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.