Autoreferat

1. Imię i nazwisko

Małgorzata Marciniak

2. Posiadane dyplomy i stopnie naukowe — z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej

1987: Dyplom studiów wyższych na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego na kierunku Informatyka o specjalności Oprogramowanie i Metody Informatyczne;

2002: Stopień doktora uzyskany w Instytucie Podstaw Informatyki Polskiej Akademii Nauk, tytuł rozprawy: *Algorytmy implementacyjne syntaktycznych reguł koreferencji zaimków dla języka polskiego w terminach HPSG*

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych

1985–1986: Instytut Fizyki PAN (zatrudniona na 1/2 etatu programisty)
od 1986: Instytut Podstaw Informatyki PAN

4. Podstawowe osiągnięcie

Moim podstawowym osiągnięciem habilitacyjnym jest praca pod tytułem *Korpusy dziedzinowe jako źródło informacji*, w której skład wchodzi książka:

— Marciniak M. *Domain corpora as a source of information*. Monograph Series, volume 4, Institute of Computer Sciences PAS

wraz z następującymi artykułami:
— Mykowiecka, A., Marciniak, M., i Kupść, A. *Rule-based Information Extraction from Patients' Clinical Data*. Journal of Biomedical Informatics, 42, 923—936, (2009)
— Marciniak, M. i Mykowiecka, A. *Construction of a Medical Corpus Based on Information Extraction Results*. Control & Cybernetics, 40(2), 337—360, (2011)
— Marciniak, M. and Mykowiecka, A. *Terminology Extraction from Medical Texts in Polish*. Journal of Biomedical Semantics, 5. (2014)
— Marciniak, M. and Mykowiecka, A. *Nested Term Recognition Driven by Word Connection Strength*. Terminology, 21(2), 180–204, (2015)

Tematem mojego podstawowego osiągnięcia habilitacyjnego są metody przetwarzania tekstów dziedzinowych w celu pozyskania informacji z takich danych. Książka *Domain corpora as a source of*

information powstała na bazie artykułów opublikowanych w materiałach konferencyjnych, w czasopismach oraz jako rozdziały książek. Informacje w nich zawarte zostały ułożone w spójną całość oraz uzupełnione o wiadomości, które nie zostały zmieszczone w artykułach. W książce przedstawiam metody analizy językowej tekstów dziedzinowych, tworzenia anotowanego korpusu tekstów oraz wykorzystania opracowanych danych do złożonych zadań związanych z pozyskiwaniem informacji ze zgromadzonych tekstów. Szczegółowo omawiam wykorzystanie korpusów tekstów dziedzinowych do regułowej ekstrakcji informacji określonych przez użytkownika oraz metody rozpoznawania terminologii specyficznej dla zebranych tekstów.

W książce skupiłam się na przetwarzaniu tekstów medycznych, a konkretnie wypisów szpitalnych. Wybór ten został podyktowany interesującymi zagadnieniami pojawiającymi się przy ich przetwarzaniu. Zwykle nie są to teksty starannie edytowane, gdyż nie są przygotowane do publikacji, a ich głównym celem jest udokumentowanie terapii prowadzonej w szpitalu. Wypisy szpitalne zawierają sporo błędów typograficznych, pomimo że zwykle są pisane z włączoną w edytorze opcją korekty pisowni. Słownictwo kliniczne różni się zdecydowanie od zawartego w ogólnych słownikach języka polskiego, np. w danych Słownika Gramatycznego Języka Polskiego (`http://sgjp.pl/leksemy/`) wykorzystywanych przez analizator *Morfeusz* (Woliński, 2006). Przykładem takiego słownictwa są nazwy substancji leczniczych i leków. Są one pisane zgodnie z zasadami międzynarodowymi lub polskimi, przy czym te pisane według norm międzynarodowych mogą odmieniać się zgodnie z normami polskimi (*Ampicillinę* vs. *Ampicylinę*). Wypisy szpitalne zawierają niektóre informacje w języku łacińskim, czasem są to całe frazy, a czasem jedynie wtrącone słowo. Ponadto teksty te zawierają wiele akronimów i skrótów, których interpretacja często odbiega od przyjętej w języku potocznym, np. *por.* w języku potocznym oznacza *porównaj*, podczas gdy w dokumentacji medycznej jest to zwykle skrót wyrazu *poradnia*.

W tekstach medycznych część informacji jest reprezentowana w formie liczbowej, np. wyniki badań, daty, przedziały czasowe lub dawki leków. W korpusach języka ogólnego (np. w Narodowym Korpusie Języka Polskiego, Przepiórkowski i in. (2012)) informacje te nie są poddawane analizie i jest im przypisywany tag *ign* — ignoruj. W przypadku omawianych przeze mnie tekstów niosą one ważne informacje, nie mogą więc być pomijane. W książce przedstawiam zagadnienia związane z tokenizacją czyli wyróżnieniem podstawowych jednostek tekstu. Narzędzia wykorzystywane do przetwarzania tekstów dzielą go według różnych zasad, które wymagają ujednolicenia. Omówione w książce problemy spójnej tokenizacji oraz tagowania[1] tekstów medycznych zostały wcześniej przedstawione w artykule konferencyjnym (Marciniak i Mykowiecka, 2011b). Automatycznie otrzymane wyniki tagera języka polskiego (trenowanego na korpusie tekstów ogólnych) zostały poprawione, za pomocą zestawu reguł. Spójna tokenizacja i poprawione interpretacje morfologiczne zostały wykorzystane przy opracowaniu korpusu tekstów medycznych, który został opisany w artykule *Construction of a Medical Corpus Based on Information Extraction Results* (Marciniak i Mykowiecka, 2011a) wchodzącym w skład głównego osiągnięcia habilitacyjnego. W omawianej pracy Agnieszka Mykowiecka zajmowała się opracowaniem struktury korpusu w XML, oraz ujednoliceniem tokenizacji, ja odpowiadałam za przygotowanie reguł poprawy tagowania i za opracowanie poziomu anatacji semantycznej.

W książce przedstawiam problem rozpoznawania tokenów złożonych takich jak: liczby dziesiętne, daty, złożone jednostki miar. Opisuję gramatykę służącą do ich rozpoznawania, którą opracowałam przy użyciu narzędzia *Spejd* (Przepiórkowski, 2008). W następującym fragmencie wypisu szpitalnego '*HbA1C:10.6% (norma 4,2 -5,7)*' wyróżniam:
— nazwę badania czyli '*HbA1C*';
— dwukropek stanowiący separator;

[1] Tagowanie jest to zadanie polegające na przypisaniu poszczególnym tokenom reprezentującym słowa: formy podstawowej, części mowy oraz wartości adekwatnych cech gramatycznych, np. przypadka, liczby, rodzaju dla rzeczownika.

— wynik badania '*10.6*', którym jest liczba zapisana zgodnie z angielskimi standardami — kropka wskazuje na część dziesiętną liczby (nie jest ani końcem zdania ani nie kończy skrótu);
— znak procentu;
— określenie norm, gdzie znak '-' nie oznacza wartości ujemnej; pomimo braku spacji powinien on zostać zinterpretowany jako znak oznaczający zakres, przy czym minimalna i maksymalna prawidłowa wartość wyrażona jest w postaci liczb dziesiętnych podanych zgodnie z polskimi zasadami.

Krótki rozdział książki poświęcam problemom anonimizacji danych szpitalnych. Celem tego rozdziału jest zebranie zasad opracowanych w USA i UE służących do ochrony wrażliwych danych zawartych w dokumentacji medycznej. Wiedza ta nie jest dostatecznie rozpowszechniona w środowisku naukowym w Polsce. W jednym z rozdziałów omówiłam przykład programu służącego do odwracalnej anonimizacji dokumentów opracowany i użyty do zebrania danych z jednego ze szpitali. Opis programu i napotkanych problemów został opublikowany wcześniej w czasopiśmie *Journal of Medical Informatics & Technologies* (Marciniak i in., 2010). Idea programu została opracowana przeze mnie na podstawie rozmów z lekarzami oraz na podstawie przykładowych wypisów z fikcyjnymi wrażliwymi danymi. Został on zaimplementowany przez Piotra Rychlika, natomiast Agnieszka Mykowiecka uczestniczyła w finalnej ocenie działania programu.

W książce omawiam dwa zadania związane z ekstrakcją informacji[2] z korpusu tekstów dziedzinowych:

— ekstrakcję wybranych informacji z nieustrukturalizowanych tekstów w języku naturalnym;
— ekstrakcję terminologii dziedzinowej na podstawie korpusu tekstów.

Pierwszym z wyżej wymienionych zagadnień przedstawionym w książce jest ekstrakcja wybranych przez eksperta informacji z wypisów szpitalnych pacjentów cukrzycowych. Do realizacji tego zadania wybrany został system SProUT (Drożdżyński i in., 2004) dostosowany do przetwarzania języka polskiego (Piskorski i in., 2004) między innymi przez dołączenie analizatora morfologicznego *Morfeusz*. SProUT umożliwia zdefiniowanie własnego słownika dziedzinowego, gdzie oprócz typowych informacji morfologicznych możliwe jest umieszczenie informacji semantycznych powiązanych z reprezentowanym hasłem.

SProUT jest to system wykorzystujący gramatyki powierzchniowe do rozpoznawania poszukiwanych fraz. Do reprezentowania informacji wykorzystywane są utypowione struktury atrybutów (TFS), gdzie typy mogą stanowić wielohierarchię. SProUT umożliwia wykorzystanie unifikacji w regułach gramatyki powierzchniowej. Ułatwia to zapis uzgodnień pomiędzy poszukiwanymi elementami oraz sposób przekazywania finalnych wyników. Wyniki otrzymane przy pomocy gramatyki ekstrakcyjnej wymagają zwykle przetworzenia poza systemem SProUT w celu np. wybrania atrybutów o wartościach niosących informację.

Pierwszym zrealizowanym systemem ekstrakcji informacji dla języka polskiego był opracowany przez nas system do analizy zawartości raportów mammograficznych (Mykowiecka i in., 2009). Podobne eksperymenty na danych medycznych dla języka angielskiego są opisane w (Hahn i in., 2002) lub (Chapman i in., 2007). Opracowanie systemu ekstrakcji informacji w języku polskim wymagało uwzględnienia specyfiki języka polskiego, takiej jak na przykład swobodny szyk słów. Niezbędne było też uwzględnienie specyfiki dziedziny i zakresu ekstrahowanych informacji. Prace nad systemem ekstrakcji informacji z raportów mammograficznych udokumentowaliśmy najpierw w postaci serii artykułów konferencyjnych, w których omówiliśmy różne zagadnienia lingwistyczne wymagające rozwiązania w trakcie procesu ekstrakcji informacji. Do ekstrakcji informacji z raportów mammograficznych wykorzystany został opis dziedziny opracowany przez ekspertów (Podsiadły-Marczykowska

[2] Pojęcie ekstrakcji informacji (ang. Information Extraction) funkcjonuje w języku polskim od kilkunastu lat. Pomimo kontrowersjiu jakie on wywołuje, używając go kieruję się zasadami opisanymi w książce Mariana Mazura „Terminologia techniczna", które głoszą, że nie należy zmieniać terminów, które się rozpowszechniły oraz że nazwy techniczne powinny być zgodne co do źródłosłowów z nazwami mającymi rozpowszechnienie międzynarodowe.

i Guzik, 2004), który został sformalizowany w postaci ontologii OWL (Mykowiecka i in., 2007) oraz przełożony na hierarchię struktur TFS wykorzystywaną w systemie SProUT (Mykowiecka i Marciniak, 2009). Wprowadzenie do problemu ekstrakcji informacji przy wykorzystaniu systemu SProUT z danych medycznych oraz wstępne idee ekstrakcji prostych informacji a następnie grupowanie ich w bloki spójnych informacji opisujących: rodzaj utkania, rozpoznane zmiany oraz rekomendacje są opisane w artykule (Marciniak i in., 2005). W artykule (Mykowiecka i in., 2005b) opisujemy finalny stan gramatyki ekstrakcyjnej z uwzględnieniem algorytmu wyodrębniania poszczególnych bloków informacji zawierających opisy zmian. Opracowana gramatyka składa się z wielu reguł rozpoznających cząstkowe informacje, czyli na przykład osobno rozpoznajemy atrybuty opisujące zmiany takie jak: kształt, rozmiar oraz rodzaj granic (cecha mówiąca czy są one wyraźne czy zatarte). Cząstkowe informacje są następnie przydzielane do opisów poszczególnych zmian na podstawie atrybutów, które są dla zmiany unikalne np. informacja gdzie się ona znajduje. W artykule (Mykowiecka i in., 2005a) omówione zostały zagadnienia związane z odniesieniami anaforycznymi w raportach mammograficznych. Zostały w nim zidentyfikowane problemy i zaproponowane rozwiązania poniższych zagadnień:

— ustalenie odniesień anaforycznych dotyczących lokalizacji ('*w tej samej piersi*') oraz zmian ('*podobna zmiana*');
— ustalenie informacji w przypadku koordynacji lokalizacji ('*w kwadrantach górnych obu piersi*');
— uwzględnienie odniesień do poprzednich raportów ('*Opisywana w badaniu poprzednim torbiel w sutku prawym o wymiarach...*').

Zaproponowane rozwiązania powyższych zagadnień polegają na identyfikacji zwrotów anaforycznych przez reguły gramatyczne, a następnie, na etapie przetwarzania wyników, na kopiowaniu odpowiednich danych (lokalizacja, opis zmiany) z poprzednio rozpoznanych bloków informacji. W przypadku odniesienia do wcześniejszych raportów odpowiednie informacje są stosownie do potrzeb pomijane lub uwzględniane w finalnym raporcie. Omawiany system został oceniony na podstawie 705 raportów, które nie były analizowane na etapie tworzenia gramatyki. Rozpoznawanie wartości atrybutów dało wynik: 98% precyzji i 97% pełności. Natomiast granice bloków były rozpoznawane z precyzją rzędu 89% i pełnością 87%.

Z innego typu problemami zetknęłam się opracowując opisany w książce system ekstrakcji informacji z wypisów pacjentów diabetologicznych. Jego celem jest ustalenie:

— przyczyny przyjęcia do szpitala;
— podstawowych danych pacjenta takich jak: płeć, wiek, waga, wzrost;
— informacji dotyczących cukrzycy, np. typ, kiedy została zdiagnozowana, wyniki istotnych badań;
— powikłań cukrzycowych;
— zaleconej diety i sposobu leczenia cukrzycy;
— dodatkowych informacji istotnych dla przebiegu leczenia takich jak np. czy pacjent zachowuje dietę i prowadzi samokontrolę poziomów cukru.

Powyższe, interesujące dla eksperta informacje, stanowią niewielki procent tekstu i muszą być wybrane spośród wielu nieistotnych. W systemie ekstrakcji informacji cukrzycowych, część z nich jest rozpoznawana, tak jak w poprzednio omówionym systemie ekstrakcji z raportów mammograficznych, bez uwzględnienia kontekstu. Przykładowo wystąpienie słowa *otyły* w tekście wskazuje niemal na pewno na tuszę pacjenta, którego dotyczy raport. Istnieje jednak duża grupa informacji, które mogą być rozpoznane jedynie w kontekście specyficznych słów. Ponieważ naszym celem jest ekstrakcja informacji o cukrzycy, więc we frazie '*Wieloletnia, źle kontrolowana, z retinopatią, cukrzyca typu 2*' słowo '*wieloletnia*' powinno być rozpoznane jako cecha cukrzycy. Natomiast we frazie '*wieloletnia choroba niedokrwienna serca*' to samo słowo wskazuje na cechę innej choroby i należy je pominąć jako nieistotne. Innym problemem, który wymagał rozwiązania było rozpoznanie informacji zanegowanej. We frazie '*nie stwierdzono późnych zmian cukrzycowych w postaci mikroangiopatii*' należy poza wystąpieniem słowa '*mikronangiopatia*' opisującym powikłanie cukrzycowe rozpoznać, że występuje ono we frazie z negacją, jest to więc informacja o braku tego powikłania. Taką in-

formację należy również wyekstrahować z danych, gdyż jest istotna dla ekspertów — wskazuje na zaawansowanie choroby cukrzycowej. Rozpoznawanie zanegowanych informacji wymaga ustalenia współwystąpienia frazy wskazującej na negację (trzeba tu uwzględnić wiele jej wariantów) oraz interesującej eksperta informacji której dotyczy negacja. Ewaluacja systemu ekstrakcji informacji z wypisów pacjentów cukrzycowych wykazała średnią precyzję 99% i pełność 96%, szczegółowa ewaluacja wraz ze wskazanie przyczyn błędów jest przedstawiona zarówno w książce jak i w artykule *Rule-based Information Extraction from Patients' Clinical Data*.

Podsumowanie prac związanych z tworzeniem regułowych systemów ekstrakcji informacji dla języka polskiego znalazło się w artykule uwzględnionym w głównym osiągnięciu habilitacyjnym *Rule-based Information Extraction from Patients' Clinical Data* opublikowanym w czasopiśmie *Journal of Biomedical Informatics*. Artykuł ten został napisany we współpracy z Agnieszką Mykowiecką i Anną Kupść, które wraz ze mną są współautorkami systemu do ekstrakcji informacji z raportów mammograficznych. System ekstrakcji z danych diabetologicznych jest mojego autorstwa. Agnieszka Mykowiecka odpowiadała za ewaluację systemu ekstrakcji z danych diabetologicznych, natomiast analiza przyczyn błędów była wykonana przeze mnie.

Wyżej omówiony system ekstrakcji informacji z danych diabetologicznych został wykorzystany do opracowania relacyjnej bazy danych, która może posłużyć: do statystycznej analizy danych oraz do wyszukiwania korelacji i interesujących przypadków. Zadanie to zostało przedstawione w artykule konferencyjnym (Marciniak i in., 2008a).

Omawiany system wykorzystałam do opracowania metody automatycznej anotacji semantycznej danych tekstowych. Polega ona na uproszczeniu gramatyki ekstrakcyjnej, która nie bierze pod uwagę kontekstów informacji. Uproszczona gramatyka ekstrahuje wiele informacji nadmiarowych. Porównanie wyników obu systemów (pełnego i uproszczonego) pozwala na precyzyjne ustalenie, które frazy są istotne oraz które fragmenty tych fraz reprezentują poszczególne informacje. Metoda ta została opisana w już wymienianym artykule (Marciniak i Mykowiecka, 2011a) i wykorzystana do opracowania poziomu semantycznego korpusu zawierającego wypisy szpitalne.

Zaanotowany semantycznie korpus posłużył do opracowania systemu ekstrakcji informacji opartego na maszynowym uczeniu, które to eksperymenty zostały omówione w artykułach konferencyjnych: (Mykowiecka i Marciniak, 2011a) oraz (Mykowiecka i Marciniak, 2011b). Mój udział polegał na automatycznym przygotowaniu zaanotowanych danych treningowych, natomiast eksperymenty polegające na automatycznym uczeniu się przeprowadziła moja współautorka Agnieszka Mykowiecka.

Ostatnim omówionym w książce zagadnieniem jest automatyczna ekstrakcja terminologii z korpusów dziedzinowych. Jest to zadanie polegające na identyfikacji fraz, zwanych terminami, typowych dla dziedziny, której dotyczą zgromadzone teksty. Pojęcie terminologii dziedzinowej jest powszechnie zrozumiałe gdyż opracowywane są słowniki terminów np.: ekonomicznych (Gęsicki i Gęsicki, 1996), literackich (Głowiński i in., 2010), sztuk pięknych (Kozakiewicz, 2014). Słowniki te zawierają hasła będące terminami dziedzinowymi wraz z opisem ich znaczenia. Terminem w słownikach języka polskiego określa się: «wyraz albo połączenie wyrazowe o specjalnym, konwencjonalnie ustalonym znaczeniu naukowym lub technicznym» (Doroszewski, 1969). Taka definicja nie wskazuje jednak drogi postępowania w przypadku chęci opracowania programu do automatycznej realizacji tego zadania ani metody oceny uzyskanych wyników. Na potrzeby automatycznej ekstrakcji terminologii dziedzinowej przyjęliśmy następującą definicję:

> Przez termin dziedzinowy rozumiemy frazę, która w tekstach dziedzinowych występuje dostatecznie często by przypuszczać, że opisuje pojęcie istotne dla dziedziny i równocześnie częstość występowanie tej frazy w tekstach spoza dziedziny jest niższa.

Większość programów do automatycznej ekstrakcji terminologii ogranicza się do rozpoznawania fraz rzeczownikowych, czyli terminami są formy nominalne '*podanie leków*' a frazy czasownikowe '*podano leki*' czy '*podać leki*' nie są uwzględniane. Wprawdzie niektóre podejścia (np. Savova i in.,

2003) biorą pod uwagę frazy czasownikowe jednak na potrzeby automatycznej ekstrakcji terminologii ograniczamy pojęcie terminu dziedzinowego do frazy rzeczownikowej.

Zadanie ekstrakcji terminologii z tekstów dziedzinowych może być realizowane z myślą o wielu zastosowaniach, począwszy od tworzenia słowników i tezaurusów terminologii dziedzinowej, tworzenia słowników wielojęzycznych wykorzystywanych do automatycznego tłumaczenia, przez indeksowanie tekstów lub powiązaną z nią ekstrakcję informacji, aż po tworzenie ontologii dziedziny, której dotyczą zgromadzone teksty. W zależności od planowanego zastosowania konstrukcja ekstrahowanych fraz może ulegać zmianom, tak by wyniki były najlepiej dopasowane do planowanego celu. W zależności od potrzeb możemy uwzględnić lub nie frazy zawierające modyfikatory przyimkowe. Podejście takie zastosowaliśmy w przypadku ekstrakcji terminologii ekonomicznej, która jest opisana w artykule (Marciniak i Mykowiecka, 2013). Istotne było wówczas uwzględnienie fraz z modyfikatorami przyimkowymi (np. '*podatek dochodowy od osoby fizycznej*') ze względu na chęć porównania wyników z terminologią zawartą w słowniku terminologii ekonomicznej SEJFEK (Savary i in., 2012), do którego hasła zostały wybrane manualnie z uwzględnieniem takich fraz. W omawianej pracy, zadanie ekstrakcji terminologii zostało wykonane na korpusie tekstów ekonomicznych zebranych z Wikipedii (Kobyliński, 2012). Natomiast w artykule (Mykowiecka i Marciniak, 2012) modyfikacje przyimkowe nie były brane pod uwagę, ze względu na cel ekstrakcji, którym było grupowanie fraz. W takim przypadku podział powyższej frazy na dwie '*podatek dochodowy*' oraz '*osoba fizyczna*' jest rozwiązaniem korzystniejszym z punktu widzenia dalszego wykorzystania wyników.

Do automatycznego rozpoznawania terminologii w tekstach polskich wybrana została metoda oparta na współczynniku C (ang. C-value) zaproponowana w artykule (Frantzi i in., 2000). Interesującą cechą tej metody jest zwrócenie uwagi na fakt istnienia terminów, które w korpusach występują wewnątrz szerszych terminów. Terminy takie są nazywane zagnieżdżonymi (ang. nested terms). Przykładowo, w wypisach szpitalnych pojęcie '*pęcherzyka żółciowego*' występuje niemal wyłącznie we frazach określających jego stan: '*pęcherzyk żółciowy prawidłowy*', '*powiększony pęcherzyk żółciowy*'. W artykule (Marciniak i Mykowiecka, 2013) dotyczącym ekstrakcji terminologii z korpusu tekstów ekonomicznych, prezentujemy wyniki pokazujące, że około 10 % fraz zdefiniowanych w słowniku SEJFEK wystąpiło w analizowanych tekstach jedynie w postaci fraz zagnieżdżonych co wskazuje, że uwzględnianie takich frazy pozytywnie wpływa na jakość tworzonych zasobów.

Metoda oparta na współczynniku C promuje frazy, które wystąpiły często i w wielu kontekstach, przy czym jako konteksty brane są pod uwagę inne terminy. Autorzy metody zwracają uwagę, że zagnieżdżona fraza, która występuje zawsze w jednym kontekście nie stanowi zapewne terminu dziedzinowego, gdyż najprawdopodobniej jest to tylko fragment jakiegoś dłuższego terminu. Frantzi i in. zdefiniowali współczynnik C, według którego tworzona jest lista terminów, której górna część stanowi zbiór terminów charakterystycznych dla analizowanych tekstów. Jest on uzależniony od częstości frazy, liczby kontekstów, w których występuje, oraz długości frazy, tak, by dłuższe frazy były promowane ze względu na ich rzadsze występowanie. Definicja współczynnika C jest podana we wzorze (1) gdzie: p oznacza rozważaną frazę; $l(p)=log_2$ *długość(p)* (jeśli chcemy uwzględnić frazy jednowyrazowe przyjmujemy dla nich wartość $l(p)$ np. 0.1); LP jest zbiorem fraz zawierających frazę p; natomiast $r(LP)$ jest liczbą różnych fraz w zbiorze LP.

$$C-value(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)), & dla\ r(LP) > 0, \\ l(p) * freq(p), & dla\ r(LP) = 0 \end{cases} \quad (1)$$

Wszystkie metody szeregowania fraz będących kandydatami na terminy dziedzinowe biorą pod uwagę częstość występowania owych fraz. Dla języka polskiego (podobnie jak dla innych języków fleksyjnych) zliczanie częstości fraz, zarówno tych co wystąpiły samodzielnie w tekście jak i tych zagnieżdżonych w innych frazach, nie jest czynnością polegającą na prostym porównywaniu napisów. Następującą frazę w narzędniku: '*zwężeniem ujścia moczowodu*' należy zidentyfikować jako wystąpienie frazy podstawowej: '*zwężenie ujścia moczowodu*'. W tej frazie wystąpiły trzy zagnieżdżone

frazy o następujących formach podstawowych: '*ujście moczowodu*', '*ujście*' oraz '*moczowód*'. Żadna z tych trzech form nie pasuje bezpośrednio ani do rozważanej frazy ani do jej formy podstawowej.

Wygodnym, choć niedoskonałym rozwiązaniem tego problemu jest wykorzystanie uproszczonej formy podstawowej, którą uzyskujemy lematyzując poszczególne elementy frazy, (Marciniak i Mykowiecka, 2013). Przykładowo, dla frazy '*zwężenie ujścia moczowodu*' jest to forma '*zwężenie ujście moczowód*', a jej trzy podfrazy mają następujące uproszczone formy podstawowe: '*ujście moczowód*', '*ujście*' oraz '*moczowód*'. Ich wystąpienie daje się zidentyfikować w uproszczonej formie podstawowej frazy nadrzędnej przy pomocy prostego porównywania napisów. Stosując to rozwiązanie należy pamiętać, że istnieją frazy, które mają różne formy podstawowe, a ich uproszczone formy podstawowe są takie same. Na przykład frazy, w których element główny jest modyfikowany frazą rzeczownikową różniącą się liczbą, np. '*zapalenie ucha*' oraz '*zapalenie uszu*' będą miały przypisaną identyczną uproszczoną formę podstawową '*zapalenie ucho*'. Problemy te wraz z metodą odtwarzania terminu w formie podstawowej wygodnej dla człowieka omawiam szczegółowo w książce.

W artykule *Terminology Extraction from Medical Texts in Polish* (Marciniak i Mykowiecka, 2014b) zwracamy uwagę na fakt braku określenia sposobu zliczania kontekstów fraz przez autorów metody. Problem ten nie pojawia się też w innych artykułach dotyczących tej metody. W artykule opisujemy i analizujemy trzy różne metody liczenia kontekstów wraz z porównaniem ich wpływu na otrzymywane wyniki. Celem różnego definiowania kontekstów była próba zmniejszenia wagi fraz, których konstrukcja gramatyczna jest poprawna, jednak są one frazami niepoprawnymi semantycznie. Są to zwykle frazy urwane (ang. truncated phrase) takie jak '*USG jamy*', które powstają z frazy '*USG jamy brzusznej*'. Niestety zmiana sposobu liczenia kontekstów nie doprowadziła do osiągnięcia pożądanego efektu.

Do eliminacji urwanych fraz zaproponowałam wykorzystanie informacji o sile wiązania słów występujących w korpusie tekstów dziedzinowych. Rozwiązanie to zakłada, że w kolejnych krokach dzielimy rozważaną frazę na najwyżej dwie frazy zagnieżdżone, a nie tak jak do tej pory ustalamy wszystkie możliwe poprawne gramatycznie frazy. Proces ten jest wykonywany rekurencyjnie aż do podziału frazy na pojedyncze wyrazy. Spośród kilku analizowanych metod liczenia siły wiązania słów wybrałam znormalizowaną informację wzajemną (Normalised Pointwise Mutual Information, NPMI, Bouma (2009)), która stosunkowo dobrze odzwierciedla siłę wiązania wyrazów zarówno dla częstych słów jak i dla tych o niskiej częstości. Definicja miary jest podana we wzorze (2). Jest ona zdefiniowana dla bigramu 'x y', gdzie 'x' i 'y' to lematy kolejnych tokenów. p(x,y) jest prawdopodobieństwem bigramu 'x y' w rozważanym korpusie a p(x), p(y) są prawdopodobieństwami bigramów 'x' i 'y'.

$$NPMI(x,y) = \left(ln \ \ \frac{p(x,y)}{p(x)p(y)} \right) \Big/ - ln \ \ p(x,y) \tag{2}$$

Wykorzystanie tej metody w korpusie wypisów szpitalnych doprowadziło do eliminacji większości urwanych fraz ze zbioru terminów, które znalazły się na początku listy uszeregowanej według współczynnika C, ewaluacja została przeprowadzona na 2000 terminów z początku listy. Metoda zaproponowana w konferencyjnym artykule (Marciniak i Mykowiecka, 2014a) została następnie opublikowana w artykule *Nested Term Recognition Driven by Word Connection Strength* (Marciniak i Mykowiecka, 2015). W tym ostatnim opisujemy wyniki testowania wpływu metody na pojawianie się niepoprawnych, urwanych terminów wydobytych z korpusu tekstów ekonomicznych. Metoda została też zastosowana dla języka angielskiego do korpusu GENIA (Kim i in., 2008). Eksperymenty te wykazały skuteczność metody do ograniczenia występowania fraz urwanych w ekstrahowanej terminologii.

W artykułach wchodzących w skład głównego osiągnięcia habilitacyjnego: (Marciniak i Mykowiecka, 2014b), i (Marciniak i Mykowiecka, 2015), Agnieszka Mykowiecka zaimplementowała metodę liczenia współczynnika C oraz gramatyki ekstrakcji terminów, ja zajmowałam się przygoto-

waniem danych oraz ewaluacją wyników. Algorytm wykorzystania NPMI do redukcji fraz urwanych jest mojego autorstwa.

5. Inne osiągnięcia

5.1. Formalny opis języka polskiego w terminach HPSG

W pierwszym okresie pracy naukowej zajmowałam się formalnym opisem języka polskiego w terminach HPSG (Head-driven Phrase Structure Grammar (Pollard i Sag, 1994)). HPSG jest generatywną teorią lingwistyczną opartą na ograniczeniach, które sprawiają, że formalizm dopuszcza jedynie konstrukcje zgodne z gramatyką. Badania te doprowadziły do napisania pracy doktorskiej pod tytułem *Algorytmy implementacyjne syntaktycznych reguł koreferencji zaimków dla języka polskiego w terminach HPSG*, którą obroniłam w 2002 roku w Instytucie Podstaw Informatyki PAN. W pracy omawiam klasyfikację zaimków polskich wraz z zasadami interpretacji zaimków osobowych oraz dzierżawczych w różnych typach fraz. Analiza materiału językowego doprowadziła do sformułowania następujących zasad teorii wiązania dla języka polskiego:

Zasada A. Zaimki anaforyczne zwrotne muszą być związane.

Zasada B. Zaimki nieanaforyczne muszą być wolne z wyjątkiem zaimków dzierżawczych w pierwszej i drugiej osobie oraz sytuacji, gdy zaimek dzierżawczy jest wiązany przez jawny podmiot frazy rzeczownikowej.

Zasady te zdefiniowałam w terminach HPSG, a następnie zaimplementowałam w testowej gramatyce. Zagadnienia opisane w pracy doktorskiej opublikowałam w dwóch artykułach w języku angielskim. Pierwszy z nich (Marciniak, 1999) stanowi rozdział książki *Slavic in Head-Driven Phrase Structure Grammar* opublikowanej przez CSLI i jest to wstępna analiza problemu, podczas gdy w artykule konferencyjnym (Marciniak, 2002) omówiłam finalne rozwiązania wraz z implementacją. Główne wyniki pracy doktorskiej są też przedstawione w rozdziale *Teoria wiązania* książki *Formalny opis języka polskiego* (Przepiórkowski i in., 2002). Mojego autorstwa jest w tej książce również rozdział dotyczący implementacji gramatyki ilustrującej poszczególne omówione zjawiska językowe.

Wśród zagadnień związanych z reprezentacją zjawisk języka polskiego w HPSG zajmowałam się również koordynacją (Kupść i in., 2000). Zajmowałam się też efektywniejszą implementacją gramatyki HPSG dla języka polskiego w systemie LKB (Copestake, 2002), co zostało omówione w artykule (Mykowiecka i Marciniak, 2008).

Prace dotyczące formalnego opisu języka polskiego zostały uzupełnione o opracowanie zbioru zdań służących do testowania parserów języka polskiego (Marciniak i in., 2000) oraz zasad tworzenia testowego korpusu wzorcowych rozbiorów HPSG zdań polskich (Marciniak i in., 2003).

5.2. Anotowany korpus dialogów telefonicznych

W latach 2006–2009 uczestniczyłam w realizacji europejskiego projektu LUNA (sopken LAnguage UNderstanding in multilinguAl communication systems, IST 033549). W ramach tego projektu został zebrany i opracowany na poziomie morfosyntaktycznym i semantycznym korpus dialogów dotyczących warszawskiej komunikacji miejskiej. Korpus posłużył do opracowania prototypu modułu rozumienia mowy, identyfikującego w wypowiedzi pojęcia z dziedziny komunikacji miejskiej. Zebrane dialogi zostały szczegółowo opracowane, poczynając od ręcznej transliteracji, poprzez analizę morfologiczną i syntaktyczną, a kończąc na kilkuetapowej analizie semantycznej. Zebrane dialogi tworzą korpus LUNA.PL. Drugi korpus LUNA-WOZ.PL został zebrany przy użycia metody "Wizard of Oz" polegającej na symulowaniu komputerowego systemu dialogowego. Oba korpusy są dostępne do prowadzenia dalszych badań nad językiem a ich szczegółowy opis znajduje się w książce *Anotowany korpus dialogów telefonicznych* pod moją redakcją.

Uczestniczyłam w pracach nad większością poziomów anotacji tekstów dialogów (transktypcja mowy i prace związane z modelem języka na potrzeby rozpoznawania mowy były realizowane przez zespół z ówczesnej PJWSTK, obecnie PJATK), przy czym główny mój wkład dotyczył poziomów analizy morfosyntaktycznej, mojego autorstwa jest rozdział książki omawiający wyodrębnianie prostych fraz w tekstach.

5.3. Narzędzia do tworzenia słowników wielowyrazowych nazw własnych

Jednym z wyników kierowanego przeze mnie projektu finansowanego przez MNiSW (567/6 PR UE/2008/7) jest opracowanie narzędzia do tworzenia słowników nazw wielowyrazowych. Zostało ono opracowane do opisu i generowania nazw własnych potrzebnych do identyfikowania nazw miejsc w Warszawie. Problem rozpoznawania takich nazw powstał w wyniku analizy dialogów projektu LUNA (Marciniak i in., 2008b). W ramach projektu opracowany został edytor *Toposław* (Marciniak i in., 2009) pierwotnie dedykowany nazwom topograficznym. Do opisu jednostek wielowyrazowych wykorzystuje on grafy zaimplementowane w *Multiflexie* (Savary, 2005), który z kolei wykorzystuje *Unitexa* (Paumier, 2002). Raz zdefiniowany graf może być wykorzystany do opisu wielu nazw. Dla osób tworzących słowniki, narzędzie ma ułatwienia polegające na wyszukiwaniu potencjalnie podobnych nazw, podpowiadaniu szkieletu grafu do dalszego rozbudowywania, oraz wskazywaniu ścieżek wykorzystywanych do generowania interesujących wariantów. Narzędzie jest opisane w artykule (Marciniak i in., 2011), a zaproponowane ułatwienia w (Woliński i in., 2009).

Tworzony przy pomocy *Toposława* słownik umożliwia opisywanie nazw wielowyrazowych, które mogą występować w wielu wariantach, na przykład '*ulica Bitwy Warszawskiej 1920 r.*' jest w praktyce skracana do '*ulica Bitwy Warszawskiej*' a potocznie mówi się o niej '*Bitwy Warszawy*'. Zaproponowane narzędzie umożliwia rozpoznawanie różnych form gramatycznych danej nazwy oraz różnych jej wariantów, umożliwiając ich powiązanie z formą bazową. Przyjęliśmy, że jest nią oficjalna nazwa zatwierdzone przez Urząd Miasta. Opracowany słownik zawiera około 8000 nazw ulic, budynków, dzielnic, pomników, itp., dla których można określić typ reprezentowanego obiektu (Marciniak i Rabiega-Wiśniewska, 2010).

Toposław jest również wykorzystywany do tworzenia słowników opisujących frazy wielowyrazowe, posłużył m.in do opracowania słownika SEJFEK — Słownika Elektronicznych Jednostek Frazeologicznych z EKonomii (Savary i in., 2012), obecnie jest wykorzystywany w projekcie NCN (DEC-2013/09/B/HS2/01222), którego celem jest opracowanie słownika frazeologizmów czasownikowych (Czerepowicka i in., 2014). Narzędzie jest rozwijane w ramach projektu CLARIN.PL.

Literatura

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. W: *Proceedings of the Biennial GSCL Conference 2009*, str. 31–40, Tübingen. Gesellschaft für Sprachtechnologie & Computerlinguistik.

Chapman, W., Chu, D., i Dowling, J. N. (2007). Context: An algorithm for identifying. contextual features from clinical text. W: *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing ACL Workshop*.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Czerepowicka, M., Kosek, I., i Przybyszewski, S. (2014). O projekcie elektronicznego słownika odmiany frazeologizmów czasownikowych. *POLONICA*, **34**, 115–123.

Doroszewski, W., red. (1958-1969). *Słownik języka polskiego*. Państwowe Wydawnictwo Naukowe, Warszawa.

Drożdżyński, W., Krieger, H.-U., Piskorski, J., Schäfer, U., i Xu, F. (2004). Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *German AI Journal KI-Zeitschrift*, **01/04**, 17–23.

Frantzi, K., Ananiadou, S., i Mima, H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, **3**, 115–130.

Gęsicki, Ł. i Gęsicki, M. (1996). *Słownik terminów ekonomiczno-prawnych*. Interfart, Łódź.

Głowiński, M., Kostkiewiczowa, T., i Okopień-Sławińska, A., red. (2010). *Słownik terminów literackich*. Ossolineum.

Hahn, U., Romacker, M., i Schultz, S. (2002). MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, str. 63–74.

Kim, J.-D., Ohtai, T., i Tsujii, J. (2008). Corpus Annotation for Mining Biomedical Events from Literature. *BMC Bioinformatics*, **9:10**.

Kobyliński, Ł. (2012). Mining Class Association Rules for Word Sense Disambiguation. W: P. Bouvry, M. A. Kłopotek, F. Leprevost, M. Marciniak, A. Mykowiecka, i H. Rybiński, red., *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, str. 307–318. Springer-Verlag, Berlin, Heidelberg.

Kozakiewicz, S., red. (2014). *Słownik terminologiczny sztuk pięknych*. Państwowe Wydawnictwo Naukowe, Warszawa.

Kupść, A., Marciniak, M., i Mykowiecka, A. (2000). Constituent coordination in Polish: An attempt at an HPSG account. W: P. Bański i A. Przepiórkowski, red., *Proceedings of the First Generative Linguistics in Poland Conference*, str. 104–115, Warsaw. Institute of Computer Science, Polish Academy of Sciences.

Marciniak, M. (1999). Toward a binding theory for Polish. W: R. D. Borsley i A. Przepiórkowski, red., *Slavic in Head-Driven Phrase Structure Grammar*, str. 125–147. CSLI Publications, Stanford, CA.

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

Marciniak, M. (2002). Anaphora binding in Polish. Theory and implementation. W: *Proceedings of DAARC2002*, Lisbon.

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

Marciniak, M. i Mykowiecka, A. (2011a). Construction of a Medical Corpus Based on Information Extraction Results. *Control & Cybernetics*, **40**(2), 337–360.

Marciniak, M. i Mykowiecka, A. (2011b). Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. W: *Proceedings of BioNLP 2011*, str. 92–100.

Marciniak, M. i Mykowiecka, A. (2013). Terminology Extraction from Domain Texts in Polish. volume 467 of *Studies in Computational Intelligence*, str. 171–185. Springer-Verlag, Cham, Heidelberg, New York, Dordrecht, London.

Marciniak, M. i Mykowiecka, A. (2014a). NPMI Driven Recognition of Nested Terms. W: *Proceedings of the 4th International Workshop on Computational Terminology*, str. 33–41. Association for Computational Linguistics and Dublin City University.

Marciniak, M. i Mykowiecka, A. (2014b). Terminology Extraction from Medical Texts in Polish. *Journal of Biomedical Semantics*, **5**.

Marciniak, M. i Mykowiecka, A. (2015). Nested Term Recognition Driven by Word Connection Strength. *Terminology*, **21**(2), 180–204.

Marciniak, M. i Rabiega-Wiśniewska, J. (2010). Elektroniczny słownik fleksyjny nazw Warszawy. *Prace Filologiczne*, **LVIII**, 271–282.

Marciniak, M., Mykowiecka, A., Kupść, A., i Węgiel, M. (2000). Klasyfikacja zjawisk syntaktycznych

na potrzeby testowego zbioru wyrażeń języka polskiego. IPI PAN Research Report 908, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Marciniak, M., Mykowiecka, A., Przepiórkowski, A., i Kupść, A. (2003). An HPSG-annotated test suite for Polish. W: A. Abeillé, red., *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, str. 129–146. Kluwer, Dordrecht.

Marciniak, M., Mykowiecka, A., Kupść, A., i Piskorski, J. (2005). Intelligent Content Extraction from Polish Medical Reports. W: L. Bolc, Z. Michalewicz, i T. Nishida, red., *International Workshop on Intelligent Media Technology for Communicative Intelligence*, volume 3490 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg.

Marciniak, M., Mykowiecka, A., i Waszczuk, J. (2008a). Automatyczne wypełnianie bazy danych pacjentów diabetologicznych na podstawie wypisów szpitalnych. W: *Proceedings of INFOBAZY 2008*.

Marciniak, M., Rabiega-Wiśniewska, J., i Mykowiecka, A. (2008b). Proper names in dialogs from the Warsaw Transportation Call Center. W: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, i K. Trojanowski, red., *Intelligent Information Systems*, Warsaw. Akademicka Oficyna Wydawnicza EXIT.

Marciniak, M., Rabiega-Wiśniewska, J., Savary, A., Woliński, M., i Heliasz, C. (2009). Constructing an electronic dictionary of Polish urban proper names. W: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, i K. Trojanowski, red., *Recent Advances in Intelligent Information Systems*, str. 233–246. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Marciniak, M., Mykowiecka, A., i Rychlik, P. (2010). Medical Text Data Anonymization. *Journal of Medical Informatics & Technologies*, **16**, 83–88.

Marciniak, M., Savary, A., Sikora, P., i Woliński, M. (2011). Toposław – a lexicographic framework for multi-word units. W: Z. Vetulani, red., *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6–8, 2009, Revised Selected Papers*, volume 6562 of *Lecture Notes in Artificial Intelligence*, str. 139–150. Springer-Verlag, Berlin.

Mazur, M. (1961). *Terminologia techniczna.* Wydawnictwa Naukowo-Techniczne, Warszawa.

Mykowiecka, A. i Marciniak, M. (2008). Phrase structure for an effective Polish HPSG grammar. W: G. Zybatow, L. Szucsich, U. Junghanns, i R. Meyer, red., *Formal Description of Slavic Languages*. Proceedings of the Fifth European Conference on Formal Description of Slavic Languages, Leipzig, 2003.

Mykowiecka, A. i Marciniak, M. (2009). Domain model for medical information extraction – the LightMedOnt ontology. W: *Aspects of Natural Language Processing*, volume 5730 of *Lecture Notes in Computer Science*. Springer-Verlag.

Mykowiecka, A. i Marciniak, M. (2011a). Automatic Semantic Labeling of Medical Texts with Feature Structures. W: I. Habernal i V. Matoušek, red., *Text, Speech and Dialogue: 14th International Conference*, volume 6836 of *Lecture Notes in Artificial Intelligence*, str. 49–56, Berlin, Heidelberg. Springer-Verlag.

Mykowiecka, A. i Marciniak, M. (2011b). Some Remarks on Automatic Semantic Annotation of a Medical Corpus. W: *Proceedings of Third Louhi Workshop on Health Documentation Text Mining and Information Analysis at AIME*.

Mykowiecka, A. i Marciniak, M. (2012). Combining Wordnet and Morphosyntactic Information in Terminology Clustering. W: *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai.

Mykowiecka, A., Marciniak, M., i Kupść, A. (2005a). Making Shallow Look deeper: Anaphora and comparisons in medical information extraction. *Archives of Control Sciences*, **15**(3), 371–380.

Mykowiecka, A., Kupść, A., i Marciniak, M. (2005b). Rule-based medical content extraction and classification. W: *Intelligent Information Processing and Web Mining Proceedings of the International IIS: IIPWM'05*, volume 31 of *Advances in Intelligent and Soft Computing*. Springer-Verlag,

Berlin, Heidelberg.

Mykowiecka, A., Marciniak, M., i Podsiadły-Marczynkowska, T. (2007). A "Data-driven" Ontology for an Information Extraction System from Mammography Reports. W: *Proceedings of 10th Intl. Protégé Conference.*

Mykowiecka, A., Marciniak, M., i Kupść, A. (2009). Rule-based Information Extraction from Patients' Clinical Data. *Journal of Biomedical Informatics*, **42**, 923–936.

Paumier, S. (2002). Manuel d'utilisation du logiciel Unitex. http://www-igm.univ-mlv.fr/unitex/manuelunitex.ps.

Piskorski, J., Homola, P., Marciniak, M., Mykowiecka, A., i Woliński, A. P. M. (2004). Information Extraction for Polish Using the SProUT Platform. W: *Intelligent Information Processing and Web Mining. Proceedings of the IIS: IIPWM"04*, volume 25 of *Advances in Intelligent and Soft Computing*, str. 225–236. Springer-Verlag, Berlin, Heidelberg.

Podsiadły-Marczykowska, T. i Guzik, A. (2004). Mammographic Ontology - Conceptual Model of the Domain. *The International Journal of Artificial Organs*, **127**.

Pollard, C. i Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar.* The University of Chicago Press, Chicago.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego.* Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski, A., Kupść, A., Marciniak, M., i Mykowiecka, A. (2002). *Formalny opis języka polskiego: Teoria i implementacja.* Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski, A., Bańko, M., Górski, R. L., i Lewandowska-Tomaszczyk, B., red. (2012). *Narodowy Korpus Języka Polskiego.* Wydawnictwo Naukowe PWN, Warszawa.

Savary, A. (2005). A formalism for the computational morphology of multi-word units. *Archives of Control Sciences*, **15**(3), 437–449.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., i Makowiecki, F. (2012). SEJFEK – a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. W: *Proceedings of Cognitive Aspects of the Lexicon (COGALEX-III), a Workshop at COLING 2012*, Mumbai, India.

Savova, G. K., Harris, M., Johnson, T., Pakhomov, S. V., i Chute, C. G. (2003). A Data-Driven Approach for Extracting "the Most Specific Term" for Ontology Development. *AMIA 2003 Annual Symposium Proceedings*, str. 579–583.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. W: M. A. Kłopotek, S. T. Wierzchoń, i K. Trojanowski, red., *Intelligent Information Processing and Web Mining*, volume 36 of *Advances in Soft Computing*, str. 503–512. Springer-Verlag, Berlin.

Woliński, M., Savary, A., Sikora, P., i Marciniak, M. (2009). Usability improvements in the lexicographic framework Toposław. W: Z. Vetulani, red., *Proceedings of the 4th Language & Technology Conference*, str. 321–325, Poznań, Poland.