji dla języka angielskiego w konkursie CoNLL-2011. Ideą działania tego systemu jest spostrzeżenie, że nie wszystkie relacje koreferencyjne są równie trudne do rozstrzygnięcia, wobec czego wcześniejsza analiza łatwiejszych z nich pozwoli zmniejszyć liczbę pozostałych do rozważenia wariantów w kolejnych krokach o coraz mniejszej pewności, a w konsekwencji – prawdopodobieństwo błędnej klasyfikacji.

Sita są kaskadą klasyfikatorów regułowych zestawionych w szereg umożliwiający wykorzystanie przez kolejne etapy analizy informacji na temat klastrów koreferencyjnych zbudowanych w etapach wcześniejszych. Implementacja systemu powstała na platformie BART (Versley 2008) zaadaptowanej do języka polskiego (Kopeć i Ogrodniczuk 2012) oraz rozszerzonej przez Baumanna i in. (2014) do

reprezentacji szeregu sit. Szczegóły prezentowanego systemu i poszczególne decyzje implementacyjne opisuje szerzej artykuł (Nitoń i Ogrodniczuk 2017); dalej prezentuję skrótowo sposób działania i charakterystykę sześciu sit składowych.

Pierwsze sito, zainspirowane konfiguracją *PreciseConstructs* opisaną przez Raghunathana i in. (2010), łączy wzmiankę z jej potencjalnym akronimem utworzonym z pierwszych wielkich liter tworzących ją słów.

Dwa kolejne sita mają zbliżoną postać: drugie łączy wzmianki nominalne zawierające dokładnie ten sam tekst, trzecie – zawierające dokładnie te same zlematyzowane postaci wzmianek, sprowadzone do postaci hasłowej metodą „słowo po słowie" za pomocą analizatora Morfeusz i tagera ujednoznaczniającego Pantera. Obecność drugiego sita wydaje się nadmiarowa, pozwala jednak skorygować ewentualne błędy tagera.

Czwarte sito wykorzystuje już informację z poziomu całego klastra, dołączając nową wzmiankę do klastra zawierającego co najmniej jedną wzmiankę, której centrum semantyczne odpowiada centrum semantycznemu nowej wzmianki oraz której słowa znaczące[11] zawierają się w zbiorze słów znaczących klastra, o ile nie przecinają się granice wzmianek[12].

Dwa ostatnie sita łączą podmioty domyślne ze zgodnymi co do liczby i rodzaju zaimkami osobowymi i konstrukcjami nominalnymi. W tym ostatnim przypadku dodatkowo wymagane jest, by wzmianka rzeczownikowa była pierwszą w zdaniu, łączone konstrukcje były zaś zawarte w tym samym lub bezpośrednio następnym zdaniu.

W tabeli 5.1 przedstawiono precyzję poszczególnych sit. Porównanie systemu sitowego z innymi rozwiązaniami (patrz rozdział 6.2) wskazuje, że w przypadku miary CEAF uzyskiwane wyniki są lepsze od systemu neuronowego, a w przypadku miary MUC – także od systemu statystycznego; dzieje się tak wskutek uwzględnienia typów wzmianek w wyborze metody klastrowania oraz dzięki wykorzystaniu dopasowania wzmianki do klastra (ang. *entity–mention*).

W związku z przekonaniem większości badaczy (np. Cristea i Postolache 2005, Durrett i Klein 2013) o korzyściach z zastosowania tzw. wiedzy ogólnej w procesie poprawiania wyników wykrywania koreferencji jednym z etapów badań była także próba integracji obszernego zasobu tego rodzaju – liczącej ponad pół miliona

[11] Wykorzystano listę polskich słów nieznaczących (ang. *stopwords*) z polskiej Wikipedii: `https://pl.wikipedia.org/wiki/Wikipedia:Stopwords`.

[12] Por. regułę *not i-within-i* z pracy (Haghighi i Klein 2009).

Tabela 5.1. Precyzja poszczególnych klasyfikatorów w modelu sitowym

Nr	Sito	Precyzja danego sita [%]
1	Akronimy	85,71%
2	Formy tekstowe	84,39%
3	Lematy	74,39%
4	Centra semantyczne	56,14%
5	Zaimki osobwe i podmioty domyślne	53,22%
6	Konstrukcje nominalne i podmioty domyślne	33,38%

znaczeń i 1,2 mln objaśnień bazy rzeczownikowych wyrażeń omownych PERIPHRASER[13], stworzonej na podstawie dostępnych zasobów (słownika `sjp.pl`[14], Słowosieci, Wikidanych i bazy objaśnień krzyżówkowych `szarada.net`[15]). Działanie wykorzystującego bazę dodatkowego sita peryfrastycznego wykazało jednak jego niską precyzję ze względu na zbyt prostą reprezentację wiedzy (np. zarówno Putin, jak i Miedwiediew mogą być w dokumentach z różnych okresów określeni jako prezydenci Rosji) oraz niewystarczające pokrycie przypadków występujących w tekstach nawet w bazie o wielomilionowym rozmiarze. Z tego względu sito peryfrastyczne nie zostało włączone do rozwiązania końcowego.

5.2.4. System neuronowy

W związku z ogromną popularnością rozwiązań wykorzystujących głębokie sieci neuronowe w zadaniach przetwarzania języka naturalnego dla języków innych niż polski (głównie angielskiego; dla zadania koreferencji to np. Clark i Manning 2016a,b, Lee i in. 2017) oraz odnoszonych przez nie sukcesów, kolejnym naturalnym krokiem była próba stworzenia systemu tego rodzaju dla polszczyzny (Nitoń i in. 2018) i porównanie go z istniejącymi rozwiązaniami innych typów.

System nazywany „neuronowym" działa na bazie wektorowych reprezentacji słów (Mikolov i in. 2011) używanych jako części składowe wektorów cech wykorzystywanych do trenowania głębokich sieci neuronowych. Elementy budowy systemu są identyczne jak w przypadku rozwiązań pozostałych rodzajów: oprócz wyboru cech

[13]Dostępnej na stronie `http://zil.ipipan.waw.pl/Periphraser`.
[14]Dostępnego na stronie `https://sjp.pl/`.
[15]Dostępnej na stronie `https://szarada.net/`.

implementacja wymaga ustalenia strategii klastrowania oraz progu powyżej którego prawdopodobieństwo zwrócone przez sieć zostanie uznane za wystarczające do podjęcia decyzji o zajściu zjawiska koreferencji.

W ramach prac przetestowane zostały oba algorytmy klastrowania: pierwszy polegał na ocenie zgodności par wzmianek (ang. *mention-based*), w którym dana wzmianka dodawana była do klastra wzmianki ją tekstowo poprzedzającej, dla której sieć zwróciła najwyższe prawdopodobieństwo koreferencji. Drugi oparty był na ocenie zgodności dodawanej wzmianki z całymi klastrami (ang. *entity-based*), w którym kandydat łączony był z klastrem o najlepszej średniej ocenie. Ocena ta obliczana była jako średnia z prawdopodobieństw koreferencyjności par utworzonych z kandydata i pozostałych wzmianek w badanym klastrze. W obu przypadkach prawdopodobieństwa musiały przekraczać ustalony wcześniej próg.

Wektor cech przykładu treningowego zawierał informacje o badanej parze wzmianek, w szczególności wektorowe reprezentacje centrów semantycznych wzmianek, ich pierwsze słowa oraz dwusłowowe konteksty, a także średnie wektorów odpowiadających pięciu słowom sprzed każdej wzmianki i po niej oraz słów je tworzących. Ponadto uwzględniono serię cech binarnych, takich jak informacja o typie wzmianki (rzeczownikowa, zaimkowa, w postaci podmiotu domyślnego, inna). Cechy odpowiadające parze to m.in. odległość dzielącą wzmianki (w słowach i wzmiankach), identyczność tekstowej reprezentacji wzmianek w postaci tekstowej i zlematyzowanej, obecność w tym samym zdaniu i akapicie oraz własność bycia akronimem.

Wektory treningowe były wejściem do w pełni połączonej głębokiej sieci neuronowej, dającej w wyniku wartość z przedziału `[0,1]` interpretowaną jako prawdopodobieństwo koreferencji badanej pary wzmianek. Sieć miała trzy warstwy ukryte o rozmiarze odpowiednio 500, 300 i 100 i używała funkcji Rectified Linear Unit (ReLu, Nair i Hinton 2010) jako funkcji aktywacji oraz funkcji sigmoidowej w warstwie wyjścia. Celem treningu było wykrycie wag minimalizujących funkcję straty mierzoną binarną funkcją entropii krzyżowej. W procesie treningu strata była minimalizowana metodą ADAM (Kingma i Ba 2015) przez dwie epoki; eksperymenty z większą liczbą epok powodowały przeuczenie sieci. W warstwach ukrytych stosowana była normalizacja (Ioffe i Szegedy 2015) i regularyzacja (ze współczynnikiem dropout o wartości 0,2) z wyłączeniem cech odpowiadających reprezentacjom wektorowym słów, traktowanym jako statyczne (Srivastava i in. 2014). Implementacja sieci wykorzystywała API KERAS (Chollet 2015) z biblioteką TENSORFLOW (Abadi i in. 2016). Systemy były trenowane na 90% tekstów

z korpusu koreferencyjnego; 10% tekstów zostało użyte jako podkorpus testowy. Podziału dokonano z uwzględnieniem zrównoważenia typów tekstów.

Eksperyment bazowy wykorzystywał reprezentacje wektorowe o rozmiarze 50, łączny rozmiar przykładu liczył zaś 1147 cech (554 dla każdej wzmianki i 39 dla pary). Zbiór treningowy miał wielkość 426 tys. par wzmianek, z równym udziałem przykładów pozytywnych i negatywnych. Eksperyment wykazał, że najlepsze wyniki osiąga system wykorzystujący cechy wzmianek (ang. *mention-based*). W kolejnym kroku sprawdzono, jak wpłynie na wynik zwiększenie rozmiaru wektorów reprezentacji słów (z 50 do 300). Nie zaobserwowano jednak żadnej znaczącej poprawy, co oznacza, że w tym zadaniu mniejszy rozmiar wektorów reprezentujących cechy słów okazał się wystarczający.

Kolejny eksperyment polegał na dodaniu większej liczby cech wzmianek i ich par do przykładów treningowych. Wykorzystano cechy zebrane w procesie implementacji wcześniejszych systemów (patrz rozdział 5.2.2), w szczególności w przypadku wzmianek zostały dodane cechy binarne określające, czy wzmianka rozpoczyna się zaimkiem wskazującym, czy zawiera zaimek osobowy itp.; w przypadku par badane były m.in. zgodność rodzajów bez rozróżniania podtypów rodzaju męskiego czy zgodność jednej ze wzmianek z akronimem stworzonym *ad hoc* z pierwszych liter drugiej.

Następna próba polegała na przetestowaniu innej architektury sieci, mianowicie tzw. sieci syjamskiej (Bromley i in. 1994), uznawanej za odpowiednią dla zadania wyszukiwania relacji podobieństwa (np. między twarzami lub podpisami dwóch osób). Rozwiązanie miało postać dwóch podsieci o współdzielonych wagach przetwarzających dane na wejściu oraz z modułu obliczającego wynik. Niestety, użycie sieci tego rodzaju nie przyniosło poprawy wyników. Typową dla zadania koreferencji dominację przykładów negatywnych nad pozytywnymi próbowano także przezwyciężyć, losując dodatkowo 600 tys. niekoreferencyjnych par wzmianek (przy zachowaniu stosunku pięciu przykładów negatywnych do jednego pozytywnego). Do puli losującej zostały także dodane wzmianki singletonowe, wcześniej niebrane pod uwagę.

Ostatni eksperyment neuronowy, którego efektem jest system dający wyniki miary F_1 CoNLL porównywalne z najlepszymi wynikami uzyskanymi za pomocą innych modeli (regułowego, statystycznego, sitowego), polegał na zbadaniu dla danej wzmianki wszystkich wzmianek w tekście, a nie wyłącznie wzmianek ją poprzedzających.

5.2.5. System hybrydowy

Wyniki eksperymentów z głębokimi sieciami neuronowymi skłoniły mnie do przeprowadzenia eksperymentu łączącego najlepszy model sitowy z najlepszym modelem neuronowym. W tym celu dane wejściowe zostały przetworzone modelem sitowym w wariancie pełnym, a następnie siecią neuronową w wariancie uwzględniającym wszystkie wzmianki (przy progu 0,95). Efektem tych prac było powstanie systemu o efektywności przewyższającej dotychczasowe modele względem niektórych miar ewaluacyjnych (szczegółową ocenę wszystkich powstałych systemów prezentuję w rozdziale 6).

Warto w tym miejscu zaznaczyć, że ze względu na prowadzone równlegle prace nad rozwojem systemów statystycznego i sitowego, uzyskane przez nie najnowsze wyniki w przypadku zadania wykrywania centrów semantycznych nie odbiegają od wyników systemu hybrydowego o więcej niż 1 punkt procentowy, a w zadaniu uwzględniającym pełne granice wzmianek nawet je przewyższają. Jednocześnie zdecydowanie wyższa złożoność obliczeniowa systemów opartych na sieci neuronowej (skutkująca czasem ewaluacji całego korpusu testowego rzędu kilkanastu sekund w przypadku modelu statystycznego i sitowego oraz kilku godzin dla modeli neuronowych) sprawia, że w praktycznych zastosowaniach znacznie lepiej sprawdzą się modele niewykorzystujące sieci neuronowych.

5.3. Dekodowanie relacji pośrednich i pomocniczych

Problem wykrywania zależności pośrednich jest uznawany za trudny i z tego względu rzadko podejmowany (patrz także rozważania w rozdziale 6.4.3); nie bez znaczenia jest także brak dostępności stosunkowo dużej ilości danych uczących, których wymaga implementacja narzędzi wykorzystujących techniki uczenia maszynowego. Taki jest też omawiany przypadek – mimo dość obszernego korpusu udział relacji pośrednich i pomocniczych rzadszych typów jest stosunkowo niewielki (rzędu 200 relacji na pół miliona segmentów, patrz tabela 4.15), podjęto zatem decyzję o ograniczeniu się do trzech typów relacji pośrednich i wspierających o najwyższej częstości – dwóch podtypów asocjacji strukturalnej (agregacji i kompozycji) oraz pomocniczej relacji predykatywnej.

Proces dekodowania relacji został zrealizowany za pomocą metod uczenia maszynowego. Zgodnie z założeniami anotacyjnymi za cel relacji były uznawane całe klastry koreferencyjne, a nie pojedyncze wzmianki; uwzględniano natomiast

kierunkowość relacji. Dane uczące zostały dobrane tak, że dla każdego przypadku pozytywnego losowanych było 300 przypadków negatywnych. Na tej podstawie drzewo decyzyjne było budowane algorytmem J48.

Wśród dystynktywnych cech relacji agregacji można wyróżnić m.in. własności liczby gramatycznej, użycia liczby w treści wzmianki (co może przekładać się na powiązanie elementu i zbioru), obecność konstrukcji odpowiadającej za tworzenie wyliczeń czy zbliżoną treść analizowanych wzmianek różniącą się tylko liczebnikiem (co z kolei odpowiada tworzeniu podzbiorów). Cechy relacji kompozycji uwzględniają m.in. własności bycia częścią ciała ze Słowosieci czy własność bycia osobą, relacje predykatywne zaś – wartości tradycyjnych kategorii morfoskładniowych oraz własność bycia frazą elektywną. Inne cechy używane wspólnie do wykrywania relacji to na przykład odległość między dowolnymi wzmiankami z klastra źródłowego i docelowego mierzona w segmentach i zdaniach, występowanie w treści centrum semantycznego wzmianek z klastrów co najmniej jednego rzeczownika, czasownika lub zaimka osobowego czy liczby wzmianek w klastrach źródłowym i docelowym.

6.

Ewaluacja szczegółowa

Ocena jakości mechanizmów automatycznego wykrywania zależności referencyjnych jest zadaniem o względnie ustabilizowanej metodologii, w naturalny sposób dzielącym się na etap dotyczący wykrywania wzmianek i etap wykrywania relacji (bezpośrednich lub pośrednich). Warto jednak zaznaczyć, że modele referencji stosowane przez różne zespoły badawcze odpowiadają różnym reprezentacjom zjawisk lingwistycznych, które przekładają się na różnice w znakowaniu tekstów. Z tego względu porównanie jakości narzędzi nawet dla tego samego języka, ale działających z wykorzystaniem różnych zasad anotacji nie jest możliwe. W związku z tym, że w ramach opisywanych prac powstało kilka narzędzi stosujących tę samą reprezentację lingwistyczną, koncentruję się na zestawieniu ich wyników.

Implementacja i ewaluacja systemów wymagała podziału korpusu bazowego na część treningową i testową. Zastosowano 10-krotną walidację krzyżową na korpusie tekstów krótkich dzielonym w stosunku 9:1 w sposób zapewniający zrównoważenie typów i rozmiarów tekstów, przy dodatkowym założeniu, że przynajmniej jeden tekst każdego typu powinien zawsze znaleźć się w próbce testowej. Oceny jakości narzędzi koreferencyjnych dokonano za pomocą narzędzia SCOREFERENCE[1] obliczającego popularne miary wykrywania wzmianek oraz miary koreferencyjne zgodnie z przyjętą metodologią ewaluacyjną. Najlepsze wyniki zostały wyróżnione wytłuszczeniem.

6.1. Wykrywanie wzmianek

Oceny jakości mechanizmu wykrywania wzmianek dokonano za pomocą dwóch strategii:

1. z uwzględnieniem dokładnych granic wzmianek (wzmianka została wykryta poprawnie, jeśli klucz zawierał wzmiankę z dokładnie tymi samymi granicami),

[1] Dostępne na stronie `http://zil.ipipan.waw.pl/Scoreference`.

2. przez porównanie wyłącznie centrów semantycznych wzmianek (wzmianka została wykryta poprawnie, jeśli klucz zawierał wzmiankę o tym samym centrum).

W tabeli 6.1 przedstawiono wyniki ewaluacji dwóch wariantów narzędzia do wykrywania wzmianek opisanych w rozdziale 5.1. Zgodnie z przewidywaniami rozwiązanie stworzone metodą uczenia maszynowego (STAT) wykazuje przewagę w stosunku do narzędzia regułowego (REG) względem wszystkich kryteriów oceny.

Tabela 6.1. Wyniki ewaluacji narzędzi do wykrywania wzmianek

Wariant systemu	**Dokładne granice**		**Centra semantyczne**	
	REG	**STAT**	**REG**	**STAT**
P	70,11%	74,34%	90,07%	92,27%
R	68,13%	69,41%	88,21%	90,21%
F_1	69,10%	**71,79%**	89,12%	**91,23%**

6.2. Wykrywanie koreferencji

Zgodnie z panującym w środowisku zwyczajem podawania wyników ewaluacji narzędzi koreferencyjnych w postaci wartości kilku metryk ewaluacyjnych (co wynika z braku powszechnej zgody na stosowanie jednej z nich) obliczam precyzję, kompletność oraz ich średnią harmoniczną F_1 dla miar MUC, B^3, CEAF-E, CEAF-M i BLANC. Za ostateczny wynik jakości systemów uznaję natomiast wartość F_1 miary CoNLL (patrz rozdział 2.6), stanowiącą średnią arytmetyczną wartości F_1 trzech pierwszych miar.

6.2.1. Wzmianki idealne

Ewaluacja systemów do wykrywania koreferencji biorących pod uwagę wyłącznie wzmianki idealne („złote", ang. *gold*) jest dziś coraz częściej traktowana wyłącznie jako uzupełniająca ocenę systemu działającego na czystym tekście (ang. *end-to-end*) ze względu na to, że sytuacja taka praktycznie nie występuje w rzeczywistości. Mimo to podanie wyników klastrowania wzmianek z klucza umożliwia ocenę jakości podsystemu wykrywania koreferencji bez zaburzeń wprowadzanych na

etapie wykrywania wzmianek, kumulującym zwykle błędy składowych narzędzi lingwistycznych. Wyniki porównania systemów zaimplementowanych w wyniku niniejszej pracy i opisanych w rozdziale 5.2 (odpowiednio: regułowego, statystycznego, sitowego, neuronowego i hybrydowego) prezentuję w tabeli 6.2.

Tabela 6.2. Wyniki wykrywania koreferencji dla wzmianek idealnych

		REG	**STAT**	**SIT**	**NEUR**	**HYBR**
MUC	P	53,35%	73,92%	69,47%	69,01%	68,60%
	R	65,85%	60,19%	66,97%	67,58%	70,83%
	F_1	58,94%	66,35%	68,19%	68,26%	**69,69%**
B^3	P	80,00%	91,14%	86,28%	86,75%	84,29%
	R	85,18%	84,14%	86,33%	86,47%	87,98%
	F_1	82,50%	**87,49%**	86,30%	86,60%	86,10%
CEAF-M	P	75,69%	82,74%	81,41%	81,38%	81,01%
	R	75,69%	82,74%	81,41%	81,38%	81,01%
	F_1	75,69%	**82,74%**	81,41%	81,38%	81,01%
CEAF-E	P	85,02%	84,63%	86,98%	86,59%	88,08%
	R	77,09%	90,91%	88,23%	87,27%	86,93%
	F_1	80,86%	**87,66%**	87,60%	86,92%	87,50%
BLANC	P	72,71%	79,88%	68,80%	73,24%	67,45%
	R	69,86%	73,15%	76,05%	76,09%	79,89%
	F_1	71,15%	**76,04%**	71,77%	74,48%	71,84%
CoNLL	F_1	74,10%	80,50%	80,70%	80,59%	**81,09%**

6.2.2. Wzmianki systemowe

W celu umożliwienia ewaluacji narzędzi do wykrywania koreferencji działających na wzmiankach zidentyfikowanych automatycznie (z wykorzystaniem najlepszego dostępnego systemu – patrz rozdział 5.1.2), obliczenia prowadzone były w dwóch wariantach:

1. z uwzględnieniem tylko tych wzmianek systemowych, które znajdują się w kluczu (przecięcia obu zbiorów);

2. z użyciem przekształcenia zgodnie z procedurą opisaną przez Marqueza (2012) i używaną podczas zadania ewaluacyjnego *Modelling Unrestricted*

Coreference in OntoNotes na konferencji CoNLL-2011 (Pradhan i in. 2011) w celu zapewnienia identyczności zbiorów umożliwiających bezpośrednie porównanie:

(a) wzmianki z klucza, które nie mają odpowiedników systemowych są dodawane do wyników systemu jako singletony;

(b) systemowe wzmianki singletonowe bez odpowiedników z klucza są usuwane;

(c) systemowe wzmianki niesingletonowe (wchodzące w skład jakiegoś klastra koreferencyjnego) są dodawane do klucza jako singletony.

W przypadku konfiguracji dla wzmianek systemowych obliczam warianty miar w konfiguracjach uwzględniających dokładne granice wzmianek (patrz tabela 6.3) lub wyłącznie głowy wzmianek (patrz tabela 6.4) w obu wariantach (z przecięciem i przekształceniem zbiorów wzmianek).

6.3. Wykrywanie wybranych zależności pośrednich i pomocniczych

Zgodnie z metodologią oceny zgodności anotatorów relacji pośrednich (Zikánová i in. 2015: rozdział 7) dokonano próby ewaluacji systemu wykrywającego wybrane relacje pośrednie i pomocnicze; w tabeli 6.5 zamieszczono jej wyniki.

Przedstawione dane nie odbiegają od najnowszych osiągnięć dla języka angielskiego uzyskanych przez Hou i in. (2018). Jej system rozpoznający łącznie relacje pośrednie różnych typów (akcji, przynależności do zbioru, atrybutu, całość – część oraz 77,6% relacji „innych") w najlepszej konfiguracji osiąga wartość 21,6% miary F_1.

6.4. Analiza błędów

W ramach analizy błędów porównano działanie powstałych systemów pod kątem ich efektywności w odniesieniu do różnych konstrukcji lingwistycznych. Przeanalizowano zbiór danych testowych o objętości równej 10% wielkości korpusu w kilku aspektach: wykrywania pełnych granic wzmianek, wyłącznie ich centrów

Tabela 6.3. Wyniki wykrywania koreferencji dla wzmianek systemowych: dokładne granice

			REG	STAT	SIT	NEUR	HYBR
PRZECIĘCIE	MUC	P	61,88%	76,64%	71,05%	72,59%	70,15%
		R	72,51%	65,78%	72,79%	71,80%	75,17%
		F_1	66,77%	70,79%	71,91%	72,17%	**72,56%**
	B^3	P	84,04%	91,69%	86,28%	88,05%	84,72%
		R	87,35%	86,09%	88,16%	87,81%	89,25%
		F_1	85,66%	**88,80%**	87,21%	87,92%	86,92%
	CEAF-M	P	79,86%	84,40%	82,66%	83,14%	82,15%
		R	79,86%	84,40%	82,66%	83,14%	82,15%
		F_1	79,86%	**84,40%**	82,66%	83,14%	82,15%
	CEAF-E	P	87,26%	86,17%	88,79%	87,94%	89,43%
		R	81,26%	91,07%	87,92%	88,26%	86,86%
		F_1	84,15%	**88,55%**	88,35%	88,09%	88,13%
	BLANC	P	77,84%	82,32%	71,04%	76,04%	69,47%
		R	74,48%	76,60%	78,97%	77,72%	80,95%
		F_1	76,02%	**79,14%**	74,28%	76,77%	73,70%
	CoNLL	F_1	78,86%	**82,71%**	82,49%	82,73%	82,54%
PRZEKSZTAŁCENIE	MUC	P	43,67%	62,22%	55,47%	56,68%	54,61%
		R	50,53%	45,84%	50,73%	50,04%	52,38%
		F_1	46,84%	52,78%	52,98%	53,12%	**53,46%**
	B^3	P	80,48%	90,27%	85,06%	86,31%	83,73%
		R	83,37%	81,60%	83,01%	82,77%	83,67%
		F_1	81,90%	**85,71%**	84,02%	84,50%	83,70%
	CEAF-M	P	73,04%	78,69%	76,75%	77,12%	76,24%
		R	73,04%	78,69%	76,75%	77,12%	76,24%
		F_1	73,04%	**78,69%**	76,75%	77,12%	76,24%
	CEAF-E	P	80,14%	79,65%	81,39%	80,87%	81,75%
		R	75,78%	87,50%	83,93%	84,31%	82,96%
		F_1	77,89%	**83,39%**	82,64%	82,55%	82,35%
	BLANC	P	69,55%	76,29%	65,68%	69,51%	64,39%
		R	65,94%	67,36%	68,92%	68,07%	70,23%
		F_1	67,51%	**70,83%**	67,11%	68,68%	66,73%
	CoNLL	F_1	68,88%	**73,96%**	73,21%	73,39%	73,17%

Tabela 6.4. Wyniki wykrywania koreferencji dla wzmianek systemowych: centra wzmianek

			REG	STAT	SIT	NEUR	HYBR
PRZECIĘCIE	MUC	P	57,66%	74,10%	68,96%	69,55%	68,16%
		R	71,41%	57,48%	67,16%	66,96%	69,85%
		F_1	63,79%	64,73%	68,05%	68,20%	**68,99%**
	B^3	P	82,29%	92,20%	87,09%	88,02%	85,59%
		R	87,61%	84,33%	87,01%	86,93%	88,11%
		F_1	84,86%	**88,09%**	87,05%	87,46%	86,83%
	CEAF-M	P	78,38%	83,00%	82,09%	82,17%	81,69%
		R	78,38%	83,00%	82,09%	82,17%	81,69%
		F_1	78,38%	**83,00%**	82,09%	82,17%	81,69%
	CEAF-E	P	86,65%	83,72%	87,02%	86,43%	87,73%
		R	78,72%	90,92%	87,89%	87,61%	86,89%
		F_1	82,49%	87,17%	**87,45%**	87,01%	87,31%
	BLANC	P	75,94%	81,84%	71,02%	75,15%	69,36%
		R	74,17%	73,85%	77,24%	76,37%	79,39%
		F_1	74,99%	**77,21%**	73,67%	75,67%	73,22%
	CoNLL	F_1	77,05%	80,00%	80,85%	80,89%	**81,04%**
PRZEKSZTAŁCENIE	MUC	P	51,69%	65,36%	61,52%	63,52%	61,00%
		R	60,23%	48,48%	56,65%	56,48%	58,91%
		F_1	55,62%	55,66%	58,98%	59,76%	**59,93%**
	B^3	P	80,81%	90,31%	85,50%	86,86%	84,12%
		R	84,17%	81,16%	83,35%	83,19%	84,23%
		F_1	82,45%	**85,49%**	84,41%	84,98%	84,17%
	CEAF-M	P	74,16%	78,60%	77,50%	78,11%	77,14%
		R	74,16%	78,60%	77,50%	78,11%	77,14%
		F_1	74,16%	**78,60%**	77,50%	78,11%	77,14%
	CEAF-E	P	81,27%	79,37%	81,98%	81,59%	82,55%
		R	75,91%	87,65%	84,59%	85,22%	83,68%
		F_1	78,50%	83,30%	83,26%	**83,36%**	83,11%
	BLANC	P	72,24%	77,12%	66,85%	71,36%	65,61%
		R	68,23%	67,99%	70,39%	69,84%	71,98%
		F_1	69,98%	**71,55%**	68,40%	70,49%	68,16%
	CoNLL	F_1	72,19%	74,82%	75,55%	**76,03%**	75,74%

Tabela 6.5. Wyniki wykrywania wybranych relacji pośrednich i pomocniczych

Relacja	P	R	F_1
Agregacja	35,90%	25,99%	30,15%
Kompozycja	30,22%	27,19%	28,63%
Predykat	33,63%	19,73%	24,86%

semantycznych oraz łączenia wzmianek w klastry. W przypadku relacji pośrednich i pomocniczych zastosowano inną metodę: przeglądania całego zbioru relacji oznaczonych jako „inne" w celu wykrycia przyczyn ewentualnych trudności w opisie relacji tego rodzaju.

6.4.1. Błędy wykrywania wzmianek

W przeanalizowanym zbiorze, zawierającym wartości z klucza nierozpoznane przez co najmniej jedno z powstałych narzędzi do identyfikacji wzmianek liczącym 20 807 pozycji, znalazło się 12 938 wzmianek (62,18%) niewykrytych przez żadne z narzędzi. Wśród nich można wyróżnić kilka klas konstrukcji lingwistycznych:

– konstrukcje z zagnieżdżoną koordynacją (*przełomie sierpnia i września*);

– wyrażenia z frazą przyimkową (*współpracy z misją polską*);

– apozycje z nazwą własną (*księdzem profesorem Bruno Rychłowskim*);

– frazy nieciągłe (*milcząca dotąd Balcerkowa*);

– wzmianki zawierające tytuły w cudzysłowach (*inscenizację „Aniołów w Ameryce")* czy wtrącenia w nawiasach (*chorej ręki (prawej)*);

– wyrażenia i przedziały liczbowe, procentowe, określenia wartości (*50 proc.*, *410 tysięcy złotych*);

– niestandardowe nazwy własne i akronimy, wykraczające poza informacje dostępne w bazach słownikowych i encyklopedycznych (*PR 2000*, *twórczość malarska J. Łydżby*);

– skróty właściwe dokumentom prawnym (*art. 8 ust. 1 pkt 1a*);

– wzmianki zawierające teksty z literówkami, błędami interpunkcyjnymi, wyrażeniami obcojęzycznymi (*wartość parametru max_connections*, *yż demograficzny z lat 80*).

Większość wymienionych konstrukcji nie jest aktualnie wykrywana przez dostępne narzędzia do przetwarzania języka polskiego i, jak się okazuje, nie poddaje się także łatwo analizie metodami statystycznymi. Niektóre kategorie błędów uda się zapewne w przyszłości poprawić powstającymi narzędziami do korekty tekstu, wykrywania wyrażeń temporalnych czy analizy nazw własnych w schemacie wykraczającym poza przyjęty przez NKJP. Część błędów wynika z obecności konstrukcji nieobsługiwanych przez używaną gramatykę; wykrycie niektórych z nich (np. wyrażeń z frazą przyimkową) jest także uważane za nietrywialne. Niektóre błędy są efektem trudności w łączeniu informacji pochodzącej z różnych warstw analitycznych (np. grup nominalnych i nazw własnych, co uniemożliwia wykrycie apozycji tego typu).

Wyłącznie przez narzędzie regułowe zostało wykrytych 3331 wzmianek (16% analizowanego zbioru), a 4538 (21,81%) — wyłącznie przez narzędzie statystyczne. Ich porównanie wykazuje, że narzędzie regułowe dobrze radzi sobie z identyfikacją fraz nominalno-przymiotnikowych, zagnieżdżonych prostych (dwuelementowych) konstrukcji skoordynowanych (*jęki i świst rózeg*), fraz w całości stanowiących tytuły, standardowych wyrażeń liczbowych, walutowych i dat (*$300*, *01.04.2001*, *16.45*) czy wyrażeń zawierających standardowe skróty (*150 KM*). Narzędzie statystyczne niekiedy nie wykrywa jako wzmianek pojedynczych segmentów, natomiast dobrze radzi sobie z frazami o dużym stopniu zagnieżdżenia (w rodzaju *konieczności upamiętnienia niezafałszowanej części historii ziem zachodnich Polski*), zawierającymi niestandardowe nazwy własne (*pewną Elizabelę*, *układu 8237A*), wyrażenia obcojęzyczne i błędy różnego rodzaju.

Analiza błędów wykrywania centrów semantycznych została przeprowadzona na 5971 wzmiankach nierozpoznanych przez co najmniej jedno z dwóch analizowanych narzędzi (regułowe lub statystyczne). Wyłącznie przez narzędzie regułowe zostało poprawnie wykrytych 788 centrów (13,20%), 12 938 centrów (62,18%) – wyłącznie przez statystyczne, 3886 centrów (65,08%) zaś przez żadne z narzędzi. Większość błędów obu narzędzi okazała się skutkiem niedospecyfikowania instrukcji anotacyjnej, zawierającej jedynie ogólną zasadę wyboru jako centrum pierwszego segmentu wzmianki w sytuacjach niejasnych.

6.4.2. Błędy wykrywania koreferencji

Badanie błędów wykrywania koreferencji przeprowadzono w dwóch wariantach – na zbiorze 784 klastrów dwuelementowych, co pozwala w łatwy sposób przenalizować etap podejmowania decyzji na podstawie cech par wzmianek, a także na zbiorze 769 klastrów o rozmiarze większym niż dwa, co z kolei może posłużyć do wykrycia różnic w sposobie klastrowania na podstawie cech wszystkich wzmianek tworzących klaster.

W zbiorze klastrów dwuelementowych 431 klastrów nie zostało wykrytych przez żadne z badanych narzędzi (wszystkich zaimplementowanych z wyjątkiem regułowego, niebranego pod uwagę ze względu na najmniejszą skuteczność). Analiza konstrukcji lingwistycznych tworzących powiązanie wykazała problemy w dekodowaniu konstrukcji:

– z koordynacją, powiązanych z liczebnikiem zbiorowym (np. *74-letni Jan S. i jego syn 40-letni Dariusz S. – obaj*);

– wymagających zastosowania wiedzy ogólnej wykraczającej poza proste relacje albo trudno kodyfikowalnej (*Noah Stone – dyrektor amerykańskiego stowarzyszenia Artists Against Piracy*, *David Beckham – rozgrywający Realu Madryt*, *emocje o walencji ujemnej – emocje negatywne*);

– wymagających złożonej analizy semantycznej całego tekstu (*pani Elżbieta – żona*, *kolejna zmiana proponowana przez ZMK – wykreślenie sformułowania o „zasadach preferencyjnych”*);

– agregacyjnych, które mogą w niektórych przypadkach zostać uznane za relację pośrednią (*nas – Agencji Leo Burnett, która przygotowała te reklamy*).

Bliższe przyjrzenie się wynikom ograniczonym do przypadków, gdy co najmniej jedno z narzędzi wykryło klaster poprawnie, pokazało, że narzędzie statystyczne lepiej od innych radzi sobie z konstrukcjami zaimkowymi, narzędzie neuronowe zaś miewa problemy z odmianą segmentów wzmianki (np. niewykryty klaster *pana artykułu – pana artykule*, odnaleziony przez wszystkie pozostałe narzędzia). Analiza zbioru klastrów wieloelementowych nie wykazała, żeby między narzędziami występowały różnice inne niż już wspomniane.

6.4.3. Analiza relacji pośrednich

Problem reprezentacji i analizy relacji pośrednich w tekście wydaje się dużo trudniejszy niż w przypadku relacji bezpośrednich; ich występowaniu często nie towarzyszą żadne powierzchniowe wyznaczniki, a dekodowanie relacji często odbywa się na podstawie pozatekstowej wiedzy odbiorcy o wzajemnych zależnościach między referentami. Gdy słyszymy, że ktoś poszedł do kina, a potem, że popcorn był drogi, nieświadomie korzystamy ze znanego nam powiązania i dodatkowo przenosimy je na egzemplarze klas pojęć użytych w konkretnej wypowiedzi. Reprezentacja tego powiązania jest jednak dużo trudniejsza niż w przypadku prostych relacji semantycznych typu wordnetowego. Potwierdzają to inni badacze, np. Versley i in. (2016) wskazują, że ani relacje WordNetu, ani metody semantyki dystrybucyjnej nie wystarczają do zdekodowania częstych relacji tego rodzaju.

Pomijam tu już zupełnie takie kwestie, jak na przykład kontekst komunikacji czy „historia komunikacyjna" budowana między konkretną parą nadawca – odbiorca (np. małżeństwem), która odbiorcy z zewnątrz może zupełnie uniemożliwić zdekodowanie już i tak zawiłego powiązania. Często pojawia się też sytuacja odwrotna, sygnalizowana przez anotatorów: dokładniejsza analiza tekstu ujawnia powiązania dla człowieka oczywiste, a więc podświadomie pomijane. Na przykład w tekście o ewakuacji szkoły anotatorzy wyróżniali – lub pomijali – powiązane wzmianki odnoszące się do nauczycieli, uczniów, lekcji, budynku, boiska. Na ten aspekt problemu zwracali uwagę na przykład Zikánová i in. (2015: s. 237), twierdząc, że możliwości łączenia elementów w tekście wydają się prawie nieograniczone ze względu na subiektywizm jego interpretacji, co stawia pytanie o sposób reprezentacji tego rodzaju zjawisk zapewniający dostatecznie wysoką zgodność anotacji.

Ze względu na tego rodzaju ograniczenia polecono anotatorom oznaczać relacje niereprezentowane w zaproponowanej taksonomii jako „inne", a w komentarzu umieszczać informację o naturze wykrytego powiązania, by ewentualnie skorzystać z niej w przyszłości. Tabela 6.6 zawiera przykłady powtarzających się kategorii z pominięciem przypadków, w których sami anotatorzy wskazywali na trudności w uogólnieniu kontekstu, w jakich wskazana przez nich relacja szczegółowa mogłaby się powtórzyć. Uwaga ta świadczy zresztą o rzadkim charakterze relacji tego rodzaju i ich dużej wariantywności. Z tego względu próba ich dokładniejszej systematyzacji nie wydaje się celowa, a podjęta decyzja o dużej granularności taksonomii wydaje się słuszna.

Tabela 6.6. Wybrane relacje pośrednie i pomocnicze wskazane przez anotatorów

Relacja	**Przykłady**
Kryterium/wartość	*Najbardziej zdziwił mnie jego wzrost. Facet miał jakieś dwa metry.*, *Przywiozła ze sobą 2 euro, czyli prawie 8 zł.*
Obiekt/cecha	*Nowy Jork – nowojorski*, *Hitler – hitlerowiec*
Osoba/ciało	*W lesie znaleziono zwłoki niezidentyfikowanego mężczyzny.*, *Kasia czuła, jak jej ciało staje się chłodne.*
Instytucja/budynek	*I LO zawsze reprezentowało wysoki poziom nauczania. (...)* *W I LO sale były wyjątkowo ciasne.*
Przedział czasowy	*grudzień – zima*, *tej samej porze – podobnej porze*
Proces/wynik	*Po 23 latach sprawa uregulowania własności drogi doprowadzona została do pozytywnego finału.*

7.

Perspektywy badań

W niniejszym rozdziale przedstawiam perspektywy prac nad umieszczeniem relacji referencyjnych w szerszym kontekście analizy międzyjęzykowej oraz analizy struktury tekstu. Prace te, których podstawą były badania opisane we wcześniejszej części książki, dotyczą badań nad rozszerzeniem opisu referencji na środowisko wielojęzyczne oraz dwóch metod opisu koreferencji jako spójnościowej relacji metatekstowej.

7.1. W stronę koreferencji uniwersalnej

Eksperyment motywowany chęcią porównania opisu relacji referencyjnych w różnych językach metodami korpusowymi przeprowadziliśmy wraz z Michalem Novákiem i Anną Nedoluzhko z Uniwersytetu Karola w Pradze, tworząc wielojęzyczny korpus równoległy tekstów angielskich, czeskich, polskich i rosyjskich, anotowanych relacją koreferencji zgodnie ze wspólną metodologią (Nedoluzhko i in. 2018). Prace te nawiązują do innych wielojęzycznych korpusów opisanych relacjami referencyjnymi, takich jak angielsko-niemiecki ParCor 1.0 (Guillou i in. 2014)[1] czy czesko-angielski PCEDT 2.0 Coref (Nedoluzhko i in. 2016), i uzupełniają analizy kontrastywne dokonywane niezależnie (np. Novák i Nedoluzhko 2015). Użycie danych równoległych umożliwia porównywanie reprezentacji referencji w różnych językach oraz usprawnianie narzędzi wykrywających relacje referencyjne z wykorzystaniem danych rzutowanych. Jedną z możliwości jest na przykład poprawianie jakości narzędzia identyfikującego relacje referencyjne w języku o mniejszej liczbie dostępnych zasobów (ang. *less-resourced*) na podstawie projekcji z innych języków; wcześniejszy zbliżony eksperyment projekcyjny dla języka polskiego wykorzystujący tłumaczenie maszynowe opisuje artykuł (Ogrodniczuk 2013).

Podstawą korpusu o nazwie PAWS (od *Parallel Annotated Wall Street Journal corpus*) jest 50 tekstów ekonomicznych z dziennika Wall Street Journal przetłumaczonych na czeski, rosyjski i polski, automatycznie zrównoleglonych na poziomie słów

[1] `http://opus.lingfil.uu.se/ParCor`.

narzędziem GIZA++ (Och i Ney 2000), sparsowanych składniowo, a następnie ręcznie anotowanych relacjami referencyjnymi. Źródłowe teksty angielskie zostały pobrane z korpusu PDTB – Penn Discourse Treebank (Prasad i in. 2008, 2014) i odpowiadają tekstom oznaczonym jako `wsj1900-49`. Konstrukcja PAWS odpowiada budowie korpusu Prague Dependency Treebank (Bejček i in. 2013)[2] stworzonego w metodologii funkcjonalno-generatywnej (Sgall i in. 1986), w ramach której analitycznej warstwie opisu towarzyszy tzw. warstwa tektogramatyczna, zawierająca drzewa zależnościowe. W korpusie reprezentowane są ponadto linki pomiędzy odpowiadającymi sobie słowami w różnych językach. Korpus PAWS liczy obecnie ok. 1000 zdań i 25 tys. słów w każdym języku.

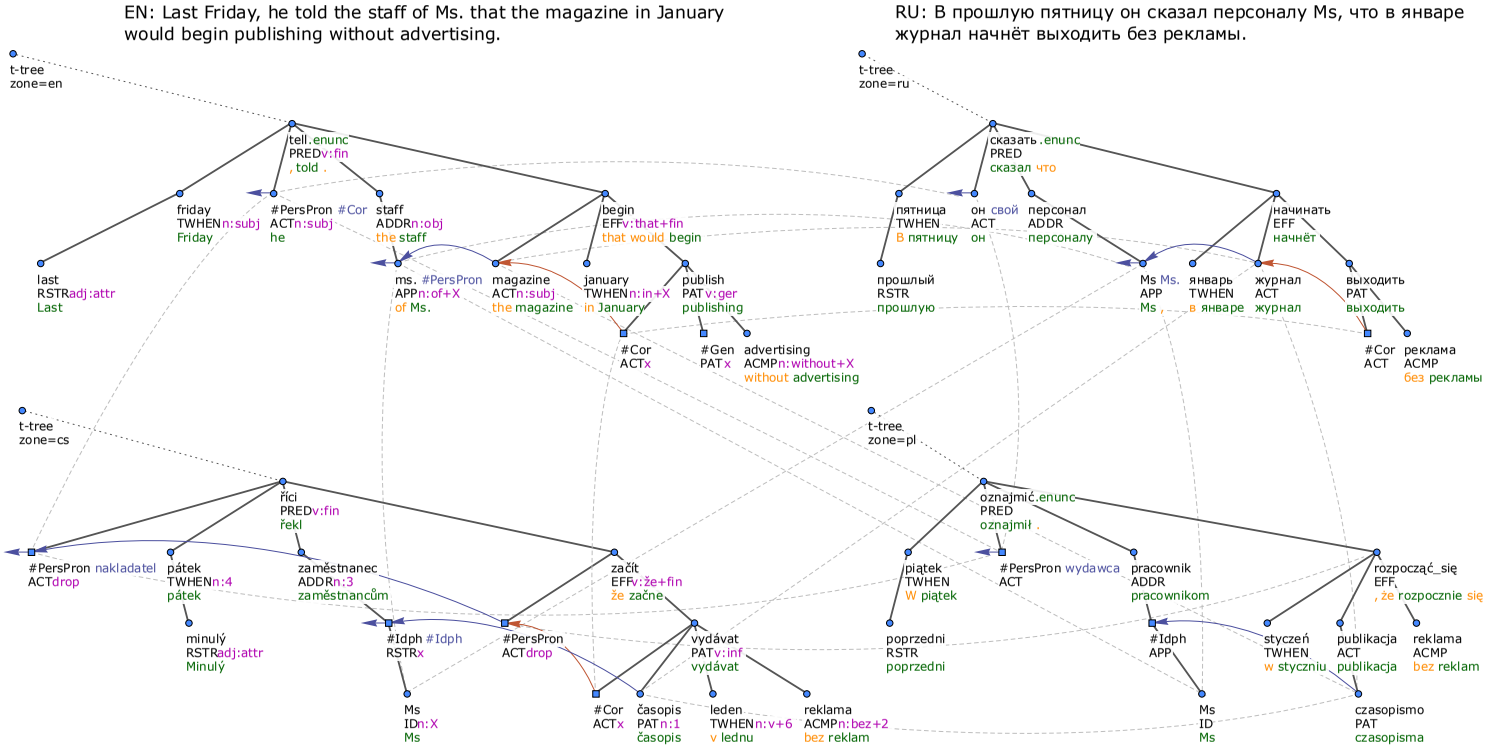
Anotacja relacji referencyjnych została przeprowadzona na warstwie tektogramatycznej i obejmuje przypadki koreferencji gramatycznej (włączając zaimki względne i zwrotne) oraz tekstowej (grupy nominalne włącznie z podmiotami domyślnymi, elipsy, koreferencja przysłówkowa, konstrukcje z rozdzielonym poprzednikiem). Dodatkowo oznaczane są przypadki odwołań do dłuższych fragmentów tekstu oraz egzoforyczne (np. *ten rok*, *niniejszy tekst*). Rysunek 7.1 przedstawia wielojęzyczny graf powiązań dla przykładowego zdania z korpusu; dla uproszczenia linki reprezentujące zrównoleglenie prezentowane są wyłącznie dla wyrażeń koreferencyjnych.

Tabela 7.1 prezentuje statystykę relacji referencyjnych w korpusie już na pierwszy rzut oka wskazującą ciekawe różnice w realizacji relacji referencyjnych w różnych językach. Przykładowo nośnikiem relacji tego rodzaju w języku czeskim są dużo częściej niż w przypadku innych języków zaimki i podmioty domyślne.

Korpus PAWS jest dostępny dla badaczy w repozytorium Lindat/Clarin i zawiera teksty w formacie tekstowym, formacie Treex – wewnętrznej reprezentacji edytora TrEd (Pajas i Štěpánek 2008) i CoNLL (patrz rozdział 4.7.1), z wyłączeniem elementów pozatekstowych, np. elips.

Stworzenie korpusu wielojęzycznego, reprezentowanego w spójnym formalizmie, stanowi próbę zniwelowania różnic w opisie tego samego zjawiska lingwistycznego w różnych językach i wpisuje się w popularny ostatnio trend zapewniania spójnej anotacji wielojęzycznej (np. Universal Dependencies; Nivre i in. 2016), umożliwiającej dokonywanie porównań, eksperymentów uczących i projekcyjnych. Temat koreferencji uniwersalnej (ang. *Universal Coreference*) jest głównym hasłem przyjętego na konferencję NAACL 2019 (Annual Conference of the North American Chapter of the Association for Computational Linguistics) warsztatu CRAC

[2] `https://ufal.mff.cuni.cz/pdt3.0`.



Rysunek 7.1. Wielojęzyczna reprezentacja przykładowego zdania z korpusu PAWS

Tabela 7.1. Statystyka relacji referencyjnych w korpusie PAWS

Liczba jednostek	Angielski	Czeski	Rosyjski	Polski
Zdania	1,078	1,078	1,078	1,078
Słowa	26,149	25,697	25,704	25,823
Węzły tektogramatyczne	18,611	20,696	18,874	18,541
Węzły koreferencyjne	4,210	4,403	4,254	3,371
koreferencja gramatyczna	729	528	749	249
koreferencja zaimkowa	620	856	525	449
koreferencja nominalna	1,361	1,496	1,610	1,568
pierwsze wzmianki	1,277	1,330	1,243	979
podział poprzednika	149	149	91	65
odwołanie do segmentu	28	23	16	12
egzofora	46	21	20	4

2019 (Second Workshop on Computational Models of Reference, Anaphora and Coreference), kontynuującego serię współkierowanych przeze mnie warsztatów CORBON[3].

7.2. Model Penn Discourse Treebank

Opis i automatyczne dekodowanie zależności referencyjnych w języku polskim wyznacza także początek prac nad całościową analizą tekstu i jego struktury, stanowiąc pomost między analizą semantyczną a analizą metatekstową. Z perspektywy komputerowej temat ten nie został jeszcze w polszczyźnie dokładnie zbadany, podczas gdy dla innych języków istnieją już zarówno liczne korpusy, jak i narzędzia do przetwarzania struktury tekstu. Najpopularniejszym korpusem tego rodzaju związanym z określonym formalizmem reprezentacji relacji dyskursywnych jest niewątpliwie PDTB – Penn Discourse Treebank (Prasad i in. 2008, 2014), którego podbudowa teoretyczna została użyta do stworzenia wielu korpusów nieanglojęzycznych, takich jak np. Hindi Discourse Treebank (Oza i in. 2009) czy Prague Discourse Treebank (Poláková i in. 2012, Rysová i in. 2016). Na podstawie hierarchii relacji PDTB powstało także wiele narzędzi do automatycznej analizy dyskursu, prezentowanych m.in. podczas dwóch konkursów w formule Shared Task (Xue i in. 2015, 2016).

[3]Patrz strona `http://corbon.nlp.ipipan.waw.pl/`

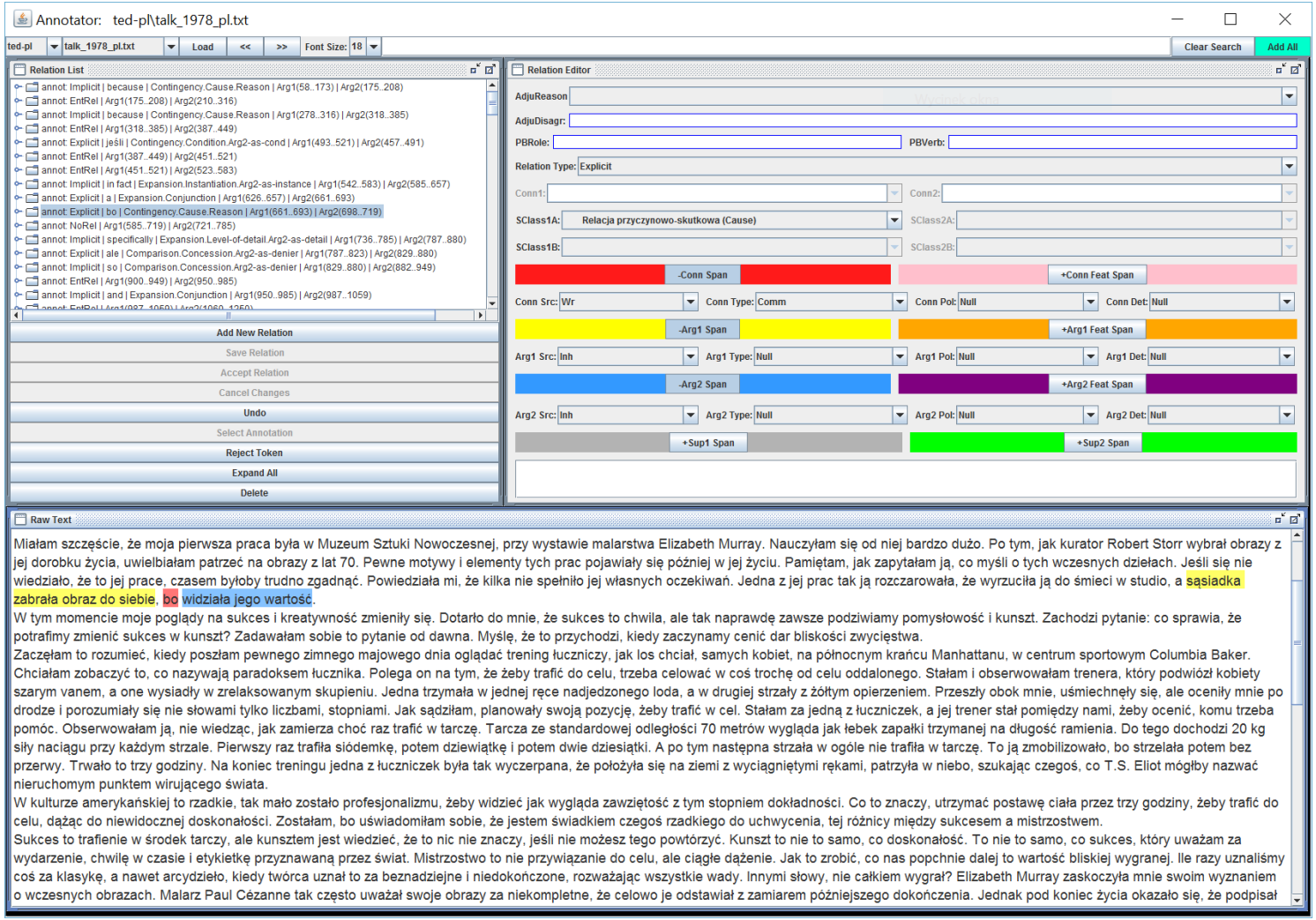
W anotacji PDTB głównym nośnikiem spójności tekstu są relacje pomiędzy **jednostkami dyskursywnymi** (ang. *discourse units*) stanowiącymi tekstowe reprezentacje sytuacji opisujących zdarzenia i stany pewnych **bytów abstrakcyjnych** (ang. *abstract objects*). Relacje wyznaczane są w sposób **jawny** (ang. *explicit*), za pomocą **znaczników dyskursywnych** (ang. *discourse markers*) – najczęściej spójników i zaimków względnych – lub **niejawnie**, odpowiadając pewnemu znacznikowi bez reprezentacji tekstowej (ang. *implicit*). Relacjom przypisuje się typy z ustalonej hierarchii; np. w zdaniu *Sąsiadka zabrała obraz do siebie, bo zdawała sobie sprawę z jego wartości.* znacznik *bo* łączy rozdzielone nim argumenty **relacją sytuacyjną** (ang. *Contingency*) **typu przyczynowo-skutkowego** (ang. *Cause*) z pierwszym argumentem w funkcji **przyczyny** (ang. *Reason*). Oprócz sensów dyskursywnych PDTB wyróżnia w tekście **relacje obiektowe** (ang. *entity-based relations*, EntRel). Są one używane, gdy nie da się określić innej relacji metatekstowej pomiędzy kolejnymi fragmentami tekstu, a nośnikiem spójności tekstu jest referent użyty w obu fragmentach. PDTB zakłada anotację relacji wyłącznie pomiędzy jednostkami bezpośrednio sąsiadującymi, podejście to jest zatem powierzchniowe w tym sensie, że wynikiem analizy tekstu nie jest pełne drzewo rozbioru dyskursu jak w przypadku innych, mniej popularnych formalizmów takich jak RST – Rhetorical Structure Theory (Mann i Thompson 1988) czy SDRT – Segmented Discourse Representation Theory (Asher i Lascarides 2003).

Wraz z badaczami zagranicznymi wykorzystaliśmy schemat PDTB do wstępnej analizy zagadnienia powiązania relacji koreferencyjnych z dyskursywnymi oraz porównania strategii wyrażania relacji tego rodzaju w językach składniowo różnych. W tym celu dokonaliśmy anotacji sześciu tekstów z wielojęzycznego korpusu prezentacji TED[4] (ang. *TED talks*, od *Technology, Entertainment and Design* — Technologia, Rozrywka i Design) relacjami PDTB w oryginalnym środowisku anotacyjnym (patrz rys. 7.2). Prezentacje TED to wystąpienia na żywo w języku angielskim, bazujące na przygotowanym wcześniej skrypcie; dotyczą szerokiego spektrum tematów, od technologicznych po kulturalne i społeczne. Ich tekstowe zapisy wraz z przekładami (wykonywanymi przez wolontariuszy i weryfikowanymi przez zawodowych tłumaczy) są dostępne w formie korpusu równoległego[5] Cettolo i in. (2012) i wykorzystywane w wielu zadaniach z dziedziny przetwarzania języka naturalnego.

Eksperyment ten (Zeyrek i in. 2019) został jednocześnie przeprowadzony dla kilku języków – angielskiego, niemieckiego, polskiego, portugalskiego, rosyjskiego

[4] `https://www.ted.com/`.

[5] Por. też `https://wit3.fbk.eu/`.



Rysunek 7.2. Przykład relacji dyskursywnej w środowisku anotacyjnym PDTB

i tureckiego (wybranych na podstawie kryterium ich różnorodności gatunkowej i rejestrowej, dostępności dla wszystkich badanych języków oraz z wykluczeniem tekstów zawierających linki egzoforyczne, tj. bezpośrednio odwołujących się do prezentowanych podczas prelekcji elementów obrazów i filmów). Dla danego języka opis każdego tekstu został stworzony przez konsylium anotatorów bez konsultacji z wersjami obcojęzycznymi.

Relacje referencyjne w schemacie PDTB różnią się od pozostałych relacji semantycznych w schemacie opisu pod względem funkcji: pierwsze łączą interpretacje obiektów abstrakcyjnych, drugie są środkiem zapewnienia spójności tekstu, przez co wspólną kategorią EntRel określane są zarówno relacje koreferencyjne, jak i asocjacyjne. Opis PDTB ogranicza się do podania faktu wystąpienia relacji; jej argumenty ani podlegające jej wzmianki nie są wskazywane. Poniższy fragment korpusu ilustruje tę zależność: zdanie drugie (wyjaśniające) jest powiązane z pierwszym wyłącznie na zasadzie wystąpienia współodwołania referencyjnego:

EN The reason, I would come to find out, was their prosthetic sockets were painful because they did not fit well. The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle.

DE Der Grund, wie ich später herausfand, waren die Prothesenschäfte, die Schmerzen verursachten, weil sie nicht gut passten. Der Prothesenschaft ist der Teil, in welchen der Amputierte seinen Stumpf steckt, der mit der eigentlichen Prothese verbunden ist.

PL Jak się wkrótce dowiedziałem, wynikało to z faktu, że leje ich protez sprawiały ból, bo nie były dobrze dopasowane. Lej protezowy to część, gdzie człowiek wkłada kikut kończyny i która jest połączona z resztą protezy.

PT A razão, como vim a saber mais tarde, era que o encaixe das próteses era doloroso por não ser um encaixe perfeito. O encaixe de uma prótese é a parte em que o amputado insere o coto do membro, e que liga com a articulação prostética.

RU Я выяснил, что причина была в том, что их культеприемные гильзы вызывали боль, потому что не подходили по размеру. Культеприемные гильзы это часть, куда инвалид вставляет свою культю и которая соединяется с протезом.

TR Sebebi, sonradan öğrendiğim üzere protez soketlerinin düzgün oturmadığı için canlarını yakmasıymış. Protez soketi, uzvu kesilmiş kişinin kesik uzvuna taktığı ve böylece uzvu protez ayağa bağladığı parçadır.

Po dokonaniu automatycznego zrównoleglenia tekstów na poziomie słów systemem GIZA++ (Och i Ney 2000), pociągającym za sobą zrównoleglenie realizacji poszczególnych argumentów, porównane zostały rodzaje relacji reprezentowanych w tekstach wielojęzycznych. Tabela 7.2 przedstawia liczbę relacji PDTB dla języków korpusu.

Tabela 7.2. Statystyka relacji w anotacji testowej wystąpień TED

Język	**Liczba słów**	**Liczba relacji danego typu**					
		Explicit	**Implicit**	**AltLex**	**EntRel**	**NoRel**	**Łącznie**
Angielski	7012	290	198	46	78	49	661
Rosyjski	5623	237	221	20	57	30	565
Polski	6520	218	195	11	104	52	580
Portugalski	7166	269	256	29	38	33	625
Niemiecki	6366	240	214	17	59	30	560
Turecki	5164	276	202	59	70	51	658

Wstępne analizy sposobu realizacji relacji dyskursywnych w różnych językach wykazują różnice w nasyceniu tekstu zależnościami referencyjnymi przy jednoczesnej zbliżonej łącznej liczbie relacji, co może wynikać z chęci zapewnienia przez tłumacza stabilniejszego powiązania spójnościowego fragmentów tekstu niż wyłącznie poprzez użycie linku referencyjnego i wprowadzanie w jego miejsce sztucznych linków metatekstowych.

7.3. Anotacja metatekstowa

W ramach projektu CLARIN-PL[6] rozpoczęto także prace nad anotacją pełnego zasobu Korpusu Koreferencyjnego relacjami metatekstowymi:

– relacjami dyskursywnymi łączącymi segmenty tekstu za pomocą metaoperatora rozszerzającego pojęcie znacznika dyskursywnego w sensie PDTB o partykuły (Grochowski i in. 2014), zaimki i ich kombinacje;

[6] `http://clarin-pl.eu/`.

– relacjami czasowymi łączącymi zdarzenia za pomocą relatorów (wykładników jednoczesności, następstwa itp.);

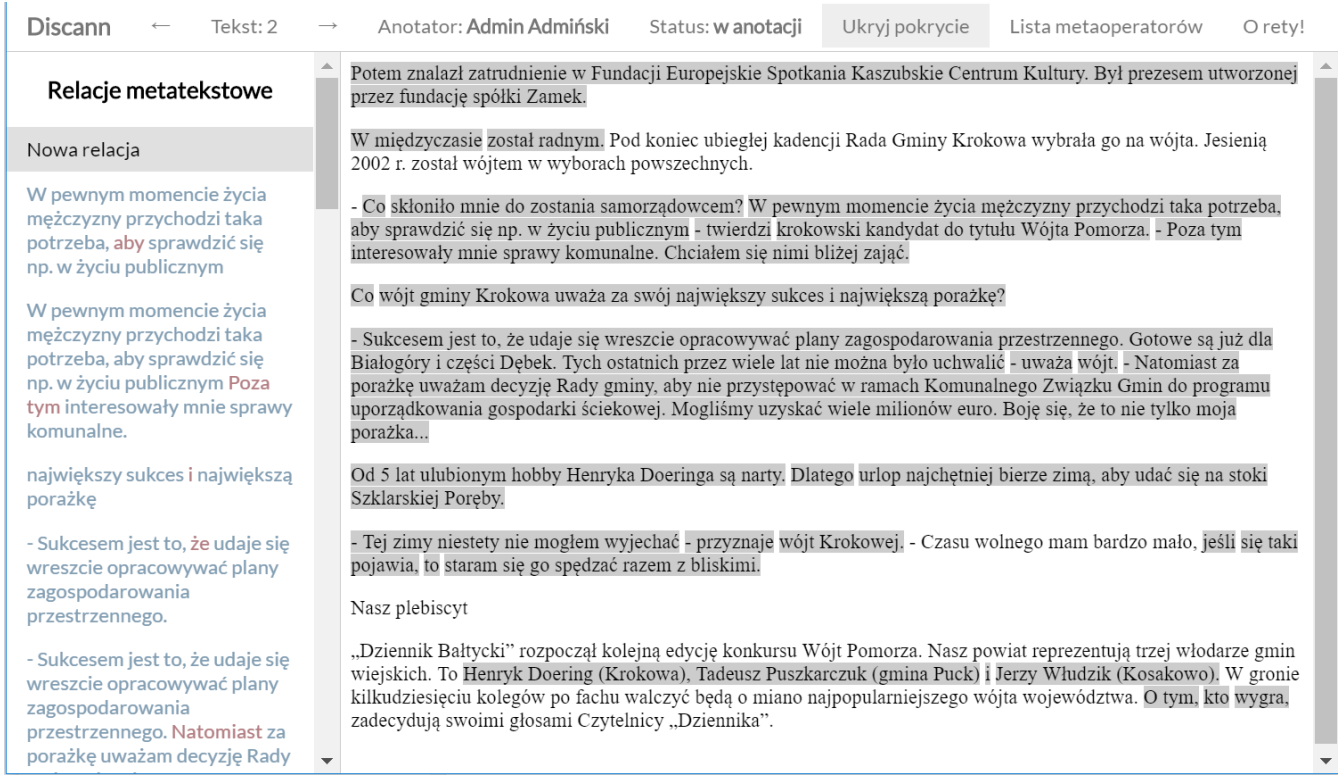
– relacjami między zapisem zdarzeń komunikacyjnych a ich autorskimi kwalifikacjami (nazwami) w tekście;

– relacjami między pytaniami a odpowiedziami, z wyróżnieniem opcjonalnego zaimka pytajnego

oraz analizą eksplicytności i implicytności przekazu treści rozumianej jako obecność między fragmentami tekstu znacznika określającego semantycznie relację między argumentami stanowiącymi reprezentacje zdarzeń w tekście.

Z punktu widzenia niniejszej pracy najistotniejszym elementem tego zadania jest dostarczenie danych do analizy wzajemnych zależności koreferencji i relacji dyskursywnych. Tryb anotacji metatekstowej w zaimplementowanej do tego celu aplikacji DISCANN[7] został zatem uzupełniony o prezentację pokrycia tekstu relacjami metatekstowymi, co umożliwia prześledzenie powiązania nieoznaczonych fragmentów tekstu z jego pozostałą częścią. Rysunek 7.3 prezentuje przykładowy tekst anotowany relacjami dyskursywnymi, w którym relacje koreferencyjne między wyrażeniami odwołującymi się do bohatera tekstu uzupełniają informację metatekstową, realizując wymaganie spójności tekstu.

[7] Dostępna na stronie `http://zil.ipipan.waw.pl/Discann`.



Rysunek 7.3. Relacje koreferencyjne uzupełniające relacje dyskursywne

Podsumowanie

Przedstawione badania stanowią pierwszą próbę wielkoskalowej analizy nominalnych relacji referencyjnych w języku polskim metodami komputerowymi. Budowa obszernego korpusu oraz narzędzi różnych typów pozwoliły osiągnąć w zakresie opisu i wykrywania koreferencji dla języka polskiego stan porównywalny ze światowym. Udało się to dzięki całemu szeregowi prac na wielu poziomach lingwistyczno-technicznych, począwszy od kwestii definicyjnych przez usprawnienia gramatyczne, integrację zasobów wiedzy oraz opracowanie algorytmów wykrywania koreferencji zgodnie z aktualnymi tendencjami implementacyjnymi.

W ramach prac zaproponowano określoną definicję koreferencji, uniezależniając ją od pojęcia anafory i stabilizując rozumienie problemu w polskiej lingwistyce komputerowej. Uporządkowano terminologię z dziedziny referencji, dokonano przeglądu polskiej i zagranicznej literatury oraz dotychczasowych badań w dziedzinie dekodowania relacji referencyjnych i ewaluacji tego procesu.

Opracowano typologię tekstowych nośników referencji, analizując kwestie wpływu frazeologii, nieokreśloności i roli semantyki struktury frazowej w opisie koreferencji. Usystematyzowano opis środków językowych służących przekazywaniu relacji referencyjnych oraz zaproponowano taksonomię relacji referencyjnych z podziałem na relacje pośrednie, pomocnicze i aspekty, wypracowując kompromis pomiędzy szczegółowymi typologiami omawianymi w literaturze a łączną reprezentacją wszystkich relacji asocjacyjnych.

Powstał jeden z największych na świecie reprezentatywny korpus zależności referencyjnych ręcznie opisany relacjami koreferencji i asocjacji, powstały na bazie tekstów Narodowego Korpusu Języka Polskiego zgodnie z uznanym standardem reprezentacji informacji lingwistycznej dla języka polskiego. Prace anotacyjne wymagały właściwego doboru tekstów, opracowania instrukcji anotacyjnej, dostosowania narzędzi i wypracowania najwłaściwszej metodologii anotacji. Podczas prac korpusowych przetestowano kilka strategii anotacyjnych i superanotacyjnych oraz wykazano przydatność superanotacji szeregowej w złożonych zadaniach semantycznych.

Przygotowano środowisko do anotacji i prezentacji danych koreferencyjnych, format reprezentacji danych zgodny z NKJP oraz formaty pośrednie, używane m.in. przez zagraniczne narzędzia ewaluacyjne. Opracowano wyszukiwarkę korpusową dla relacji koreferencyjnych, umożliwiającą powiązanie opisu referencyjnego z innymi warstwami analizy lingwistycznej.

Zaimplementowano dwa systemy do wykrywania wzmianek – regułowy i statystyczny oraz kilka systemów do dekodowania koreferencji – regułowy, statystyczny, sitowy i neuronowy. Przetestowano wiele konfiguracji tych systemów, w tym system hybrydowy, wykorzystując m.in. bazę wyrażeń omownych, dane słownika walencyjnego oraz rozszerzając formalną gramatykę języka polskiego na potrzeby narzędzi składowych. Sprawdzono przydatność kilkuset cech wzmianek do wykrywania koreferencji oraz kilka konfiguracji sieci neuronowej. Przeprowadzono prototypową konfigurację metody do wykrywania relacji asocjacyjnych skutkującą powstaniem systemu do dekodowania relacji agregacyjnych, kompozycyjnych i predykatywnych.

Przeprowadzono ewaluację powstałych systemów zgodnie z powszechnie używaną do tego celu metodologią, potwierdzając efektywność powstałych systemów porównywalną z uzyskiwaną dla innych języków na świecie. Przeprowadzono także analizę błędów narzędzi, co umożliwi ich przyszły rozwój.

We współpracy z partnerami zagranicznymi rozpoczęto badania wiążące prace nad opisem i dekodowaniem relacji referencyjnych dla języka polskiego z pracami prowadzonymi dla innych języków – angielskiego, czeskiego, niemieckiego, portugalskiego, rosyjskiego i tureckiego. Celem tych działań było z jednej strony zainicjowanie prac nad tzw. uniwersalnym modelem koreferencji, z drugiej – wsparcie badań nad spójnością tekstu i powiązanie relacji referencyjnych z relacjami metatekstowymi.

Wyniki przedstawionych prac mają już teraz zastosowanie praktyczne: zostały użyte m.in. w systemie ekstrakcyjnego streszczania tekstu do zastępowania wzmianek odpowiadającymi im podmiotami domyślnymi w celu zapewnienia lepszej płynności tekstu wynikowego. W perspektywie streszczania mieszczą się także działania pokrewne: zastępowanie podmiotów domyślnych wyrażeniami pełnoznacznymi w celu podniesienia zrozumiałości tekstu czy wygładzanie tekstu poprzez zmianę kolejności wzmianek w łańcuchu anaforycznym, w tym na przykład zastępowanie fraz nominalnych ich odpowiednikami zaimkowymi.

Mam nadzieję, że chociaż niektóre zaprezentowane wyniki staną się źródłem refleksji dla czytelników i umożliwią rozpoczęcie nowych badań inżynieryjno-lingwistycznych w polszczyźnie.

English summary

The book presents the summary of corpus-based research on coreference resolution for Polish. Subsequent phases of the work are described in a project-based manner, reflected in the order of chapters: after outlining the initial assumptions (chapter 1) and the state of current theoretical and practical knowledge in the field (chapter 2), the model of general reference is constructed (chapter 3) and used for annotation of the corpus of reference relations (chapter 4). The corpus provides the data for development and evaluation of automatic reference detection tools (chapters 5 and 6). The concluding chapters put the research in a broader context of discourse modelling (chapter 7) and provide a summary of the results and development perspectives.

1 Introduction

The introductory chapter brings the basic definitions of reference, anaphora, coreference and bridging, establishing the Polish terminology of the field. The motivation for the work is presented, situating the phenomenon of linguistic reference at the heart of semantic text structure and pointing out that despite numerous existing theoretical works, computational processing of reference in Polish was not systematically carried out until the recent years. Only the latest advances of corpus linguistics in Poland and the development of numerous syntactic and semantic natural language processing tools (morphological analysers, named entity recognizers, dependency parsers, etc.) made it possible to verify existing hypotheses on a much larger scale and create tools for end-to-end coreference resolution.

The scope of the work is limited to decoding nominal coreference, i.e. clustering textual fragments (*mentions*) expressed with nominal constructs (phrases, pronouns, named entities, elided subjects) into equivalence classes based on their reference to different discourse-world entities. Since coreferent mentions can be expressed with a variety of linguistic means (such as synonyms, hypernyms, neologisms, proper names or even idiolectal expressions), this task is considered one of the most difficult in natural language processing.

The methodology of this work is based on a corpus-based approach which requires establishing a formal model of reference relations, carrying out annotation of language data following this representation and using the data set for development, training and formal evaluation of implemented tools.

2 Related work

Chapter 2 starts with a short synthesis of findings from the Polish theoretical literature influencing our computational-linguistic work. The broad concept of *referencing elements* being instantiated in texts not only as single phrases, but also as larger pieces of text (whole sentences or even paragraphs) has been borrowed from Klemensiewicz (1937). The basic categorization of the reference expressions, implemented as nouns, nominal groups, proper names or pronouns was inspired by Bellert (1971).

Portions of the two most extensive taxonomies of reference relations presented by Topolińska (1984) and Paduczewa (1992) were adopted as the starting point for the final classification. The line between referencing and non-referencing expressions was drawn following the approach of Langacker (2008), Vater (2009) and Kunz (2010), assuming the prevalence of the discourse world over the real world in decoding reference, i.e. treating all nominal phrases as potentially referential. A description of indirect relations combining several existing classifications was also presented, starting from the best known one by Clark (1977) and several others, summarized in a paper by Gardent et al. (2003).

Lexical features of referential relations were inspired by the work of Pisarkowa (1969), Fontański (1986) and Grzegorczykowa (1996). Numerical features reflecting word order and inter-sentential position of mentions were taken from Szwedek (1975), Honowska (1984) and Duszak (1986). Following Gajda (1990), we analyzed the density of reference expressions depending on the text genre and similarly to Dobrzyńska (1996) we investigated the variation of stylistic quality of text (here measured by its readability).

The chapter also contains an extensive list of the largest foreign corpora annotated with referential relations and existing Polish reference-related corpora (Filak 2006, Marciniak 2010, Broda et al. 2012b). Previous attempts of anaphora and coreference resolution for Polish (Mitkov and Styś 1997, Kulików et al. 2004, Abramowicz et al. 2006, Filak 2006, Broda et al. 2012b, Kaczmarek and Marcińczuk 2017) are

investigated together with our own translation- and projection-based experiment (Ogrodniczuk 2013).

Different variants of foreign computer-based implementations of reference influencing our target solution are presented in order to outline the history of linguistic engineering work to date. Starting with early attempts using syntactic rules (Hobbs 1976), centering theory (Grosz 1977, Sidner 1979, Brennan et al. 1987) or knowledge poor approaches (Mitkov and Styś 1997), we shortly present supervised machine learning algorithms (see, e.g., Connolly et al. 1994, McCarthy and Lehnert 1995, Kehler 1997, Soon et al. 1999, 2001, Ng and Cardie 2002, Rahman and Ng 2009), high-precision sieve-based methods (see, e.g., Haghighi and Klein 2009, Raghunathan et al. 2010, Lee et al. 2011), hybrid approaches (Denis and Baldridge 2008, Chen and Ng 2012, Ratinov and Roth 2012) and deep neural network solutions (Lee et al. 2017, Zhang et al. 2018), with the best results presently achieved for English.

The chapter concludes with a short summary of evaluation methods and metrics currently used: MUC (Vilain et al. 1995), B^3 (Bagga and Baldwin 1998), CEAF-E (Luo 2005) and CoNLL (Pradhan et al. 2012) together with the two types of clustering algorithms: *mention-pair* (Aone and Bennett 1995) and *entity-based* (Luo et al. 2004).

3 Model of reference and its corpus representation

Chapter 3 presents the model of reference composed of a precise definition of mention types and scope, followed by a broad taxonomy of referential relations. Mentions are defined as generalized nominal groups (nouns with their syntactic constituents, personal pronouns, demonstratives introducing non-relative clauses, elided subjects, gerunds), possessive pronouns and undefined, negative or universal pronouns (often forming pseudoconferential clusters). The schema involves marking nested phrases and discontinuities.

The taxonomy of referential relations (see Fig. 1) is supplemented with the concept of facets representing subjectivity, uncertainty or impartiality of the parties involved in communication. Additionally, the typology presents non-referential auxiliary relations used to support the process of decoding reference.

Chapter 4 presents the process of construction of the corpus of reference relations. After establishing the annotation strategy and sampling texts from the National

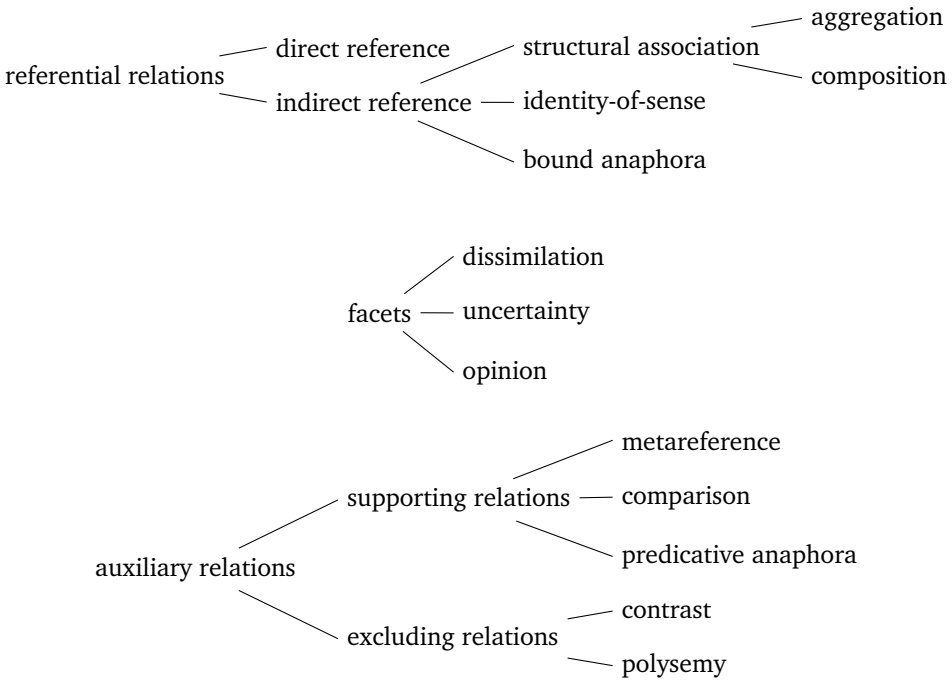


Figure 1. Referential relations, facets and auxiliary relations

Corpus of Polish (Przepiórkowski et al. 2012) the data set was manually annotated, following the designed typology and made available to download, browse and search. As a result, the Polish Coreference Corpus with 530K tokens was created — currently one of the largest coreference corpora in the world. Its basic statistics are presented in Table 1.

4 Implementation and evaluation

Chapter 5 describes the development of the mention detection and coreference resolution tools. Two mention detectors were implemented, both using input from various natural language processing components for Polish: a morphological analyser, named entity recognizer, and a shallow parser. The rule-based detector collected mention candidates from available sources and removed redundancies. The machine-learning detector used a number of lexical, grammatical and numerical features.

Table 1. Basic statistics of the Polish Coreference Corpus

Tokens	532,166
Mentions	185,802
nominal groups	137,645
named entities	20,847
zero subjects	18,807
personal pronouns	9,185
ellipses	488
Clusters	21,865
Bridging relations	
aggregation	9,213
composition	2,008
bound anaphora	317
identity-of-sense	224
Supporting relations	
predicative relation	2,350
metareference	154
comparison	101
Excluding relations	
contrast	225
polysemy	2
Facets	
no facet	16,455
dissimilation	336
opinion	188
uncertainty	66

Implemented coreference resolvers used even more decoding methodologies. The baseline rule-based system was built over a number of manually created rules (gender-number match, nesting prevention, same surface form agreement, etc.), taking into account compatibility of mention pairs and clusters. The machine-learning system used 147 mention-pair learning features representing the surface, syntactic, semantic, meta-textual and anaphoric information. The sieve-based system ordered the classifiers according to their precision and contained six simple compatibility rules. The neural system used word embeddings as components of feature vectors to train the deep neural network model.

Chapter 6 presents detailed results of the evaluation of implemented components over the test portion of the corpus, following the 10-fold cross-validation method. Mention detectors were evaluated in two settings: by comparing only semantic heads of mentions and by comparing the complete borders of mentions (see Table 2).

Table 2. Evaluation of mention detectors

Approach	**Semantic heads**	**Exact boundaries**
Rule-based	89.12%	69.10%
Machine-learning	**91.23%**	**71.79%**

Coreference resolvers were evaluated using traditional metrics in five variants, taking into account gold mentions only or system mentions in two main configurations with respect to mention borders (semantic heads only or complete borders) and the strategy of processing twinless mentions: INTERSECT (considering only mentions present in both gold and system sets) versus TRANSFORM (following the procedure by Pradhan et al. (2011), used in CoNLL-2011 shared task). Table 3 presents the values of CoNLL F_1 measure for all implemented systems.

Table 3. Evaluation of implemented coreference resolution systems

Approach	**Gold mentions**	**Semantic heads**	**Exact boundaries**	**Semantic heads**	**Exact boundaries**
		INTERSECT		**TRANSFORM**	
Rule-based	74.10%	77.05%	78.86%	72.19%	68.88%
Machine-learning	80.50%	80.00%	82.71%	74.82%	**73.96%**
Sieve	80.70%	80.85%	82.49%	75.55%	73.21%
Neural	80.59%	80.89%	**82.73%**	**76.03%**	73.39%
Hybrid	**81.09%**	**81.04%**	82.54%	75.74%	73.17%

5 Reference in discourse

Chapter 7 presents several experiments investigating reference in a broader context of discourse relations. First was the comparison of the description of Prague Discourse Treebank-compatible reference relations in Czech, Russian and Polish with the newly implemented Parallel Annotated Wall Street Journal corpus (Nedoluzhko et al. 2018). The results show variation of referential properties in different languages, both in frequency of the use of referential groups and in types of reference.

Similarly, annotation of discourse relations, including coreference, with the Penn Discourse Treebank methodology for English, German, Polish, Portuguese, Russian and Turkish was carried out, showing differences in realization of discourse relations in different languages (Zeyrek et al. 2019).

Another task focused on annotation of the Polish Coreference Corpus with event-linking time relations, communication events and relations between questions and responses to analyze explicitness and implicitness of representations of events

in the text. The results also show variation in coverage of the text with different metatextual relations which may help investigate how reference influences textual coherence and cohesion.

6 Conclusions

The concluding chapter summarizes the most important findings from the work. The presented study constitutes the first attempt of computer-based large-scale analysis of the nominal referential relations in Polish. The construction of a large corpus and decoding tools made it possible to achieve for Polish the results comparable with global developments. This was made possible by applying a series of improvements on many linguistic and technical levels, starting with the clarification of the notion of reference, anaphora and coreference, through reconstruction of the formal grammar of Polish, integration of external resources and development of new detection algorithms.

Our research, with standardization of the Polish terminology in the field of coreference, the proposed categorization of text-based reference markers and typology of referential relations contributed to the description of the problem in Polish computational linguistics. Creation of one of the world's largest representative corpora of referential relations manually annotated with coreference and bridging relations based on the texts of the reference corpus required proper selection of texts, preparation of annotation guidelines, adaptation of tools and development of the annotation methodology. Several annotation strategies were tested, demonstrating the usefulness of serial adjudication in complex semantic tasks.

The environment for annotation and presentation of the coreference data was prepared with several corpus representation formats, visualisation of referential relations and a search engine linking reference with other layers of linguistic analysis. Rule-based and machine-learning mention detectors were implemented, as well as several solutions for decoding coreference — rule-based, statistical, neural, sieve and hybrid systems, tested in various configurations and supported with external resources, such as the database of periphrastic expressions, valency dictionary or customized formal grammar of Polish. A prototype configuration for detecting associative relations was carried out to decode aggregation, composition and predictive relations. The evaluation of the resulting systems has been conducted in accordance with the commonly used metrics and methodology. In addition, qualitative analysis of created decoders was performed to reduce errors.

In cooperation with foreign partners, we have started cross-lingual research on reference relations linking the work on Polish with other languages. These activities constitute the first step towards the universal multilingual description of coreference.

The results of the presented work are also practical: they have been used, among others, in the automatic summarization system to improve text fluency by means of mention substitution.

Bibliografia

Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M., Levenberg J., Monga R., Moore S., Murray D.G., Steiner B., Tucker P., Vasudevan V., Warden P., Wicke M., Yu Y. i Zheng X. (2016). *TensorFlow: A system for large-scale machine learning* [w:] *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, s. 265–283.

Abramowicz W., Filipowska A., Piskorski J., Węcel K. i Wieloch K. (2006). *Linguistic Suite for Polish Cadastral System* [w:] Calzolari N., Choukri K., Gangemi A., Maegaard B., Mariani J., Odijk J. i Tapias D. (red.), *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 2006)*, s. 2518–2523, Genua. European Language Resources Association.

Acedański S. (2010). *A Morphosyntactic Brill Tagger for Inflectional Languages* [w:] Loftsson H., Rögnvaldsson E. i Helgadóttir S. (red.), *Advances in Natural Language Processing*, t. 6233 serii *Lecture Notes in Computer Science*, s. 3–14. Springer.

Aone C. i Bennett S.W. (1994). *Discourse tagging tool and discourse-tagged multilingual corpora* [w:] *Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR)*, s. 71–77.

Aone C. i Bennett S.W. (1995). *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies* [w:] *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL 1995, s. 122–129, Stroudsburg. Association for Computational Linguistics.

Asher N. i Lascarides A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge, Wielka Brytania.

Bagga A. i Baldwin B. (1998). *Algorithms for Scoring Coreference Chains* [w:] *Proceedings of the Workshop on Linguistic Coreference at the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, s. 563–566, Granada.

Bański P. i Przepiórkowski A. (2009). *Stand-off TEI annotation: the case of the National Corpus of Polish* [w:] *Proceedings of the 3rd Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, s. 64–67.

Baumann J., Kühling X. i Ruder S. (2014). *Rule-based coreference resolution with BART*. Plakat podsumowujący niepublikowany raport.

Baumann S. i Riester A. (2012). *Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects* [w:] Elordieta G. i Prieto P. (red.), *Prosody and Meaning*, t. 25 serii *Interface Explorations*, s. 119–162, Mouton De Gruyter.

Bayerl P.S. i Paul K.I. (2011). *What determines inter-coder agreement in manual annotations? A meta-analytic investigation*. „Computational Linguistics", 37(4), s. 699–725.

Beigman Klebanov B. i Beigman E. (2009). *From Annotator Agreement to Noise Models*. „Computational Linguisics", 35(4), s. 495–503.

Bejček E., Hajičová E., Hajič J., Jínová P., Kettnerová V., Kolářová V., Mikulová M., Mírovský J., Nedoluzhko A., Panevová J., Poláková L., Ševčíková M., Štěpánek J. i Zikánová Š. (2013). *Prague Dependency Treebank 3.0*. Uniwersytet Karola w Pradze, ÚFAL.

Bellert I. (1971). *O pewnym warunku spójności tekstu* [w:] Mayenowa M.R. (red.), *O spójności tekstu*, t. XXI, s. 47–76, Zakład Narodowy im. Ossolińskich, Wrocław.

Bengtson E. i Roth D. (2008). *Understanding the Value of Features for Coreference Resolution* [w:] *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2008, s. 294–303, Stroudsburg. Association for Computational Linguistics.

Bennet E.M., Alpert R. i Goldstein A.C. (1954). *Communications Through Limited-Response Questioning*. „Public Opinion Quarterly", 18(3), s. 303–308.

Bhardwaj V., Passonneau R.J., Salleb-Aouissi A. i Ide N. (2010). *Anveshan: A Framework for Analysis of Multiple Annotators' Labeling Behavior* [w:] *Proceedings of the 4th Linguistic Annotation Workshop (LAW 2010)*, s. 47–55, Stroudsburg. Association for Computational Linguistics.

Björkelund A., Eckart K., Riester A., Schauffler N. i Schweitzer K. (2014). *The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution* [w:] Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Rejkiawik. European Language Resources Association.

Black M. (1949). *Language and philosophy: Studies in method*. Cornell University Press.

Bobrow D.G. (1964). *A Question-answering System for High School Algebra Word Problems* [w:] *Proceedings of the Fall Joint Computer Conference, Part I*, (AFIPS 1964), s. 591–614, Nowy Jork. ACM.

Brennan S.E., Friedman M.W. i Pollard C.J. (1987). *A centering approach to pronouns* [w:] *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, ACL 1987, s. 155–162, Stroudsburg. Association for Computational Linguistics.

Broda B., Burdka Ł. i Maziarz M. (2012a). *IKAR: An Improved Kit for Anaphora Resolution for Polish* [w:] *Proceedings of COLING 2012: Demonstration Papers*, s. 25–32.

Broda B., Marcińczuk M., Maziarz M., Radziszewski A. i Wardyński A. (2012b). *KPWr: Towards a Free Corpus of Polish* [w:] Calzolari N., Choukri K., Declerck T., Dogan M.U., Maegaard B., Mariani J., Odijk J. i Piperidis S. (red.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, s. 3218–3222, Stambuł. European Language Resources Association.

Bromley J., Guyon I., LeCun Y., Säckinger E. i Shah R. (1994). *Signature Verification using a "Siamese" Time Delay Neural Network* [w:] Cowan J.D., Tesauro G. i Alspector J. (red.), *Advances in Neural Information Processing Systems 6*, s. 737–744. Morgan-Kaufmann.

Brouwer M., Brugman H. i Kemps-Snijders M. (2017). *MTAS: A Solr/Lucene based Multi Tier Annotation Search solution* [w:] *Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings 136*, s. 19–37. Linköping University Electronic Press.

Bunescu R. (2003). *Associative Anaphora Resolution: A Web-Based Approach* [w:] *Proceedings of the EACL-2003 Workshop on the Computational Treatment of Anaphora*, s. 47–52, Budapeszt.

Burnard L. i Bauman S., red. (2007). *Guidelines for Electronic Text Encoding and Interchange (TEI P5)*. The TEI Consortium.

Cahill A. i Riester A. (2012). *Automatically Acquiring Fine-Grained Information Status Distinctions in German* [w:] *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, s. 232–236. Association for Computational Linguistics.

Cai J. i Strube M. (2010). *Evaluation Metrics for End-to-end Coreference Resolution Systems* [w:] *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL 2010, s. 28–36, Stroudsburg. Association for Computational Linguistics.

Carnap R. (1947). *Meaning and Necessity*. University of Chicago Press, Chicago.

Caselli T. i Prodanof I. (2006). *Annotating Bridging Anaphors in Italian: in Search of Reliability* [w:] *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, s. 1173–1176, Genua. European Language Resources Association.

Cettolo M., Girardi C. i Federico M. (2012). *WIT3: Web Inventory of Transcribed and Translated Talks* [w:] *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012)*, s. 261–268.

Chamberlain J., Poesio M. i Kruschwitz U. (2016). *Phrase Detectives Corpus 1.0: Crowdsourced Anaphoric Coreference* [w:] Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, s. 2039–2046, Portorož. European Language Resources Association.

Chen C. i Ng V. (2012). *Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution* [w:] *Joint Conference on EMNLP and CoNLL: Proceedings of the Shared Task*, s. 56–63.

Chollet F. (2015). *Keras*. `https://keras.io`.

Ciura M., Grund D., Kulików S. i Suszczańska N. (2004). *A System to Adapt Techniques of Text Summarizing to Polish* [w:] Okatan A. (red.), *International Conference on Computational Intelligence*, s. 117–120, Stambuł. International Computational Intelligence Society.

Clark H.H. (1977). *Bridging* [w:] Johnson-Laird P. i Wason P.C. (red.), *Thinking: Readings in Cognitive Science*, s. 411–420. Cambridge University Press.

Clark K. i Manning C.D. (2016a). *Deep Reinforcement Learning for Mention-Ranking Coreference Models* [w:] *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, s. 2256–2262, Austin. Association for Computational Linguistics.

Clark K. i Manning C.D. (2016b). *Improving coreference resolution by learning entity-level distributed representations* [w:] *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, s. 643–653, Berlin. Association for Computational Linguistics.

Connolly D., Burger J.D. i Day D.S. (1994). *A Machine Learning Approach to Anaphoric Reference* [w:] *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, s. 255–261, ACL.

Cristea D. i Postolache O.D. (2005). *How to Deal with Wicked Anaphora?* „Current Issues in Linguistic Theory", 263, s. 17–46.

Cunningham H., Maynard D., Bontcheva K. i Tablan V. (2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications* [w:] *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, s. 168–175.

Data-Bukowska E. (2008). *O funkcjonowaniu zaimkowych odniesień anaforycznych w języku polskim – analiza z perspektywy językoznawstwa kognitywnego*. „Studia Linguistica Universitatis Iagellonicae Cracoviensis", 125, s. 51–65.

Day D., Aberdeen J., Caskey S., Hirschman L., Robinson P. i Vilain M. (1998). *Alembic Workbench Corpus Development Tool* [w:] *Proceedings of the 1st International Conference on Language Resource and Evaluation*, s. 1021–1028.

Day D., Mchenry C., Kozierok R. i Riek L. (2004). *Callisto: A configurable annotation workbench* [w:] *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*, s. 2073–2076, Lizbona.

Denis P. i Baldridge J. (2008). *Specialized Models and Ranking for Coreference Resolution* [w:] *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, s. 660–669, Honolulu, Hawaii. Association for Computational Linguistics.

Dobrzyńska T. (1996). *Tekst i jego odmiany: zbiór studiów* [w:] Dobrzyńska T. (red.), *Tekst – w perspektywie stylistycznej*, s. 125–143, Instytut Badań Literackich PAN, Warszawa.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S. i Weischedel R. (2004). *The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation* [w:] Lino M.T., Xavier M.F., Ferreira F., Costa R. i Silva R. (red.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, s. 837–840, Lizbona. European Language Resources Association.

Dubisz S., red. (2006). *Uniwersalny słownik języka polskiego PWN*. Wydawnictwo Naukowe PWN, Warszawa. t. 1–4.

Durrett G. i Klein D. (2013). *Easy Victories and Uphill Battles in Coreference Resolution* [w:] *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, s. 1971–1982, Seattle, Washington. Association for Computational Linguistics.

Duszak A. (1986). *Niektóre uwarunkowania semantyczne szyku wyrazów w zdaniu polskim*. „Polonica", XII(12), s. 59–74.

Eckart K., Riester A. i Schweitzer K. (2012). *A Discourse Information Radio News Database for Linguistic Analysis* [w:] Chiarcos C., Nordhoff S. i Hellmann S. (red.), *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, s. 65–76, Springer, Berlin, Heidelberg.

Evans G. (1977). *Pronouns, Quantifiers, and Relative Clauses (I)*. „Canadian Journal of Philosophy", VII(3), s. 467–536.

Fall J. (1994). *Anafora i jej zatarte granice*. „Studia Semiotyczne", XIX/XX, s. 163–191.

Fan J., Barker K. i Porter B. (2005). *Indirect Anaphora Resolution as Semantic Path Search* [w:] *Proceedings of 3rd International Conference on Knowledge Capture (K–CAP'05)*, s. 153–160, ACM.

Fauconnier G. (1985). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. MIT Press, Cambridge.

Fauconnier G. i Turner M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, Nowy Jork.

Filak T. (2006). *Zastosowanie metod automatycznego uczenia do rozstrzygania problemu anafory*. Praca magisterska, Wydział Informatyki i Zarządzania Politechniki Wrocławskiej, Wrocław.

Fleiss J.L. (1971). *Measuring Nominal Scale Agreement Among Many Raters*. „Psychological Bulletin", 76, s. 378–382.

Fontański H. (1986). *Anaforyczne przymiotniki wskazujące w języku polskim i rosyjskim: problem użycia*. Prace naukowe Uniwersytetu Śląskiego w Katowicach. Uniwersytet Śląski.

Fort K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley.

Fort K. i Sagot B. (2010). *Influence of Pre-annotation on POS-tagged Corpus Development* [w:] *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, s. 56–63, Stroudsburg. Association for Computational Linguistics.

Fraurud K. (1990). *Definiteness and the Processing of Noun Phrases in Natural Discourse*. „Journal of Semantics", 7, s. 395–433.

Frege G. (1892). *Über Sinn und Bedeutung*. „Zeitschrift für Philosophie und philosophische Kritik", 100, s. 25–50.

Gajda S. (1982). *Podstawy badań stylistycznych nad językiem naukowym*. Państwowe Wydawnictwo Naukowe, Wrocław.

Gajda S. (1990). *Współczesna polszczyzna naukowa: język czy żargon?* Instytut Śląski, Opole.

Gardent C., Manuélian H. i Kow E. (2003). *Which bridges for bridging definite descriptions?* [w:] *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora (LINC'03)*, s. 69–76.

Gardent C., Manuélian H. i Pontoise C. (2005). *Création d'un corpus annoté pour le traitement des descriptions définies*. „Traitement Automatique des Langues", 46(1), s. 115–140.

Głowińska K. (2012). *Anotacja składniowa* [w:] Przepiórkowski A., Bańko M., Górski R.L. i Lewandowska-Tomaszczyk B. (red.), *Narodowy Korpus Języka Polskiego*, s. 107–127. Wydawnictwo Naukowe PWN, Warszawa.

Górski R.L. i Łaziński M. (2012). *Reprezentatywność i zrównoważenie korpusu* [w:] Przepiórkowski i in. (2012), s. 25–36.

Grochowski M. (1976). *O pojęciu elipsy*. „Pamiętnik Literacki", LXVII(1), s. 121–136.

Grochowski M., Kisiel A. i Żabowska M. (2014). *Słownik gniazdowy partykuł polskich*. Polska Akademia Umiejętności, Kraków.

Grosz B.J. (1977). *The Representation and Use of Focus in Dialogue Understanding*. Rozprawa doktorska, University of California, Berkeley.

Grosz B.J., Weinstein S. i Joshi A.K. (1995). *Centering: A Framework for Modeling the Local Coherence of Discourse*. „Computational Linguistics", 21(2), s. 203–226.

Gruszczyński W. i Ogrodniczuk M., red. (2015). *Jasnopis, czyli mierzenie zrozumiałości polskich tekstów użytkowych*. Wydawnictwo ASPRA-JR, Warszawa.

Grzegorczykowa R. (1990). *Wprowadzenie do semantyki językoznawczej*. Państwowe Wydawnictwo Naukowe, Warszawa.

Grzegorczykowa R. (1996). *Polskie leksemy z wbudowaną informacją anaforyzacyjną* [w:] Grochowski M. (red.), *Anafora w strukturze tekstu*, s. 71–77. Wydawnictwo Energeia, Warszawa.

Guillou L., Hardmeier C., Smith A., Tiedemann J. i Webber B. (2014). *ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT* [w:] Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, s. 3191–3198, European Language Resources Association.

Gundel J.K., Hedberg N. i Zacharski R. (1993). *Cognitive Status and the Form of Referring Expressions in Discourse*. „Language", 69(2), s. 274–307.

Haghighi A. i Klein D. (2007). *Unsupervised Coreference Resolution in a Nonparametric Bayesian Model* [w:] Carroll J.A., van den Bosch A. i Zaenen A. (red.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, s. 848–855, Association for Computational Linguistics.

Haghighi A. i Klein D. (2009). *Simple Coreference Resolution with Rich Syntactic and Semantic Features* [w:] *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, s. 1152–1161, Singapur. Association for Computational Linguistics.

Hahn U., Strube M. i Markert K. (1996). *Bridging Textual Ellipses* [w:] *Proceedings of the 16th Conference on Computational Linguistics – Volume 1*, COLING '96, s. 496–501, Stroudsburg. Association for Computational Linguistics.

Hajnicz E., Nitoń B., Patejuk A., Przepiórkowski A. i Woliński M. (2015). *Internetowy słownik walencyjny języka polskiego oparty na danych korpusowych*. „Prace Filologiczne", LXV, s. 95–110.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. i Witten I.H. (2009). *The WEKA Data Mining Software: An Update*. „ACM SIGKDD Explorations Newsletter", 11(1), s. 10–18.

Harabagiu S.M., Bunescu R.C. i Ştefan T.M. (2001). *COREFDRAW: a tool for annotation and visualization of coreference data* [w:] *Proceedings of the 13th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2001)*, s. 273–279.

Hendrickx I., Bouma G., Daelemans W., Hoste V., Kloosterman G., Mineur A.M., Van J., Vloet D. i Verschelde J.L. (2008). *A Coreference Corpus and Resolution System for Dutch* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S. i Tapias D. (red.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, s. 144–149, Marakesz. European Language Resources Association.

Hendrickx I., De Clercq O. i Hoste V. (2011). *Analysis and Reference Resolution of Bridge Anaphora Across Different Text Genres* [w:] *Proceedings of the Eighth International Conference on Anaphora Processing and Applications (DAARC 2011)*, s. 1–11, Berlin, Heidelberg. Springer-Verlag.

Hirst G. (1981). *Anaphora in Natural Language Understanding: A Survey*, t. 119 serii *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg/Nowy Jork.

Hobbs J.R. (1976). *Pronoun Resolution*. Technical report, Department of Computer Science, City College, City University of New York.

Hobbs J.R. (1978). *Resolving Pronoun References*. „Lingua", 44, s. 311–338.

Hodosh M., Young P., Rashtchian C. i Hockenmaier J. (2010). *Cross-caption coreference resolution for automatic image understanding* [w:] *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL 2010*, s. 162–171, Stroudsburg. Association for Computational Linguistics.

Holen G.I. (2013). *Critical Reflections on Evaluation Practices in Coreference Resolution* [w:] *Proceedings of the 2013 NAACL HLT Student Research Workshop*, s. 1–7, Atlanta, Georgia. Association for Computational Linguistics.

Honowska M. (1984). *Grzybnia zaimkowa. Przyczynek do zagadnień spójności tekstu*. „Polonica", X, s. 111–120.

Hou Y., Markert K. i Strube M. (2013). *Global Inference for Bridging Anaphora Resolution* [w:] *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, s. 907–917, Atlanta, Georgia. Association for Computational Linguistics.

Hou Y., Markert K. i Strube M. (2018). *Unrestricted bridging resolution*. „Computational Linguistics", 44(2), s. 237–284.

Hovy D., Berg-Kirkpatrick T., Vaswani A. i Hovy E. (2013). *Learning Whom to Trust with MACE* [w:] *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, s. 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Iida R., Komachi M., Inui K. i Matsumoto Y. (2007). *Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations* [w:] *Proceedings of the Linguistic Annotation Workshop (LAW 2007)*, s. 132–139, Stroudsburg. Association for Computational Linguistics.

Ioffe S. i Szegedy C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* [w:] Bach F.R. i Blei D.M. (red.), *ICML*, t. 37 serii *JMLR Workshop and Conference Proceedings*, s. 448–456. JMLR.org.

Janssen T. (1980). *Coreference and Interreference in Anaphoric Relations: Grammatical Semantics or Pragmatics?* [w:] van der Auwera J. (red.), *The Semantics of Determiners*, t. 24 serii *Routledge Library Editions: Linguistics*, s. 67–80. Croom Helm London, University Park Press Baltimore.

Kaczmarek A. i Marcińczuk M. (2015a). *Evaluation of Coreference Resolution Tools for Polish from the Information Extraction Perspective* [w:] *The 5th Workshop on Balto-Slavic Natural Language Processing*, s. 24–33, Hissar, Bułgaria. INCOMA Ltd. Shoumen.

Kaczmarek A. i Marcińczuk M. (2015b). *Heuristic Algorithm for Zero Subject Detection in Polish* [w:] Král P. i Matoušek V. (red.), *Proceedings of the 18th International Conference on Text, Speech, and Dialogue (TSD 2015)*, LNAI 9302, s. 378–386. Springer International Publishing.

Kaczmarek A. i Marcińczuk M. (2017). *A preliminary study in zero anaphora coreference resolution for Polish*. „Cognitive Studies", (17), s. 1–13.

Karttunen L. (1976). *Discourse Referents* [w:] McCawley J.D. (red.), *Syntax and Semantics 7: Notes from the Linguistic Underground*, s. 363–386. Academic Press, Nowy Jork.

Kehler A. (1997). *Probabilistic Coreference in Information Extraction* [w:] *Proceedings of the 2nd Conference on Empirical Methods in NLP (EMNLP-2)*, s. 163–173.

Kingma D.P. i Ba J.L. (2015). *Adam: A Method for Stochastic Optimization* [w:] *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015)*.

Klemensiewicz Z. (1937). *Składnia opisowa współczesnej polszczyzny kulturalnej*. Polska Akademia Umiejętności, Kraków.

Klemensiewicz Z. (1948). *Syntaktyczny stosunek nawiązania*. „Sprawozdania z Czynności i Posiedzeń PAU", XLVIII(6), s. 214–217.

Klemensiewicz Z. (1950). *O syntaktycznym stosunku nawiązania*. „Slavia", XIX, s. 13–27.

Klemensiewicz Z. (1953). *Zarys składni polskiej*. Państwowe Wydawnictwo Naukowe, Warszawa.

Klemensiewicz Z. (1982). *O syntaktycznym stosunku nawiązania* [w:] Kałkowska A. (red.), *Składnia, stylistyka, pedagogika językowa*, Biblioteka Filologii Polskiej: Językoznawstwo, s. 241–257, Państwowe Wydawnictwo Naukowe, Warszawa.

Kopeć M. (2014). *Zero subject detection for Polish* [w:] *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, s. 221–225, Gothenburg, Sweden. Association for Computational Linguistics.

Kopeć M. (2018). *Summarization of Polish Press Articles Using Coreference*. Rozprawa doktorska, Instytut Podstaw Informatyki PAN.

Kopeć M. i Ogrodniczuk M. (2012). *Creating a Coreference Resolution System for Polish* [w:] Calzolari N., Choukri K., Declerck T., Dogan M.U., Maegaard B., Mariani J., Odijk J. i Piperidis S. (red.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, s. 192–195, Stambuł. European Language Resources Association.

Korzen I. i Buch-Kromann M. (2011). *Anaphoric Relations in the Copenhagen Dependency Treebanks* [w:] Dipper S. i Zinsmeister H. (red.), *Beyond Semantics Corpus-based Investigations of Pragmatic and Discourse Phenomena*, t. 3, s. 83–98, Göttingen. Ruhr-Universität Bochum, Sprachwissenschaftliches Institut.

Krasavina O. i Chiarcos C. (2007). *PoCoS – Potsdam Coreference Scheme*. [w:] Boguraev B., Ide N., Meyers A., Nariyama S., Stede M., Wiebe J. i Wilcock G. (red.), *Proceedings of the Linguistic Annotation Workshop*, s. 156–163. Association for Computational Linguistics.

Kripke S. (2001). *Nazywanie a konieczność*. Fundacja Aletheia, Warszawa.

Krug M., Puppe F., Jannidis F., Macharowsky L., Reger I. i Weimar L. (2015). *Rule-based Coreference Resolution in German Historic Novels* [w:] *Proceedings of the 4th Workshop on Computational Linguistics for Literature*, s. 98–104, Denver, Colorado. Association for Computational Linguistics.

Kulików S., Romaniuk J. i Suszczańska N. (2004). *A syntactical analysis of anaphora in the Polsyn parser* [w:] Kłopotek M.A., Wierzchoń S.T. i Trojanowski K. (red.), *Intelligent Information Processing and Web Mining*, t. 25 serii *Advances in Soft Computing*, s. 444–448. Springer Berlin Heidelberg.

Kunz K., Lapshinova-Koltunski E. i Martínez J.M. (2016). *Beyond identity coreference: Contrasting indicators of textual coherence in English and German* [w:] Ogrodniczuk M. i Ng V. (red.), *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON@NAACL-HLT 2016)*, s. 23–31, San Diego. The Association for Computational Linguistics.

Kunz K.A. (2010). *Variation in English and German Nominal Coreference: A Study of Political Essays*. Saarbrücker Beiträge zur Sprach- und Translationswissenschaft. Peter Lang, Frankfurt/Berlin/Berno/Bruksela/Nowy Jork/Oxford/Wiedeń.

Lakoff G. i Johnson M. (1988). *Metafory w naszym życiu*. PIW, Warszawa.

Langacker R.W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press.

Lapshinova-Koltunski E. i Kunz K. (2014). *Annotating cohesion for multillingual analysis*. [w:] *Proceedings of the 10th Joint ACL–ISO Workshop on Interoperable Semantic Annotation*, s. 57–64, Rejkiawik. European Language Resources Association.

Lapshinova-Koltunski E., Kunz K.A. i Nedoluzhko A. (2016). *From interoperable annotations towards interoperable resources: A multilingual approach to the analysis of discourse* [w:] Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, s. 991–997, Portorož. European Language Resources Association.

Lassalle E. i Denis P. (2011). *Leveraging Different Meronym Discovery Methods for Bridging Resolution in French* [w:] Hendrickx I., Lalitha Devi S., Branco A. i Mitkov R. (red.), *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011). Revised Selected Papers*, s. 35–46. Springer Berlin Heidelberg.

Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M. i Jurafsky D. (2011). *Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task* [w:] *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, CoNLL Shared Task 2011, s. 28–34, Stroudsburg. Association for Computational Linguistics.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M. i Jurafsky D. (2013). *Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules*. „Computational Linguistics", 39(4), s. 885–916.

Lee K., He L., Lewis M. i Zettlemoyer L. (2017). *End-to-end Neural Coreference Resolution* [w:] *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, s. 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Leech G. (1997). *Introducing corpus annotation*. [w:] Garside R., Leech G. i McEnery T. (red.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Pearson Education, s. 1–18. Longman, London.

Luo X. (2005). *On Coreference Resolution Performance Metrics* [w:] *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT 2005, s. 25–32, Vancouver. Association for Computational Linguistics.

Luo X., Ittycheriah A., Jing H., Kambhatla N. i Roukos S. (2004). *A Mention-synchronous Coreference Resolution Algorithm Based on the Bell Tree* [w:] *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, s. 135–142, Stroudsburg. Association for Computational Linguistics.

Luo X., Pradhan S., Recasens M. i Hovy E. (2014). *An Extension of BLANC to System Mentions* [w:] *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, s. 24–29, Baltimore, Maryland. Association for Computational Linguistics.

Lyons J. (1977). *Semantics*, t. 1. Cambridge University Press.

Maillard M. (1974). *Essai de typologie des substituts diaphoriques*. „Langue française", 21(1), s. 55–71.

Mann W.C. i Thompson S.A. (1988). *Rhetorical structure theory: Toward a functional theory of text organization*. „Text, Interdisciplinary Journal for the Study of Discourse", 8(3), s. 243–281.

Marcińczuk M., Kocoń J. i Broda B. (2012). *Inforex – a web-based tool for text corpus management and semantic annotation* [w:] Calzolari N., Choukri K., Declerck T., Dogan M.U., Maegaard B., Mariani J., Odijk J. i Piperidis S. (red.), *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, Stambuł. European Language Resources Association.

Marciniak M. (2001). *Algorytmy implementacyjne syntaktycznych reguł koreferencji zaimków dla języka polskiego w terminach HPSG*. Rozprawa doktorska, Instytut Podstaw Informatyki PAN, Warszawa.

Marciniak M., red. (2010). *Anotowany korpus dialogów telefonicznych*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Marciszewski W. (1983). *Spójność strukturalna a spójność semantyczna* [w:] Dobrzyńska T. i Janus E. (red.), *Tekst i zdanie*, s. 183–189. Zakład Narodowy im. Ossolińskich, Wrocław.

Markert K., Nissim M. i Modjeska N.N. (2003). *Using the Web for Nominal Anaphora Resolution* [w:] *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, s. 39–46, Budapeszt.

Màrquez L., Recasens M. i Sapena E. (2012). *Coreference Resolution: An Empirical Study Based on SemEval-2010 Shared Task 1*. „Language Resources and Evaluation", 47, s. 1–34.

Matysiak I. (2007). *Information Extraction Systems and Nominal Anaphora Analysis Needs* [w:] *Proceedings of the International Multiconference on Computer Science and Information Technology*, s. 183–192.

Maziarz M., Piekot T., Poprawa M., Broda B., Radziszewski A. i Zarzeczny G. (2012). *Język raportów ewaluacyjnych*. Ministerstwo Rozwoju Regionalnego. Departament Koordynacji Polityki Strukturalnej, Warszawa.

Maziarz M., Marcińczuk M., Oleksy M., Piasecki M., Radziszewski A., Nowak J., Wardyński A. i Wieczorek J. (2016). *KPWr annotation guidelines – coreference*. CLARIN-PL digital repository.

McCarthy J.F. i Lehnert W.G. (1995). *Using Decision Trees for Coreference Resolution* [w:] *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, s. 1050–1055, Montreal.

McKelvie D., Isard A., Mengel A., Møller M.B., Grosse M. i Klein M. (2001). *The MATE workbench – An annotation tool for XML coded speech corpora*. „Speech Communication", 33(1–2), s. 97–112.

Mikolov T., Deoras A., Povey D., Burget L. i Černockỳ J. (2011). *Strategies for training large scale neural network language models* [w:] *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, s. 196–201.

Mill J.S. (1843). *A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, t. 1. John W. Parker, Londyn.

Mitkov R. i Styś M. (1997). *Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish* [w:] *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*, s. 74–81.

Mitkov R., Belguith L. i Styś M. (1998). *Multilingual Robust Anaphora Resolution* [w:] *Proceedings of the 3rd International Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, s. 7–16, Granada.

Mitkov R., Evans R., Orăsan C., Ha L.A. i Pekar V. (2007). *Anaphora Resolution: To What Extent Does It Help NLP Applications?* [w:] Branco A. (red.), *Anaphora: Analysis, Algorithms and Applications*, s. 179–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Moosavi N.S. i Strube M. (2016). *Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric* [w:] *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, s. 632–642, Berlin. Association for Computational Linguistics.

Morton T. i LaCivita J. (2003). *WordFreak: An Open Tool for Linguistic Annotation* [w:] *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, s. 17–18.

Müller C. i Strube M. (2001). *MMAX: A Tool for the Annotation of Multi-modal Corpora* [w:] *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, s. 45–50.

Müller C. i Strube M. (2006). *Multi-level annotation of linguistic data with MMAX2* [w:] Braun S., Kohn K. i Mukherjee J. (red.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, s. 197–214, Peter Lang, Frankfurt.

Muzerelle J., Lefeuvre A., Antoine J.Y., Schang E., Maurel D., Villaneau J. i Eshkol I. (2013). *ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement* [w:] *Proceedings of the 20th Conference Traitement Automatique des Langues Naturelles (TALN 2013)*, s. 555–563, Les Sables d'Olonne.

Nair V. i Hinton G.E. (2010). *Rectified Linear Units Improve Restricted Boltzmann Machines* [w:] Fürnkranz J. i Joachims T. (red.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, s. 807–814, Hajfa. Omnipress.

Nedoluzhko A., Mírovský J., Ocelák R. i Pergler J. (2009). *Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank* [w:] *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, s. 1–16, AU-KBC Research Centre, Anna University, Chennai.

Nedoluzhko A., Novák M., Cinková S., Mikulová M. i Mírovský J. (2016). *Coreference in Prague Czech-English Dependency Treebank* [w:] *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, s. 169–176, Portorož. European Language Resources Association.

Nedoluzhko A., Novák M. i Ogrodniczuk M. (2018). *PAWS: A Multi-lingual Parallel Treebank with Anaphoric Relations* [w:] *Proceedings of the NAACL-HLT Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2018)*, s. 68–76, Nowy Orlean. Association for Computational Linguistics.

Ng V. i Cardie C. (2002). *Improving Machine Learning Approaches to Coreference Resolution* [w:] *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 2002, s. 104–111, Stroudsburg. Association for Computational Linguistics.

Nissim M., Dingare S., Carletta J. i Steedman M. (2004). *An Annotation Scheme for Information Status in Dialogue* [w:] *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, s. 1023–1026, Lizbona. European Language Resources Association.

Nitoń B. (2016). *Evaluation of Uryupina's coreference resolution features for Polish* [w:] Vetulani Z., Uszkoreit H. i Kubis M. (red.), *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7–9, 2013. Revised Selected Papers*, t. 9561 serii *Lecture Notes in Artificial Intelligence*, s. 354–367, Switzerland. Springer International Publishing.

Nitoń B., Morawiecki P. i Ogrodniczuk M. (2018). *Deep Neural Networks for Coreference Resolution for Polish* [w:] Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, s. 395–400, European Language Resources Association.

Nitoń B. i Ogrodniczuk M. (2017). *Multi-pass Sieve Coreference Resolution System for Polish* [w:] Gracia J., Bond F., McCrae J.P., Buitelaar P., Chiarcos C. i Hellmann S. (red.), *Proceedings of the 1st Conference on Language, Data and Knowledge (LDK 2017)*, t. 10318 serii *Lecture Notes in Artificial Intelligence*, s. 1–15, Springer Berlin Heidelberg.

Nivre J., de Marneffe M., Ginter F., Goldberg Y., Hajic J., Manning C.D., McDonald R.T., Petrov S., Pyysalo S., Silveira N., Tsarfaty R. i Zeman D. (2016). *Universal Dependencies v1: A Multilingual Treebank Collection* [w:] Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, s. 1659–1666, Portorož. European Language Resources Association.

Novák M. i Nedoluzhko A. (2015). *Correspondences between Czech and English Coreferential Expressions*. „Discours: Revue de linguistique, psycholinguistique et informatique", 16, s. 1–41.

Och F.J. i Ney H. (2000). *Improved Statistical Alignment Models* [w:] *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL 2000, s. 440–447, Stroudsburg. Association for Computational Linguistics.

O'Donnell M.J. (2008). *The UAM CorpusTool: Software for corpus annotation and exploration* [w:] Bretones Callejas C.M. (red.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*, s. 1433–1447, Universidad de Almería.

Ogrodniczuk M. (2013). *Translation- and Projection-Based Unsupervised Coreference Resolution for Polish* [w:] Kłopotek M.A., Koronacki J., Marciniak M., Mykowiecka A. i Wierzchoń S.T. (red.), *Proceedings of the 20th International Conference Intelligent Information Systems*, t. 7912 serii *Lecture Notes in Computer Science*, s. 125–130, Springer-Verlag, Berlin, Heidelberg.

Ogrodniczuk M. (2017). *Lingwistyka komputerowa dla języka polskiego: dziś i jutro*. „Język Polski", XCVII(1), s. 18–28.

Ogrodniczuk M. i Kopeć M. (2011a). *End-to-end coreference resolution baseline system for Polish* [w:] Vetulani Z. (red.), *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2011)*, s. 167–171, Poznań. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

Ogrodniczuk M. i Kopeć M. (2011b). *Rule-based coreference resolution module for Polish* [w:] *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, s. 191–200, Faro.

Ogrodniczuk M. i Nitoń B. (2017). *Improving Polish Mention Detection with Valency Dictionary* [w:] *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, s. 17–23, Walencja. Association for Computational Linguistics.

Ogrodniczuk M. i Zawisławska M. (2016). *Bridging Relations in Polish: Adaptation of Existing Typologies* [w:] Ogrodniczuk M. i Ng V. (red.), *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, s. 16–22, San Diego. Association for Computational Linguistics.

Ogrodniczuk M., Wójcicka A., Głowińska K. i Kopeć M. (2014). *Detection of Nested Mentions for Coreference Resolution in Polish* [w:] Ogrodniczuk M. i Przepiórkowski A. (red.), *Advances in Natural Language Processing: Proceedings of the 9th International Conference on NLP, PolTAL 2014*, t. 8686 serii *Lecture Notes in Computer Science*, s. 270–277, Warszawa. Springer International Publishing.

Ogrodniczuk M., Głowińska K., Kopeć M., Savary A. i Zawisławska M. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter, Berlin/Boston/Monachium.

Orăsan C. (2003). *PALinkA: a highly customizable tool for discourse annotation* [w:] *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, s. 39–43, Sapporo.

Orăsan C., Cristea D., Mitkov R. i Branco A. (2008). *Anaphora Resolution Exercise: An Overview* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S. i Tapias D. (red.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, s. 2801–2805, Marakesz. European Language Resources Association.

Oza U., Prasad R., Kolachina S., Sharma D.M. i Joshi A.K. (2009). *The Hindi Discourse Relation Bank* [w:] *Proceedings of the 3rd Linguistic Annotation Workshop (LAW 2009)*, s. 158–161, Singapur. The Association for Computer Linguistics.

Paduczewa J. (1992). *Wypowiedź i jej odniesienie do rzeczywistości. (Referencyjne aspekty znaczenia zaimków)*. PWN, Warszawa.

Pajas P. i Štěpánek J. (2008). *Recent Advances in a Feature-rich Framework for Treebank Annotation* [w:] *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1*, s. 673–680, Stroudsburg. Association for Computational Linguistics.

Panevová J., Hajičová E. i Sgall P. (2000). *Coreference in Annotating a Large Corpus* [w:] Gavrilidou M., Carayannis G., Markantonatou S., Piperidis S. i Stainhaouer G. (red.), *Proceedings of the 2nd International Conference on Language Resources*, t. I, s. 497–500, Ateny. European Language Resources Association.

Pasek J. (1991). *Anafora* [w:] Pelc J. (red.), *Prace z pragmatyki, semantyki i metodologii semiotyki*, Biblioteka myśli semiotycznej, s. 275–286, Ossolineum, Wrocław.

Piasecki M., Szpakowicz S. i Broda B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Pisarek J. (2012). *Językowe mechanizmy nawiązania w tekstach publicystycznych na przykładzie felietonów „Tygodnika Powszechnego"*. Rozprawa doktorska, Wydział Polonistyki Uniwersytetu Jagiellońskiego, Kraków.

Pisarkowa K. (1969). *Funkcje składniowe polskich zaimków odmiennych*. Prace Komisji Językoznawstwa nr 22. Zakład Narodowy im. Ossolińskich. Polska Akademia Nauk, Oddział w Krakowie.

Poesio M. (2000). *The GNOME annotation scheme manual*. Technical report, University of Essex, United Kingdom.

Poesio M. i Artstein R. (2008). *Anaphoric Annotation in the ARRAU Corpus* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S. i Tapias D. (red.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, s. 1170–1174, Marakesz. European Language Resources Association.

Poesio M., Vieira R. i Teufel S. (1997). *Resolving Bridging References in Unrestricted Text* [w:] *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts (ANARESOLUTION '97)*, s. 1–6, Stroudsburg. Association for Computational Linguistics.

Poesio M., Ishikawa T., Schulte im Walde S. i Vieira R. (2002). *Acquiring Lexical Knowledge for Anaphora Resolution* [w:] *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, s. 1220–1224, Las Palmas.

Poesio M., Mehta R., Maroudas A. i Hitzeman J. (2004). *Learning to Resolve Bridging References* [w:] *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004), Main Volume*, s. 143–150, Barcelona.

Poesio M., Chamberlain J., Kruschwitz U., Robaldo L. i Ducceschi L. (2015). *Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation* [w:] *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI 2015)*, s. 4202–4206, Buenos Aires. AAAI Press.

Poláková L., Jínová P., Zikánová Š., Hajičová E., Mírovský J., Nedoluzhko A., Rysová M., Pavlíková V., Zdeňková J., Pergler J. i Ocelák R. (2012). *Prague Discourse Treebank 1.0*. Biblioteka Cyfrowa LINDAT/CLARIN w Instytucie Lingwistyki Formalnej i Stosowanej, Uniwersytet Karola.

Poon H. i Domingos P. (2008). *Joint Unsupervised Coreference Resolution with Markov Logic* [w:] *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, s. 650–659, Honolulu, Hawaii. Association for Computational Linguistics.

Posturzyńska-Bosko M. (2015). *Instrumenty spójności tekstu w dziele „Le livre des fais et bonnes meurs du sage roy Charles V" średniowiecznej francuskiej pisarki Christine de Pizan*. „Acta Universitatis Lodziensis. Studia Indogermanica Lodziensia. Supplementary Series", (4), s. 183–194.

Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R. i Xue N. (2011). *CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes* [w:] *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, CoNLL Shared Task 2011, s. 1–27, Stroudsburg. Association for Computational Linguistics.

Pradhan S., Moschitti A., Xue N., Uryupina O. i Zhang Y. (2012). *CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes* [w:] *Proceedings of the 16th Conference on Computational Natural Language Learning (CoNLL 2012)*, s. 1–40, Jeju.

Pradhan S.S., Ramshaw L., Weischedel R., MacBride J. i Micciulla L. (2007). *Unrestricted Coreference: Identifying Entities and Events in OntoNotes* [w:] *Proceedings of the 1st IEEE International Conference on Semantic Computing (ICSC 2007)*, s. 446–453, Irvine. IEEE Computer Society.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A. i Webber B. (2008). *The Penn Discourse TreeBank 2.0* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S. i Tapias D. (red.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, s. 2961–2968, Marakesz. European Language Resources Association.

Prasad R., Webber B. i Joshi A. (2014). *Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation*. „Computational Linguistics", 40(4), s. 921–950.

Presspublica (2002). *Rzeczpospolita Corpus* [zasób elektroniczny]. `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.

Przepiórkowski A. (2004). *The IPI PAN Corpus: Preliminary version*. Instytut Podstaw Informatyki PAN, Warszawa.

Przepiórkowski A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski A. i Buczyński A. (2007). *Spejd: Shallow Parsing and Disambiguation Engine* [w:] Vetulani Z. (red.), *Proceedings of the 3rd Language & Technology Conference*, s. 340–344, Poznań.

Przepiórkowski A., Bańko M., Górski R.L. i Lewandowska-Tomaszczyk B., red. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Pustejovsky J. i Stubbs A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., Pekin/Cambridge/Farnham/Kolonia/Sewastopol/Tokio.

Radziszewski A. (2012). *Metody znakowania morfosyntaktycznego i automatycznej płytkiej analizy składniowej języka polskiego*. rozprawa doktorska, Politechnika Wrocławska.

Raghunathan K., Lee H., Rangarajan S., Chambers N., Surdeanu M., Jurafsky D. i Manning C. (2010). *A Multi-pass Sieve for Coreference Resolution* [w:] *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2010, s. 492–501, Stroudsburg. Association for Computational Linguistics.

Rahman A. i Ng V. (2009). *Supervised Models for Coreference Resolution* [w:] *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP 2009, s. 968–977, Stroudsburg. Association for Computational Linguistics.

Rahman A. i Ng V. (2012). *Translation-Based Projection for Multilingual Coreference Resolution* [w:] *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2012)*, s. 720–730, Montreal. Association for Computational Linguistics.

Ratinov L. i Roth D. (2012). *Learning-based Multi-sieve Co-reference Resolution with Knowledge* [w:] *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, s. 1234–1244, Stroudsburg. Association for Computational Linguistics.

Recasens M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation*. Rozprawa doktorska, Department of Linguistics, University of Barcelona, Barcelona.

Recasens M. i Hovy E. (2011). *BLANC: Implementing the Rand Index for Coreference Evaluation*. „Natural Language Engineering", 17(4), s. 485–510.

Recasens M. i Martí M.A. (2010). *AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan*. „Language Resources and Evaluation", 44(4), s. 315–345.

Recasens M., Martí A. i Taulé M. (2007). *Where Anaphora and Coreference Meet. Annotation in the Spanish CESS-ECE Corpus* [w:] *Proceedings of RANLP 2007*, s. 504–509, Borowec.

Recasens M., Hovy E. i Martí M.A. (2010). *A Typology of Near-Identity Relations for Coreference (NIDENT)* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M. i Tapias D. (red.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, s. 149–156, Valletta. European Language Resources Association.

Recasens M., Hovy E. i Martí M.A. (2011). *Identity, non-identity, and near-identity: Addressing the complexity of coreference*. „Lingua", 121(6), s. 1138–1152.

Rehbein I. i Ruppenhofer J. (2017). *Detecting annotation noise in automatically labelled data* [w:] *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, s. 1160–1170, Vancouver. Association for Computational Linguistics.

Riester A., Lorenz D. i Seemann N. (2010). *A Recursive Annotation Scheme for Referential Information Status* [w:] Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M. i Tapias D. (red.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, s. 717–722, Valletta. European Language Resources Association.

Roesiger I. i Teufel S. (2014). *Resolving Coreferent and Associative Noun Phrases in Scientific Text* [w:] *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, s. 45–55, Gothenburg, Sweden. Association for Computational Linguistics.

Russell B. (1905). *On Denoting*. „Mind", 14, s. 479–493.

Rysová M., Synková P., Mírovský J., Hajičová E., Nedoluzhko A., Ocelák R., Pergler J., Poláková L., Pavlíková V., Zdeňková J. i Zikánová Š. (2016). *Prague Discourse Treebank 2.0*.

Sasano R. i Kurohashi S. (2009). *A Probabilistic Model for Associative Anaphora Resolution* [w:] *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, s. 1455–1464, Singapur. Association for Computational Linguistics.

Schäfer U., Spurk C. i Steffen J. (2012). *A fully coreference-annotated corpus of scholarly papers from the ACL anthology* [w:] *Proceedings of COLING 2012: Posters*, s. 1059–1070, Mumbai, India. The COLING 2012 Organizing Committee.

Schulte im Walde S. (1998). *Resolving Bridging Descriptions in High-Dimensional Space*. Praca magisterska, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Searle J.R. (1975). *The Logical Status of Fictional Discourse*. „New Literary History", 6(2), s. 319–332.

Sgall P., Hajičová E. i Panevová J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Sidner C.L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Technical report, Massachusetts Institute of Technology, Cambridge.

Soon W.M., Ng H.T. i Lim C.Y. (1999). *Corpus-Based Learning for Noun Phrase Coreference Resolution* [w:] *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, s. 285–291, College Park. The Association for Computer Linguistics.

Soon W.M., Ng H.T. i Lim D.C.Y. (2001). *A Machine Learning Approach to Coreference Resolution of Noun Phrases*. „Computational Linguistics", 27(4), s. 521–544.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I. i Salakhutdinov R. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. „Journal of Machine Learning Research", 15, s. 1929–1958.

Stede M. i Neumann A. (2014). *Potsdam Commentary Corpus 2.0: Annotation for Discourse Research* [w:] Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, s. 925–929, Rejkiawik. European Language Resources Association.

Stenetorp P., Topić G., Pyysalo S., Ohta T., Kim J.D. i Tsujii J. (2011). *BioNLP Shared Task 2011: Supporting Resources* [w:] *Proceedings of BioNLP Shared Task 2011 Workshop*, s. 112–120, Portland, Oregon. Association for Computational Linguistics.

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S. i Tsujii J. (2012). *brat: a Web-based Tool for NLP-assisted Text Annotation* [w:] *Proceedings of the Demonstrations Session at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, s. 102–107, Avignon. Association for Computational Linguistics.

Stoyanov V., Gilbert N., Cardie C. i Riloff E. (2009). *Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art* [w:] *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, s. 656–664, Suntec, Singapur. Association for Computational Linguistics.

Stroińska M. (1992). *Styl bezosobowy a spójność referencjalna w dyskursie* [w:] Dobrzyńska T. (red.), *Typy tekstów: zbiór studiów*, s. 15–25. Instytut Badań Literackich Polskiej Akademii Nauk, Warszawa.

Stuckardt R. (2001). *Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm*. „Computational Linguistics", 27(4), s. 479–506.

Stührenberg M., Goecke D., Diewald N., Mehler A. i Cramer I. (2007). *Web-based Annotation of Anaphoric Relations and Lexical Chains* [w:] *Proceedings of the Linguistic Annotation Workshop*, s. 140–147. Association for Computational Linguistics.

Szkudlarek-Śmiechowska E. (2003). *Wskaźniki nawiązania we współczesnych tekstach polskich (na materiale współczesnej nowelistyki polskiej)*. Acta Universitatis Lodziensis: Folia linguistica. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Szwedek A. (1975). *Coreference and Sentence Stress in English and Polish*. „Poznań Studies in Contemporary Linguistics", 3, s. 209–213.

Topolińska Z. (1976). *Wyznaczoność (tj. charakterystyka referencyjna) grupy imiennej w tekście polskim*. „Polonica", 3(2), s. 33–72.

Topolińska Z. (1977). *„Referencja", „koreferencja", „anafora"*. „Slavica Slovaca", 12(3), s. 225–232.

Topolińska Z. (1984). *Składnia grupy imiennej* [w:] Grochowski M., Karolak S. i Topolińska Z. (red.), *Składnia*, Gramatyka współczesnego języka polskiego, s. 301–389. Państwowe Wydawnictwo Naukowe, Warszawa.

Trofimiec S. (2007). *Konstrukcje anaforyczne jako wskaźniki nawiązania w tekstach prasowych*. „Język Polski", LXXXVII(1), s. 24–28.

Uryupina O. (2007). *Knowledge Acquisition for Coreference Resolution*. Rozprawa doktorska, Saarland University.

Vater H. (2009). *Wstęp do lingwistyki tekstu. Struktura i rozumienie tekstów*, t. 2. Atut, Wrocław.

Versley Y. (2008). *Vagueness and referential ambiguity in a large-scale annotated corpus*. „Research on Language and Computation", 6, s. 333–353.

Versley Y., Ponzetto S.P., Poesio M., Eidelman V., Jern A., Smith J., Yang X. i Moschitti A. (2008). *BART: A Modular Toolkit for Coreference Resolution* [w:] *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, s. 9–12, Columbus. Association for Computational Linguistics.

Versley Y., Poesio M. i Ponzetto S.P. (2016). *Using Lexical and Encyclopedic Knowledge* [w:] Poesio M., Stuckardt R. i Versley Y. (red.), *Anaphora Resolution. Algorithms, Resources, and Applications*, s. 393–429. Springer.

Vetulani Z. (2014). *Polnet – Polish WordNet* [w:] Vetulani Z. i Mariani J. (red.), *Human Language Technology Challenges for Computer Science and Linguistics*, s. 408–416, Cham. Springer International Publishing.

Vieira R. i Teufel S. (1997). *Towards resolution of bridging descriptions* [w:] *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, s. 522–524, Madryt. Association for Computational Linguistics.

Vilain M., Burger J., Aberdeen J., Connolly D. i Hirschman L. (1995). *A Model-Theoretic Coreference Scoring Scheme* [w:] *Proceedings of the 6th Message Understanding Conference (MUC-6)*, s. 45–52, Columbia. Association for Computational Linguistics.

Wajszczuk J. (1978). *Syntaktyczny stosunek nawiązania (na materiale współczesnego języka rosyjskiego)*. Rozprawa doktorska, Uniwersytet Warszawski, Warszawa.

Waszczuk J., Głowińska K., Savary A., Przepiórkowski A. i Lenart M. (2013). *Annotation tools for syntax and named entities in the National Corpus of Polish*. „International Journal of Data Mining, Modelling and Management", 5(2), s. 103–122.

Webber B., Prasad R., Lee A. i Joshi A. (2016). *A Discourse-Annotated Corpus of Conjoined VPs* [w:] *Proceedings of the 10th Linguistic Annotation Workshop (LAW-X 2016)*, s. 22–31. Association for Computational Linguistics.

Wierzbicka A. (2010). *Semantyka: jednostki elementarne i uniwersalne*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.

Winkler W. (1999). *The State of Record Linkage and Current Research Problems*. Technical report. Statistical Research Report Series, No. RR1999/04. U.S. Bureau of the Census, Washington, D.C.

Woliński M. (2006). *Morfeusz – a practical tool for the morphological analysis of Polish* [w:] Kłopotek M.A., Wierzchoń S.T. i Trojanowski K. (red.), *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining 2006 Conference*, s. 511–520, Ustroń. Springer.

Woliński M. (2014). *Morfeusz Reloaded* [w:] Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. i Piperidis S. (red.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, s. 1106–1111, Rejkiawik. European Language Resources Association.

Xue N., Ng H.T., Pradhan S., Prasad R., Bryant C. i Rutherford A. (2015). *The CoNLL-2015 Shared Task on Shallow Discourse Parsing* [w:] Xue N., Ng H.T., Pradhan S., Prasad R., Bryant C. i Rutherford A. (red.), *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL 2015): Shared Task*, s. 1–16, Pekin. ACL.

Xue N., Ng H.T., Pradhan S., Rutherford A., Webber B.L., Wang C. i Wang H. (2016). *CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing* [w:] Xue N., Ng H.T., Pradhan S., Rutherford A., Webber B.L., Wang C. i Wang H. (red.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL 2016): Shared Task*, s. 1–19, Berlin. ACL.

Zeyrek D., Mendes A., Grishina Y., Kurfalı M., Gibbon S. i Ogrodniczuk M. (2019). *TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style*. „Language Resources and Evaluation". (w druku).

Zhang R., Nogueira dos Santos C., Yasunaga M., Xiang B. i Radev D. (2018). *Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering* [w:] *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, s. 102–107, Melbourne. Association for Computational Linguistics.

Zhou G. i Su J. (2004). *A High-Performance Coreference Resolution System using a Constraint-based Multi-Agent Strategy* [w:] *Proceedings of COLING 2004*, s. 522–528, Genewa. COLING.

Zikánová S., Hajičová E., Hladká B., Jínová P., Mírovský J., Nedoluzhko A., Poláková L., Rysová K., Rysová M. i Václ J. (2015). *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Instytut Lingwistyki Formalnej i Stosowanej, Uniwersytet Karola, Praga.

Skorowidz

agregacja 62
anafora 20
    leniwa 64
    typu E 62
    związana 62
anotacja
    równoległa 69
    szeregowa 72
asocjacja 20, 58
    strukturalna 62
aspekt 58, 64
    dysymilacja 65
    niepewność 65
    opinia 65

centrum semantyczne 77

dekodowanie koreferencji 20
diafora 20
dysymilacja 65

egzofora 56

homofora 56

identyczność
    odniesienia 64
    sensu 64
interreferencja 20

katafora 20
kategorialność 64
klaster koreferencyjny 19
kompozycja 62
koreferencja 19, 58

łańcuch koreferencyjny 19

metareferencja 63
metryka ewaluacyjna
    B^3 51
    BLANC 53
    CEAF 52
    CoNLL/MELA 49
    MUC 50

polisemia 64
porównanie 63

quasi-anafora 61

referencja 19
    bezpośrednia 20, 58
    pośrednia 20, 58
relacja
    definicyjna 63
    kategorialności 64
    kontrastu 64
    negatywna 59, 63
    polisemii 64
    predykatywna 63
    wspierająca 58, 63
    wykluczająca 59, 63
rozdzielony poprzednik 62
rozpodobnienie 65

singleton 19

wyrażenie dominujące 77
wzmianka 19

Skorowidz terminów angielskich

aggregation 62
anaphora 20
    associative anaphora 20
    bound anaphora 62
    E-type 62
    lazy anaphora 64
annotation
    parallel 69
    serial 72
association 58

bridging 20

cataphora 20
clustering algorithm
    entity-based 44
    mention-pair 41
comparison 63
composition 62
contrast 64
coreference 19, 58
    chain 19
    cluster 19
    resolution 20

diaphora 20
discourse world 19

evaluation metric
    B^3 51
    BLANC 53
    CEAF 52
    CoNLL/MELA 49
    MUC 50
excluding relation 59, 63
exophora 56

facet 58

homophora 56

identity-of-reference 64
identity-of-sense 64
interreference 20

mention 19
metareference 63

near-identity 29

polysemy 64
predicative anaphora 63

quasi-anaphora 61

reference 19
    direct 20, 58
    indirect 20, 58

scorer 41
singleton 19
split antecedent 62
structural association 62
supporting relation 58, 63

Wykaz powstałych narzędzi i zasobów

Lista narzędzi i zasobów wynikowych powstałych w ramach opisywanych prac:

- korpus zależności referencyjnych (ang. *Polish Coreference Corpus*): `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`
- BARTEK, narzędzie statystyczne do wykrywania koreferencji: `http://zil.ipipan.waw.pl/Bartek`
- rozszerzona wersja narzędzia BRAT, używanego do wizualizacji relacji referencyjnych: `http://zil.ipipan.waw.pl/brat4ref`
- DISCANN, aplikacja do anotacji metatekstowej: `http://zil.ipipan.waw.pl/Discann`
- DISTSYS, narzędzie do zarządzania przydziałem plików anotatorom: `http://zil.ipipan.waw.pl/DistSys`
- MENTIONDETECTOR, narzędzie regułowe do wykrywania wzmianek: `http://zil.ipipan.waw.pl/Discann`
- MMAX4CORE, wersja narzędzia MMAX zmodyfikowana na potrzeby zadania anotacji zależności referencyjnych: `http://zil.ipipan.waw.pl/MMAX4CORE`
- MMAX4REF, wersja zmodyfikowana na potrzeby anotacji relacji pośrednich: `http://zil.ipipan.waw.pl/MMAX4REF`
- RULER, narzędzie regułowe do wykrywania koreferencji: `http://zil.ipipan.waw.pl/Ruler`
- SCOREFERENCE, narzędzie do ewaluacji zadania wykrywania wzmianek i dekodowania koreferencji: `http://zil.ipipan.waw.pl/Scoreference`

Interfejsy dostępowe korpusu zależności referencyjnych i zaimplementowanych narzędzi:

- przeglądarka zawartości korpusu:
  `http://cothec.nlp.ipipan.waw.pl/`
- wyszukiwarka w korpusie zależności referencyjnych:
  `http://pcc.nlp.ipipan.waw.pl/`
- wersje demonstracyjne narzędzi do wykrywania wzmianek i dekodowania koreferencji zintegrowane z Multiserwisem:
  `http://multiservice.nlp.ipipan.waw.pl/pl/`