https://doi.org/10.18778/7525-352-8.10

Automatyczne wydobywanie wiedzy o semantyce języka naturalnego z korpusów tekstu[1]

Maciej Piasecki
Politechnika Wrocławska

1. Geneza automatyzacji

W tytule pracy zawarte jest twierdzenie odwołujące się do założenia, że automatyczne wydobywanie wiedzy o semantyce języka naturalnego jest możliwe i przydatne. Pytanie, czy jest tak w rzeczywistości? Znaczenie wyrażeń językowych lub, co gorsza, wypowiedzi jest zjawiskiem złożonym i mocno powiązanym z szeroko rozumianym kontekstem ich wystąpienia. W przypadku wypowiedzi językowej zapisanej w postaci fragmentu tekstu w korpusie[2], dostępna wiedza o jej kontekście użycia jest bardzo uboga, np. często znamy autora, ale sytuacja, adresat, czas, miejsce itd. są zwykle nieokreślone lub określone bardzo nieprecyzyjnie. Rodzą się pytania, jak bardzo ogranicza to możliwość poznania znaczenia wyrażeń językowych poprzez analizę zawężoną do zapisów wypowiedzi językowych zawartych w korpusach tekstu, oraz w jakich zastosowaniach i obszarach tak pozyskana wiedza o znaczeniu może być pomocna?

Mając na uwadze powyższe uzasadnione wątpliwości, rozpocznijmy od krótkiej analizy, dlaczego warto podejmować wysiłki nad opracowywaniem metod automatycznego wydobywania wiedzy o znaczeniu z korpusów tekstu. Przez automatyczne wydobywanie wiedzy rozumieć będziemy proces pozyskiwania wiedzy, realizowany przez program komputerowy, którego działanie nie wymaga ingerencji człowieka.

[1] Praca naukowa finansowana ze środków na naukę w latach 2005–2012 przez Ministerstwo Nauki i Szkolnictwa Wyższego jako projekty badawcze nr 3 T11C 018 29 oraz N N516 068637. Chciałbym bardzo podziękować za cenne uwagi do wcześniejszych wersji niniejszej pracy: dr Elżbiecie Hajnicz, dr Agnieszce Piaseckiej, dr Ewie Rudnickiej, prof. dr hab. Piotrowi Stalmaszczykowi oraz dr Robertowi Trypuzowi.

[2] W niniejszej pracy będziemy używać pojęcia korpus w szerokim znaczeniu, jako obszerny zbiór tekstów reprezentujących przykłady użycia języka naturalnego (por. McEnery, 2003; Lewandowska-Tomaszczyk, 2005).

Jedynym źródłem danych, jakie dostarczamy do programu, są korpusy tekstu, co oznacza, że program z założenia nie wykorzystuje gotowych już źródeł wiedzy opisujących semantykę języka naturalnego, takich jak np. leksykony semantyczne czy też sformalizowane opisy znaczenia konstrukcji językowych. W przypadku programu, który by korzystał z istniejącego już źródła wiedzy semantycznej, a tak będzie w przypadku systemu rozszerzającego istniejący tezaurus omawianego w sekcji 7, oczekujemy, że program na podstawie korpusów tekstu rozszerzy początkowy zasób wiedzy, np. o opis znaczenia słów, które nie występowały w tezaurusie dostarczonym do programu na starcie. Rezultatem procesu wydobywania powinna być wiedza opisująca wybrane aspekty semantyki wyrażeń językowych, wyabstrahowana z dostępnych danych (np. z wystąpień wyrażeń językowych w korpusie) i zapisana w formie określonej przez przyjęty model opisu (inaczej formalizm opisu).

Już na tym etapie wyłaniają się dwie istotne przyczyny, dla których scharakteryzowany powyżej proces automatycznego wydobywania wiedzy semantycznej może być przydatny. Pierwsza ma charakter zewnętrzny w stosunku do lingwistyki, a zarazem utylitarny. W różnych zastosowaniach technologii przetwarzania języka naturalnego, np. w ramach systemów inteligentnego wyszukiwania i wydobywania informacji, bardzo pomocne są wielkie słowniki opisujące znaczenia poszczególnych słów. Pożądane jest, aby stosowany w przetwarzaniu tekstu słownik zapewniał jak największe pokrycie słownictwa występującego w danej dziedzinie, najlepiej bliskie 100%, w tym opis słownictwa specjalistycznego, oraz by był na bieżąco aktualizowany wraz ze zmianami w stosowanej terminologii. Budowa, a następnie nieustanna aktualizacja takiego słownika byłaby bardzo pracochłonna. Metody automatyczne, gdy są zastosowane w oparciu o duże korpusy tekstu, które relatywnie łatwo można zebrać w dobie elektronicznych mediów, mogą wspomóc budowanie słowników tego typu lub nawet mogą być zastosowane do automatycznego generowania pewnego rodzaju słownika bezpośrednio z danych korpusowych.

Druga istotna przyczyna przemawiająca za rozwijaniem metod automatycznych to potencjalna szansa uzyskania innej perspektywy w spojrzeniu na semantykę języka naturalnego. Metody automatyczne mogą dać obraz semantycznych zjawisk językowych w skali masowej. Obraz taki jest uzyskiwany w sposób obiektywny (w zakresie sposobu wytworzenia) z dużego zbioru danych i zależny wyłącznie od tego dostarczonego zbioru danych, czyli korpusu. Uzyskany model (sformalizowany opis znaczenia) może pokazać zależności semantyczne pomiędzy wyrażeniami językowymi — rodzaj struktury semantycznej, która jest często widoczna dopiero po przyjęciu globalnej i ilościowej skali analizy. Stąd też metody automatyczne są potencjalnie użytecznym narzędziem w badaniach lingwistycznych.

2. Formalizacja opisu znaczeń leksykalnych

Każdy program komputerowy czy system informatyczny, niezależnie od zastosowanego sposobu jego konstrukcji (np. języka programowania), jest rodzajem automatu opierającego swoje działanie o skończony zbiór prostych operacji, które może wykonać, i stanów, w których może się znaleźć. Automat taki realizowany jest fizycznie poprzez stany wewnętrzne układów elektronicznych komputera. Oznacza to, że cały proces automatycznego wydobywania wiedzy semantycznej musi być opisany w spo-

sób sformalizowany, tak aby możliwe było jego przełożenie na działania w obrębie programu, czyli precyzyjnie określone zmiany stanów automatu. Program komputerowy definiowany jest poprzez podanie:
— precyzyjnej, formalnej definicji danych wejściowych,
— precyzyjnej, formalnej definicji rezultatu działania (wyjścia) — zarówno formatu, jak i warunków, jakie ma spełniać
— oraz, ostatecznie, poprzez napisanie kodu programu określającego sposób przeprowadzenia operacji generujących rezultat zgodnie ze zdefiniowanymi warunkami na podstawie danych wejściowych.

W konstrukcji metody automatycznego wydobywania wiedzy semantycznej musimy zawsze założyć pewien sformalizowany sposób opisu tej wiedzy, tzn. musimy precyzyjnie określić metajęzyk, którym będziemy się posługiwać. Do opisu wiedzy semantycznej jako metajęzyka możemy użyć dowolnego języka, nawet języka naturalnego, jednak musimy w sformalizowany sposób określić, jak będziemy się nim posługiwać, np. jak będziemy konstruować wyrażenia opisujące znaczenia. Co więcej, musimy później precyzyjnie opisać[3], jak wyrażenia opisujące znaczenia będą budowane na podstawie danych wejściowych pozyskanych z korpusu.

Sformalizowany opis znaczenia wypowiedzi językowej opiera się w większości przypadków na schemacie translacji wyrażeń językowych do wyrażenia zapisanego w wybranym formalnym języku reprezentacji znaczenia (zob. np. Piasecki, 2004). Procedura translacji może uwzględniać również kontekst użycia wyrażenia językowego w ramach danej wypowiedzi, czyli opisywać, w szerszy lub węższy sposób, różne aspekty znaczenia, wkraczając nawet w obszary pragmatyki. Zarówno język reprezentacji znaczenia, jak i sposób translacji są ustalane arbitralnie w ramach przyjętej teorii semantycznej. Najczęściej język reprezentacji znaczenia jest rodzajem języka logicznego, a reprezentację semantyczną zdania lub tekstu stanowią formuły logiczne. Automatyczne wydobywanie wiedzy semantycznej w odniesieniu do poziomu tekstu czy zdania oznaczałoby automatyczną konstrukcję metody translacji wyłącznie na podstawie danych zebranych z korpusu. Nie są szerzej znane próby rozwiązania tak postawionego problemu na skalę praktyczną, tj. dla znaczącego podzbioru języka naturalnego. Stosowane metody automatycznego wydobywania opisów semantycznych wyrażeń językowych o wewnętrznej strukturze składniowej, tzn. wyrażeń zbudowanych z więcej niż jednej jednostki leksykalnej, ograniczają się jedynie do:
— przyjęcia założenia schematu predykatowo-argumentowej budowy wyrażenia[4],
— scharakteryzowania ograniczeń dotyczących własności syntaktycznych i kategorii semantycznych (stosowana jest różnorodna terminologia, por. sekcja 8) wyrażeń językowych wypełniających poszczególne pozycje argumentowe określonych wyrażeń predykatywnych (np. czasowników) — w przypadku kategorii semantycznych, raczej należy mówić o preferowanych kategoriach niż ściśle ograniczonych zbiorach dla poszczególnych pozycji.

3 Wymagany stopień precyzji w posługiwaniu się metajęzykiem bardzo utrudnia zastosowanie języka naturalnego w tej roli, np. konieczne byłoby precyzyjne opisanie znaczeń słów i reguł ich łączenia, czyli *de facto* wykonania tego samego, do czego chcemy wykorzystać język naturalny. Spowodowane jest to koniecznością manipulowania wyrażeniami metajęzyka poprzez program — rodzaj automatu.

4 Wyrażenie o budowie predykatowo-argumentowej jest konstruowane wokół głównego elementu — elementu predykatywnego (np. czasownika), który otwiera określoną liczbę pozycji argumentowych (np. wypełnianych przez rzeczowniki lub frazy przyimkowe).

Do powyższego zagadnienia wrócimy w sekcji 8. Tam też pokrótce omówione zostaną stosowane metody. W dalszej części pracy skoncentrujemy jednak uwagę przede wszystkim na metodach automatycznego wydobywania opisów znaczenia wyrażeń językowych, które są zawarte w słownikach, tj. jedno- i wielosłowowych jednostek leksykalnych.

Z uwagi na różnorodność stosowanej terminologii przyjmijmy na potrzeby niniejszej pracy następujące definicje:

— lematem nazywać będziemy morfologiczną formę wyrazową wybraną jako reprezentant całej grupy form wyrazowych tej samej klasy gramatycznej, w rozumieniu Przepiórkowskiego (2004), opisywanych przez wspólny zbiór kategorii gramatycznych i różniących się jedynie co do wartości przyjmowanych dla poszczególnych kategorii, np. program jako reprezentant grupy: programu, programie, programach… lub maszyna parowa (wielowyrazowy lemat) jako reprezentant grupy: maszyny parowej, maszynie parowej, maszyn parowych…; zakładamy, że wybór określonej formy wyrazowej jako podstawowej jest arbitralny i uwarunkowany jedynie pewną tradycją, np. dla rzeczownika lematem będzie forma liczby pojedynczej w mianowniku, a dla przymiotnika forma liczby pojedynczej w mianowniku i rodzaju męskoosobowym;
— jednostką leksykalną będzie para: lemat i jedno ze znaczeń reprezentowanych poprzez wystąpienia danego lematu w wyrażeniach językowych[5] (por. Derwojedowa i in., 2008; Piasecki i in., 2009c), np. lemat *kolejka* i jego znaczenie oznaczane w Słowosieci 1.0 (Derwojedowa i in., 2009) poprzez symbol `kolejka 3` — rodzaj kolei, środka transportu.

W ramach formalnego języka reprezentacji semantycznej jednostkom leksykalnym odpowiadają zwykle predykaty logiczne. Znaczenie predykatu jest ustalane poprzez jego interpretację w postaci określonej relacji, w ramach przyjętego modelu formalnego, np.:

— dla jednostki leksykalnej ⟨*kolejka*, `kolejka 3`⟩ odpowiadający jej predykat jednoargumentowy (załóżmy, że nazwany *kolejka*_3) otrzyma interpretację w postaci podzbioru obiektów określonego typu wydzielonego spośród wszystkich obiektów modelu, np. *kolejka*_3(o_i) ma wartość prawdy, jeżeli obiekt o_i należy do podzbioru obiektów przypisanego do *kolejka*_3 jako jego interpretacja;
— dla jednostki leksykalnej ⟨*dotykać*, `dotykać 2`⟩ uzyskamy predykat dwuargumentowy *dotykać*_2, interpretowany jako zbiór par obiektów wyznaczający określoną relację — relacja ta reprezentuje określone znaczenie *dotykać* (w tym przypadku składa się z par obiektów wchodzących w fizyczny kontakt), np. *dotykać*_2(o_i, o_j) ma wartość prawdy, jeżeli para obiektów ⟨o_i, o_j⟩ należy do zbioru par przypisanych do *dotykać*_2 jako jego interpretacja.

Zgodnie z ideą opisu znaczenia poprzez translację (por. Piasecki, 2004), interpretacja predykatu, czyli przypisane do predykatu zbiory obiektów lub par lub trójek itd. (w zależności od liczby argumentów predykatu), stanowi jednocześnie formalny opis znaczenia jednostki leksykalnej.

Dla definicji interpretacji języka logicznego jest obojętne, jakie konkretne zbiory obiektów, par, trójek itd. przypiszemy do poszczególnych symboli predykatywnych.

5 Dla uproszczenia zakładamy, że poszczególne formy wyrazowe należące do zbioru reprezentowanego przez lemat nie różnią się pod względem znaczenia. Mówiąc o wystąpieniu lematu w tekście, mamy na myśli wystąpienie jednej z form wyrazowych ze zbioru reprezentowanego przez lemat.

Jednak w perspektywie użycia tego modelu interpretacji do reprezentacji znaczeń leksykalnych konkretna struktura modelu, rozumiana jako zbiór relacji, ma kluczowe znaczenie. Aby ograniczyć możliwe postacie interpretacji do tych, które są zgodne ze znaczeniami wyrażanymi w języku naturalnym, wprowadzono mechanizm postulatów znaczeniowych (ang. *meaning postulates*) (np. Dowty i in., 1981). Postulat znaczeniowy jest rodzajem aksjomatu definiującego ograniczenia nałożone na możliwe interpretacje określonego symbolu predykatywnego lub grup symboli. Postulat znaczeniowy ma zapewnić pewien rodzaj zgodności interpretacji symbolu predykatywnego z opisem rzeczywistości przedstawionym w jednostce leksykalnej reprezentowanej przez dany symbol. Poszczególne postulaty znaczeniowe często przybierają postać złożonego wyrażenia logicznego. Budowa ich spójnego zbioru jest sporym wyzwaniem dla człowieka, a tym bardziej dla jakiegokolwiek programu komputerowego. Konstrukcja zbioru postulatów znaczeniowych w ramach opisu znaczącego podzbioru języka naturalnego metodami semantyki formalnej nie jest, jak dotychczas, znana w literaturze.

Natomiast w ramach sztucznej inteligencji były podejmowane próby budowy wielkich baz wiedzy, w których wiedza była wyrażana za pomocą języka formalnego, w mniejszym lub większym stopniu odwołującego się do języka logiki. W takiej reprezentacji pojęcia są etykietowane lematami języka naturalnego (najczęściej języka angielskiego), co skutkuje pewną relacją pomiędzy bazą wiedzy a systemem znaczeń leksykalnych. Bardzo często jednak relacja ta nie jest definiowana wprost. Przykładem wielkiej bazy wiedzy tego typu może być baza budowana od kilkudziesięciu lat w ramach projektu Cyc® (Siegel i in., 2004). Baza ta ma odzwierciedlić wiedzę przeciętnego dorosłego człowieka. Język reprezentacji wiedzy w Cyc® nie jest językiem logiki, chociaż wywodzi się z logiki predykatów. Występujące tam postulaty znaczeniowe, które specyfikują znaczenie pojęć, nie mają charakteru aksjomatów logicznych.

Postulat znaczeniowy to przykład skrajnej formalizacji opisu znaczeń leksykalnych. Na drugim biegunie leży technika opisu znaczeń w języku naturalnym, znana z tradycyjnych słowników papierowych. W ramach hasła słownikowego etykietowanego lematem poszczególne jednostki leksykalne są opisane w osobnych punktach. Każda jednostka jest opatrzona krótkim opisem w języku naturalnym. Najczęściej zakres używanych konstrukcji składniowych i słownictwa jest ograniczony w opisach zawartych w hasłach słownikowych. Opisy są podawane wraz ze zbiorami przykładów użycia danej jednostki leksykalnej. Taki sposób reprezentacji znaczeń leksykalnych nie jest przydatny jako format wyjściowy dla automatycznego wydobywania wiedzy semantycznej, ponieważ nie jest to format sformalizowany. Synteza poprawnych definicji słownikowych (np. w sensie Grochowski, 1993; Apresjan, 2000; Bańko, 2001) jest poważnym wyzwaniem dla metod inżynierii języka naturalnego. Jak już wspominaliśmy, konieczny byłby w tym celu szczegółowy model struktur językowych i ich znaczeń. Poprawne definicje znaczeń językowych nie pojawiają się również w ogólnym korpusie, chyba że częścią korpusu jest tekst słownika.

Pomiędzy obydwoma skrajnymi podejściami do stopnia formalizacji opisów znaczeń leksykalnych, tj. zastosowaniem z jednej strony języka logiki, a z drugiej języka naturalnego, można odnaleźć szereg innych propozycji, potencjalnie lepiej wspierających potrzeby i możliwości metod automatycznego wydobywania. W ramach semantyki komponentowej (ang. *componential semantics*) (por. Saint-Dizier, Viegas, 1995b; Apresjan, 2000) znaczenie jednostki leksykalnej jest opisywane za pomocą

określonego zbioru cech i ich wartości, które odróżniają ją od innych jednostek leksykalnych. Znaczenie może być również reprezentowane jako wyrażenie w sformalizowanym języku reprezentacji. Wyrażenia takie składają się z podstawowych komponentów znaczeniowych (nazywanych też prymitywami znaczeniowymi, semami lub atomami znaczeniowymi, por. Polański, 1993)[6] połączonych określonym zestawem operatorów (najczęściej jest to język logiczny). Krótki przegląd prac można znaleźć u D. Dowty'ego (1979). Zbiór współczesnych prac, w których pojawiają się idee analizy komponentowej znaczenia leksykalnego jest obszerny, np. Wierzbicka (2006), Pustejovsky (1991) czy też Mel'čuk (1988).

W klasycznych tezaurusach, np. Roget's Thesaurus (Roget, 1856), semantyczne relacje leksykalne, takie jak synonimia[7], hiperonimia[8] czy meronimia[9], łączą jednostki leksykalne w semantyczne grupy stanowiące podstawę opisów znaczeń leksykalnych (por. Piotrowski, 1994). Tezaurusy mogą osiągać znaczne rozmiary, zapewniając pokrycie opisem dużych podzbiorów całego leksykonu. Jednostki leksykalne są różnicowane przede wszystkim w oparciu o ich położenie w sieci relacji. Zbiory jednostek leksykalnych, z którymi dana jednostka pozostaje w relacji synonimii, hiperonimii, hiponimii[10], meronimii itd. stanowią podstawę opisu jej znaczenia[11]. Opis znaczenia jednostki leksykalnej poprzez sieć relacji dostarcza jedynie częściowej wiedzy w porównaniu z analizą komponentową (tzn. metodami semantyki komponentowej), ponieważ w tym pierwszym przypadku znamy jedynie ograniczenia nałożone na znaczenie jednostki leksykalnej, wyrażone poprzez semantyczne relacje leksykalne, w których ona występuje. Natomiast analiza komponentowa może być teoretycznie dowolnie szczegółowa. Dlatego też w tezaurusach jednostki leksykalne bywają opatrzone glosą lub przykładami, np. w Princeton WordNet — wielkim elektronicznym tezaurusie języka angielskiego (Fellbaum, 1998; Miller i in., 2006). Ponieważ roz-

6 Podstawowe komponenty znaczeniowe reprezentują znaczenia pierwotne dla danego języka reprezentacji, które nie są opisywane za pomocą wyrażeń złożonych i są zadane z góry w ramach danego języka reprezentacji znaczeń leksykalnych.

7 Istnieje wiele różnych definicji synonimii (por. Apresjan, 2000). Część z nich opiera się na identyczności zbioru bytów reprezentowanych przez dwie jednostki leksykalne. Inne na możliwości zastępowania użyć jednej jednostki leksykalnej przez drugą. Można również określać synonimię jako dwustronną hiperonimię (por. opis hiperonimii w sekcji 7 oraz Derwojedowa i in., 2007b).

8 Hiperonimia to semantyczna relacja leksykalna zachodząca pomiędzy jednostką leksykalną nadrzędną — o bardziej ogólnym znaczeniu (denotacji) i szerszej pojęciowo — a jednostką podrzędną — o węższym znaczeniu, denotacji, bardziej specyficznym znaczeniu, np. *przedmiot 5 — książka 1* w Słowosieci 1.0 (Derwojedowa i in., 2009; por. definicję hiperonimii w Derwojedowa i in., 2007b).

9 Meronimia to relacja zróżnicowana, obejmująca wiele podtypów, często rozumiana bardzo szeroko, ale dominującym aspektem jej znaczenia jest relacja cząstkowości pomiędzy bytem reprezentowanym przez meronim a bytem reprezentowanym przez holonim, np. *okładka 1* jest w relacji meronimii do *książka 1* (holonimu) w Słowosieci 1.0 (Derwojedowa i in., 2009; por. szczegółową definicję podtypów meronimii w Derwojedowa i in., 2007b). Pojęciowo odwrotną relacją do meronimii jest holonimia, jednak nie w każdym przypadku fakt zachodzenia meronimii pociąga za sobą wystąpienie holonimii (por. Derwojedowa i in., 2008).

10 Relacja odwrotna do hiperonimii (por. Derwojedowa i in., 2007b).

11 Ze względu na problemy z definicją synonimii, w Słowosieci podstawowymi relacjami są hiperonimia/hiponimia oraz holonimia/meronimia, natomiast synonimia jest definiowana wtórnie, w oparciu o relacje podstawowe. Jednostki leksykalne są grupowane w zbiory prawie synonimów, wtedy gdy współdzielą powiązania wyznaczone przez relacje podstawowe (Piasecki i in., 2009c).

wiązanie to sprowadza się praktycznie do wprowadzenia opisu w języku naturalnym, glosy nie nadają się jako format wyjściowy dla algorytmów wydobywania wiedzy semantycznej z korpusu.

Budowa dużych leksykonów zgodnie z paradygmatem semantyki komponentowej jest bardzo pracochłonna i wraz ze wzrostem rozmiaru słownika narasta problem zapewnienia spójności opisu. Nie istnieją sformalizowane słowniki oparte na tym paradygmacie, które by opisywały znaczący podzbiór słownictwa.

Z drugiej strony, ograniczona wiedza zawarta w sieciach semantycznych relacji leksykalnych jest wystarczająca dla wielu zastosowań inżynierii języka naturalnego, np. Princeton WordNet doczekał się setek zastosowań (por. Rosenzweig i in., 2007; Piasecki i in., 2009c). Tezaurusy tego typu stały się bardzo cennym zasobem dla systemów informatycznych przetwarzających język naturalny, np. systemów inteligentnego wyszukiwania informacji, zarządzania wiedzą czy też automatycznego tłumaczenia. Koncepcja reprezentacji znaczenia jednostek leksykalnych poprzez sieć semantycznych relacji leksykalnych zaowocowała rozwojem wielkich wordnetów, tj. elektronicznych tezaurusów wzorowanych na Princeton WordNet, np. powstały między innymi powiązane wordnety dla wielu języków europejskich: EuroWordNet (Vossen, 2002) i BalkaNet (Tufiş i in., 2004) oraz ostatnio również dla języka polskiego Słowosieć 1.0 (ang. nazwa plWordNet 1.0) (Derwojedowa i in., 2009; Derwojedowa i in., 2008; Piasecki i in., 2009c).

Semantyczne relacje leksykalne pojawiają się również jako elementy sformalizowanego opisu znaczeń leksykalnych w wielu podejściach (np. Pustejovsky, 1991; Mel'čuk, 1988).

Podsumowując, można zauważyć dwa podstawowe typy elementów składowych sformalizowanych opisów znaczeń leksykalnych: komponenty znaczeniowe oraz instancje semantycznych relacji leksykalnych12. Komponenty znaczeniowe posiadają jedynie znaczenie jako elementy konstrukcyjne złożonych opisów. Automatyczna konstrukcja takich opisów na podstawie korpusu jest procesem bardzo trudnym (por. Baldwin, 2007). Dlatego też w dalszej części pracy skoncentrujemy uwagę na automatycznym wydobywaniu instancji semantycznych relacji leksykalnych

3. Przesłanki dostępne dla metod automatycznych

W analizie korpusu pod kątem poszukiwania informacji, która jest istotna dla budowy opisu znaczeń leksykalnych, a która jednocześnie jest możliwa do wykorzystania przez metody automatyczne, możemy przyjąć jedną z dwóch podstawowych perspektyw:

— analizy pojedynczych wystąpień lematów
— analizy danych statystycznych dotyczących dystrybucji lematów w całym korpusie.

W myśl pierwszej z nich, analiza szczegółowej informacji charakteryzującej współwystąpienia par lub większych grup lematów[13] w zakresie wiążących je struktur

12 Poprzez instancję semantycznej relacji leksykalnej rozumiemy tu parę jednostek leksykalnych powiązanych określoną relacją — relację można opisać poprzez zbiór par.

13 W tym miejscu i w dalszej części pracy, pisząc o analizie wystąpień lematów w korpusie, dokonujemy świadomie pewnych skrótów myślowych. Po pierwsze, lemat reprezentuje jedną lub więcej jednostek leksykalnych, które są przedmiotem opisu znaczenia. Odnosząc pewne cechy semantyczne do lematu,

leksykalno-składniowych, lub nawet semantycznych, pozwoli nam na określenie semantycznych relacji leksykalnych wiążących analizowane lematy, np. analizując powiązania składniowe rzeczowników z konkretnymi przymiotnikami, możemy wyciągać pewne wnioski co do cech powiązanych ze znaczeniem poszczególnych rzeczowników. W przypadku analizy statystycznej przyjmujemy perspektywę globalną: analizujemy poszczególne wystąpienia lematów z potencjalnie niższą dokładnością w zakresie rozpoznawania struktur językowych, w których występują, ale jednocześnie zakładamy, że popełniane przy tym błędy zostaną zniwelowane dzięki statystycznej analizie zgromadzonego materiału.

3.1. Opisy znaczeń zawarte w tekście

W tych gatunkach tekstu, w których najważniejsza jest komunikacja informacji, wprowadzenie nowego terminu jest, a przynajmniej powinno być, skojarzone z jego opisem. Opis taki może przybierać formę definicji wiążącej lemat reprezentujący termin definiowany z wyrażeniem definiującym. Takie nieformalne definicje mogą być nawet bliskie wymogom stawianym definicji słownikowej (np. Grochowski, 1993; Apresjan, 2000; Bańko, 2001). W przypadkowo wybranym fragmencie polskiej Wikipedii (Fundacja Wikimedia, 2009) (wybranej jako źródło w całkowicie zamierzony sposób) możemy odnaleźć rodzaj definicji odnoszącej się do lematu lemat:

> Formalnie jednak każdy lemat jest pełnoprawnym twierdzeniem, a zaklasyfikowanie pewnego twierdzenia jako lematu wynika jedynie ze sposobu jego użycia w innym, obszerniejszym kontekście.

Warto zauważyć, że powyższa definicja odnosi się do innej jednostki leksykalnej, nazwijmy[14] ją `lemat _ 2`, reprezentowanej przez lemat *lemat*, niż jednostka najczęściej reprezentowana przez *lemat* w tej pracy, nazwijmy ją `lemat _ 1`. Odpowiednio `lemat _ 2` odnosi się do pojęcia matematycznego, a `lemat _ 1` do abstrakcyjnego elementu struktury językowej.

Fragment tekstu użyty powyżej jako przykład pochodzi ze źródła szczególnego, jednak Wikipedia, z racji swojej dostępności i wielkości, jest często wykorzystywana jako korpus tekstów. W ogólnym korpusie tego typu konstrukcje językowe zbliżone do definicji będą jednak rzadkie. Znacznie częściej można napotkać wyrażenia wiążące parę lematów w sposób sygnalizujący występowanie określonej semantycznej relacji leksykalnej pomiędzy parą jednostek leksykalnych reprezentowanych przez te lematy, np. para: *pracownia–miejsce* powiązane relacją hiperonimii w przykładzie z Korpusu IPI PAN (Przepiórkowski, 2004):

> *Pracownia* nie tylko będzie służyła młodej artystce, ale także stanie się *miejscem* warsztatów i spotkań artystycznych.

czynimy to z myślą o grupie jednostek leksykalnych reprezentowanych przez niego. Po drugie, lemat nie ma bezpośredniej reprezentacji w tekście, a jedynie pośrednią, poprzez wyrazy odpowiadające formom wyrazowym reprezentowanym przez dany lemat. Mówiąc o wystąpieniu lematu, mamy na myśli wystąpienie jednej z form wyrazowych reprezentowanych przez niego.

14 Nie jest opisana jeszcze w Słowosieci.

W wielu konstrukcjach językowych zawierających parę lematów, o których wiadomo, że są powiązane pewną semantyczną relacją leksykalną, można zidentyfikować fragmenty struktury leksykalno-składniowej, które są charakterystyczne dla tego typu konstrukcji. Takie charakterystyczne elementy składowe i zależności w obrębie struktury składniowej można opisać sformalizowanym wzorcem. Wzorzec taki definiuje elementy leksykalno-strukturalne konstrukcji językowej, po której możemy ją rozpoznać. Dla podanego powyżej przykładu możemy sformułować następujący prosty wzorzec (abstrahujemy tu od konkretnego formalnego języka definicji wzorców):

$$\texttt{NLem1}_{\text{przypadek=mianownik}} \ldots \texttt{lemat=}\textit{stać się} \ldots \texttt{NLem2}_{\text{przypadek=narzędnik}}$$

Powyższy wzorzec odpowiada nieskończonej liczbie zdań, w tym wielu takim, w których lematy identyfikowane na pozycjach `NLem1` i `NLem2` nie pozostają z sobą w relacji hiperonimii[15]. Można jednak uzyskać większą dokładność przesłanki, zwiększając szczegółowość opisu konstrukcji językowej zawierającej oba lematy. Aby wykorzystać w ramach przetwarzania automatycznego informację wyrażaną przez tego typu konstrukcje językowe, niezbędne jest zastosowanie pewnych mechanizmów wydobywania tego typu związków pomiędzy relacją semantyczną a elementami struktury leksykalno-syntaktycznej oraz rozpoznawania tego typu struktur w korpusie. Do zagadnienia tego wrócimy w sekcji 5.

3.2. Hipoteza dystrybucyjna

Struktura wyrażenia językowego jest w większości wypadków zdeterminowana własnościami jego składników. Kolokacje *sensu stricto*, które charakteryzują się brakiem kompozycyjności w warstwie składni bądź semantyki i które wyłamują się z powyższej zależności, nie wydają się stanowić znaczącej części wystąpień wyrażeń językowych w tekście.

Zależność struktury całości od składników obejmuje również zależność od własności semantycznych składników, tzn. złożone wyrażenia językowe w pewien sposób determinują własności wyrażeń składowych, z których są zbudowane. Analiza wystarczająco dużej liczby wystąpień wyrażeń językowych powinna doprowadzić do identyfikacji regularności w użyciu poszczególnych wyrażeń językowych, które wynikają z właściwości semantycznych tych wyrażeń. Ta ogólna idea pojawiła się w przynajmniej kilku lingwistycznych teoriach, jednak w wyraźny sposób została sformułowana przez Harrisa (1968) w postaci tzw. hipotezy dystrybucyjnej (za Sahlgren, 2001 — tłumaczenie własne):

> Znaczenie bytów językowych i znaczenie relacji gramatycznych pomiędzy nimi jest powiązane z ograniczeniami nałożonymi na kombinacje tych bytów w relacji do innych bytów językowych.

Pojawiające się w hipotezie pojęcie bytu językowego nie zostało precyzyjnie określone. Bytem może być dowolny element wyróżniany w opisie języka. Ze względu

15 Elżbieta Hajnicz w swojej recenzji wstępnej wersji tej pracy podpowiedziała doskonały przykład: „Miejsce to nie tylko będzie służyło młodej artystce, ale stanie się pracownią i dla jej przyjaciół".

jednak na charakter metod omawianych w niniejszej pracy, w dalszych rozważaniach ograniczymy się jedynie do odniesienia hipotezy do wyrażeń językowych, czyli bytów językowych obserwowanych bezpośrednio w tekście. Biorąc pod uwagę ograniczenia w możliwości automatycznego wydobywania opisów semantycznych wyrażeń językowych, wspomniane w sekcji 2, zawęzimy w dalszych rozważaniach klasę rozpatrywanych wyrażeń językowych jedynie do tych, które reprezentują w tekście wystąpienia jedno- lub wielowyrazowych lematów.

W wypadku analizy dużego korpusu na poziomie tekstu trudno jest znaleźć nietrywialne ograniczenia na wystąpienia poszczególnych lematów, które byłyby zawsze spełnione. Dlatego bardziej praktyczne jest postrzeganie ograniczenia z hipotezy dystrybucyjnej jako rodzaju preferencji o różnej sile.

Przyjęło się opisywać zbiór ograniczeń nałożonych na dystrybucję danego lematu metodą nie wprost poprzez określenie kontekstów tekstowych[16] (dalej kontekstów), w jakich dany lemat może wystąpić (por. Widdows, 2004). Konteksty są opisywane przez cechy odnoszące się do wyrażeń językowych i struktur językowych reprezentowanych przez poszczególne konteksty. Cechą może być wystąpienie określonego wyrażenia lub relacji strukturalnej (tj. syntaktycznej lub semantycznej) w danym kontekście, np. `wystąpienie w roli argumentu określającego instrument w strukturze czasownika` *`uderzyć`*. Jednakże i w tym wypadku powinniśmy przypisać do każdej cechy siłę powiązania odzwierciedlającą jak bardzo typowe jest dla danego lematu występowanie w kontekstach charakteryzowanych przez daną cechę. Gdy jako cechę przyjmujemy nie tylko sam fakt wystąpienia pewnego wyrażenia językowego lub relacji w danym kontekście, ale również bierzemy pod uwagę częstość tych wystąpień w danym kontekście, to możemy mówić o parze cecha–wartość[17], a wtedy siłę powiązania musimy odnosić również do pary cecha–wartość.

Aby zastosować hipotezę dystrybucyjną jako podstawę analizy znaczenia leksykalnego, musimy odpowiedzieć na dwa pytania:

— Jakiego rodzaju cechy będą używane do opisu kontekstów?
— Czym jest kontekst oraz jak duży wycinek tekstu obejmuje?

W przypadku pierwszego pytania istnieją trzy podstawowe sposoby zdefiniowania cechy:

1) samo wystąpienie opisywanego lematu *x* w określonym kontekście, np. w pewnym dokumencie *d*, *de facto* konkretny kontekst staje się cechą,
2) współwystąpienie *x* z określoną formą wyrazową *y* (lub określonym lematem) w ramach tego samego kontekstu,
3) wystąpienie *x* jako elementu składowego w ramach wystąpienia określonej relacji leksykalno-syntaktycznej — czyli *x* pojawia się jako część instancji (por. przypis 11) danej relacji łączącej *x* z pewnymi wyrażeniami, całość instancji relacji

16 Używamy tu pojęcia kontekst w bardzo wąskim zakresie, stosowanym w leksykografii (por. Bańko, 2001), w którym kontekst to wyrażenia językowe lub całe zdania, które mogą poprzedzać lub następować po wystąpieniu analizowanego lematu. W definicjach słownikowych tak rozumiany kontekst jest używany do ujednoznacznienia jednostki leksykalnej, do której w danym momencie się odnosimy.

17 W najprostszym wypadku wartością jest częstość (tj. liczba) wystąpień wyrażenia lub relacji, do którego odnosi się cecha w danym kontekście. Ta pierwotna wartość może być jednak później poddawana różnorodnym przekształceniom, które będą omawiane w sekcji 6.

musi zawierać się w danym kontekście, np. wystąpienie *x* jako argumentu określonego wyrażenia predykatywnego *z*, w którym część lub wszystkie pozostałe argumenty są ustalone na określone wyrażenie językowe y_1, … y_n (por. szczegółowe przykłady podane w sekcji 6).

Rozmiar kontekstu może być utożsamiony z:
— rozmiarem całego tekstowego obiektu,
— oknem tekstowym — czyli wycinkiem tekstu obejmującym wystąpienie analizowanego lematu i określoną liczbę tokenów[18] w otoczeniu,
— lub określoną konstrukcją składniową, np. frazą lub zdaniem.

Obliczenie wartości cech dla kontekstów, w których występuje dany lemat *x* oraz ocena statystyczna siły powiązania wystąpień *x* z wystąpieniami poszczególnych par cecha–wartość skutkuje budową statystycznego modelu dystrybucji *x* w korpusie. Zgodnie z hipotezą dystrybucyjną porównanie dystrybucji dwóch lematów pozwala na ocenę stopnia powiązania ich znaczeń. Uproszczeniem byłoby mówić, że porównanie dystrybucji daje w wyniku ocenę podobieństwa znaczeń. Jak zobaczymy w sekcji 6, związek pomiędzy miarą numeryczną wyłaniającą się z porównania a naturą związku semantycznego pomiędzy porównywanymi lematami zależy od rodzaju zdefiniowanych cech, a nawet samego sposobu porównywania dystrybucji lematów.

4. Paradygmaty wydobywania relacji semantycznych

Prace nad automatycznym wydobywaniem semantycznych relacji leksykalnych z korpusów prowadzone są już od kilkunastu lat (np. Crouch i Yang, 1992; Hearst, 1992; Grefenstette, 1993). Semantyczne relacje leksykalne mogą być obserwowane w korpusie na podstawie obydwu typów przesłanek omówionych w sekcji 3: definicji pojawiających się w tekście (por. sekcja 3.1) oraz powiązania znaczeniowego obserwowanego na podstawie dystrybucji. W tym drugim wypadku można oczekiwać, że powiązane znaczeniowo lematy reprezentują jednostki leksykalne powiązane pewną relacją semantyczną. W toku rozwoju metod wyłoniły się dwa podstawowe paradygmaty:
— wzorców leksykalno-syntaktycznych — omawianych w sekcji 5,
— semantyki dystrybucyjnej — omawianych w sekcji 6.

Można również zaobserwować stopniowe zacieranie się różnicy pomiędzy statystyczną naturą metod opartych na dystrybucji i metod opartych na wzorcach identyfikujących poszczególne wystąpienia określonych struktur językowych (np. Pantel, Pennacchiotti, 2006; Kurc, Piasecki, 2008), jak również budowanie kompleksowych metod łączących w sobie wyniki algorytmów różnego typu (np. Piasecki i in., 2009d, c). Wybrane podejścia tego typu zostaną omówione w sekcji 7.

5. Metody oparte na wzorcach leksykalno-syntaktycznych

W tradycyjnym słowniku poszczególne hasła zawierają definicje sformułowane w języku naturalnym, a nie języku formalnym, co uniemożliwia bezpośrednie odczytanie za-

18 Token w tym kontekście to podstawowy, niepodzielny segment (por. Przepiórkowski, 2004), który jest wyróżniany w tekście. Token może być wyrazem, ale również liczbą czy też dowolnym symbolem wyrazowym i niewyrazowym.

wartej w nich wiedzy za pomocą metod automatycznych (por. sekcję 2). Jednak zwykle definicje słownikowe posiadają podobną strukturę, są napisane w podobny sposób (co redukuje zakres pojawiających się konstrukcji składniowych) oraz wiążą definiowany lemat z innymi lematami w sposób sygnalizujący określone relacje semantyczne, np. relację hiperonimii. Na tej podstawie rozwinęło się szereg metod wydobywania semantycznych relacji leksykalnych ze słowników w wersji elektronicznej (ang. *machine readable dictionaries*), opartych na ręcznie konstruowanych wzorcach w postaci wyrażeń regularnych[19], pozwalających na zidentyfikowanie w treści definicji słownikowych wystąpień lematów powiązanych z lematem opisywanym za pomocą określonej relacji semantycznej. Były również podejmowane próby zastosowania tego podejścia w odniesieniu do języka polskiego (np. Martinek, 1997; Ceglarek, Rutkowski, 2006).

W przeciwieństwie do definicji słownikowych zmienność struktur językowych jest w korpusie nieograniczona. Miejsca wystąpień fragmentów tekstu zbliżonych do definicji muszą być najpierw wykryte, a ponadto w ogólnym przypadku nie należy spodziewać się w korpusie wystąpień poprawnych definicji znaczeń leksykalnych. Jednak Hearst (1992) pokazała, że zastosowanie wzorców podobnych do tych użytych do przetwarzania słowników, czyli o sile ekspresji wyrażeń regularnych, może przynieść wartościowe rezultaty również w odniesieniu do tekstu w korpusie, np.:

```
NP0 ….. such as {NP1, NP2 . . . . (and | or ) NPn
```

Powyższy wzorzec opisuje wystąpienie w tekście kilku fraz nominalnych, w taki sposób, że pierwsza `NP0` jest oddzielona wyrażeniem *such as*, a kolejne tworzą frazę złożoną współrzędnie. Zgodnie z interpretacją przypisaną do wzorca na podstawie analizy przykładów `NP0` reprezentuje hiperonim, natomiast `NP1 – NPn` jego hiponimy[20]. Opierając się na analizie zdań z korpusu zawierających znane pary hiperonimiczne, Hearst (1992, 1998) zaproponowała pięć produktywnych wzorców o relatywnie dobrej dokładności. Dokładność została określona jako stosunek liczby wydobytych poprawnych par hiperonimicznych do liczby wszystkich wydobytych par. Dla zaprezentowanego powyżej wzorca 61 z 106 wydobytych par lematów z *Grolier Encyclopedia* występowało w ówczesnej wersji Princeton WordNetu (Hearst, 1992), co jest wynikiem relatywnie dobrym, biorąc pod uwagę dodatkowo fakt, że rozmiar ówczesnej wersji Princeton WordNetu był ograniczony. Warto jednak podkreślić, że Hearst (1992) wykorzystała płytki parser[21], który w przypadku wielu języków, np. języka polskiego, jest niedostępny lub nie zapewnia jeszcze takiego pokrycia przetwarzanych konstrukcji językowych, jakie jest odpowiednie dla analizy dużego korpusu.

Przydatność wzorców w wydobywaniu semantycznych relacji leksykalnych opiera się na założeniu, że możliwe jest skonstruowanie wzorców dostatecznie precyzyjnych,

19 Wyrażenie regularne pod względem siły ekspresji jest równoważne gramatyce regularnej, która dopuszcza jedynie reguły o prostym schemacie dołączania pojedynczych elementów i ma bardzo ograniczoną siłę ekspresji.

20 Za pomocą wzorca wydobywana jest para lematów (uzyskanych w wyniku sprowadzenia poszczególnych fraz nominalnych do lematów), natomiast relacja hiperonimii wiąże parę jednostek leksykalnych przez nie reprezentowanych.

21 Program do automatycznej analizy składni. Płytki parser nie buduje kompletnego opisu struktury składniowej zdania, a jedynie wyznacza granice poszczególnych elementów składowych (por. Piasecki, 2008; Przepiórkowski, 2008).

aby wyciągać wnioski co do faktu powiązania określonych jednostek leksykalnych na podstawie pojedynczych przykładów ich współwystępowania w korpusie.

Mimo relatywnie wysokiej dokładności metody opartej na wzorcach pojawiają się jednak problemy wynikające z samej jej natury. Poważnym wyzwaniem jest odróżnienie metaforycznych użyć od literalnych, np. u Hearst (1992) *aeroplane* (*samolot*) został zidentyfikowany jako hiponim *target* (*cel*) czy też *Washington* jako hiponim *nationalist* (*nacjonalista*). Rozwiązanie tego problemu wymagałoby głębokiej analizy semantycznej i pragmatycznej. Kolejnym poważnym problemem jest bardzo ograniczona częstość występowania wyrażeń odpowiadających wzorcom w korpusie, np. tylko 46 wyrażeń dopasowanych do prezentowanego wcześniej wzorca zostało wydobytych z korpusu „New York Times" liczącego 20 milionów słów (Hearst, 1992).

Próby rozszerzenia metody ręcznie konstruowanych wzorców na semantyczne relacje leksykalne inne niż hiperonimia nie powiodły się, np. Berland i Charniak (1999) osiągnęli dużo gorsze wyniki dla meronimii. Dalsza droga rozwoju wiodła w kierunku metod półautomatycznej konstrukcji wzorców (np. Jacquemin, 2001), łączenia wzorców z maszynowym uczeniem się (np. Girju i in., 2006) czy ostatecznie rezygnacji z zasady wydobywania instancji relacji z pojedynczych wystąpień na rzecz statystycznej analizy wystąpień wielu bardziej ogólnych wzorców w dużym korpusie budowanym z zasobów Internetu (Pantel, Pennacchiotti, 2006).

Większość znanych metod wydobywania relacji semantycznych w oparciu o wzorce została zastosowana do tekstów w języku angielskim. W przypadku języka o słabo ograniczonym szyku i bogatej fleksji ograniczona siła ekspresji wzorców opartych na wyrażeniach regularnych może sprawiać problemy. Dlatego też, na potrzeby półautomatycznej metody rozszerzania polskiego wordnetu, zastosowaliśmy wzorce wyrażone w języku definiowania złożonych ograniczeń morfo-syntaktycznych, nazwanym JOSKIPI (Piasecki, 2006; Piasecki, Radziszewski, 2009), który jest wykorzystywany w tagerze morfo-syntaktycznym języka polskiego o nazwie TaKIPI (Piasecki, 2007; Piasecki, 2008). JOSKIPI zapewnia między innymi operatory umożliwiające: sprawdzanie własności morfo-syntaktycznych poszczególnych wyrazów, kompatybilności morfo-syntaktycznej[22] poszczególnych wyrazów i sekwencji wyrazów lub zastosowanie zdefiniowanego złożonego ograniczenia do całej sekwencji wyrazów o nieustalonej z góry długości (np. przeszukiwanie do spełnienia podanego ograniczenia lub granicy zdania). Skonstruowane zostało sześć produktywnych wzorców (por. Piasecki i in., 2009c), schemat jednego z nich został zaprezentowany poniżej[23]:

```
NP1 (Adj|Adv|Noun|,)*
  (base∈{i, oraz) (base∈{inny, pozostały, nmb=pl) (Adj|Adv)*
NP2(cas=cas(NP1))
```

Powyższy wzorzec identyfikuje lemat (potencjalnie wielowyrazowy) frazy `NP1` jako hiponim lematu frazy `NP2`.

22 Wynikającej zarówno z uzgodnienia wartości poszczególnych kategorii gramatycznych, jak i narzucenia wymaganej wartości.

23 Prezentowana postać to pojęciowy, uproszczony schemat, w którym wyrażenia JOSKIPI zostały zastąpione etykietami opisującymi poszczególne wyrazy i grupy wyrazów. Szczegółowy opis przykładowego wzorca można znaleźć w pracy Piasecki i in., 2009c, 105.

Skonstruowane zostały jeszcze dwa inne wzorce o podobnej dokładności i charakterze rozpoznawanych związków. Zostały one oparte na wyznacznikach leksykalnych: *taki jak* i *w tym*. Wszystkie trzy wzorce były stosowane jako jeden połączony wzorzec złożony na zasadzie rozpoznania jednej z opisywanych konstrukcji. Poniżej został zaprezentowany przykład konstrukcji językowej zgodnej ze wzorcem opartym na wyznaczniku leksykalnym taki jak:

> *Betondour* doskonale nadaje się do wykończania podłóg w pomieszczeniach **takich jak** *garaże*, *warsztaty samochodowe*, *magazyny*, *sklepy*, *pomieszczenia produkcyjne*, *piwnice* czy wykonane z betonu schody.

Opisany powyżej wzorzec złożony został zastosowany do wydobycia par hiperonimicznych z trzech korpusów:

— *Korpusu IPI PAN* (zawierającego około 254 milionów tokenów) (Przepiórkowski, 2004);
— korpusu elektronicznej edycji gazety „Rzeczpospolita" od stycznia 1993 do marca 2002 (około 113 milionów tokenów) („Rzeczpospolita", 2008);
— oraz korpusu dużych plików tekstowych w języku polskim (około 214 milionów tokenów) pobranych z Internetu; w korpusie znalazły się tylko dokumenty zawierające mały procent wyrazów nierozpoznanych przez analizator morfologiczny Morfeusz (Woliński, 2006) (zastosowano również dodatkową kontrolę ręczną tekstu dla przypadków granicznych) i niepowielające się z dokumentami z żadnego z pozostałych dwóch korpusów.

W dalszej części pracy, w przypadku łącznego użycia wszystkich trzech korpusów, będziemy mówić o połączonym korpusie. Aby zwiększyć dokładność działania wzorców, ograniczyliśmy zastosowanie wzorca jedynie do tych przypadków, w których obydwa lematy (potencjalnie wielowyrazowe[24]) z wydobywanej pary znajdowały się na wcześniej zdefiniowanej liście. Na liście tej umieściliśmy wszystkie lematy przewidziane do umieszczenia w nowej, rozszerzonej wersji Słowosieci, tj. wersji 1.0 (Derwojedowa i in., 2009; Piasecki i in., 2009c). W celu poprawy efektywności działania programu wydobywającego pary, wzorce były stosowane jedynie do tych zdań z połączonego korpusu, w których występowała przynajmniej jedna para lematów z zadanej listy.

Na przygotowanej liście lematów znalazło się 13 285 nominalnych lematów pochodzących z: wczesnej wersji Słowosieci (tzw. jądra Słowosieci) (Derwojedowa i in., 2008) — 5340, małego słownika polsko-angielskiego (Piotrowski i Saloni, 1999), ogólnego słownika języka polskiego (PWN, 2007) — dwuwyrazowowe lematy oraz Korpusu IPI PAN (Przepiórkowski, 2004) — tylko najczęstsze lematy nominalne, które dopełniły listę do założonego wstępnie rozmiaru.

Ocena dokładności wydobywania par hiperonimicznych osiągniętej przy użyciu omawianego złożonego wzorca została przedstawiona w tabeli 1. Natomiast w tabeli 2 zaprezentowane zostały przykłady wydobytych par. Dokładność została oszacowana na podstawie ręcznej oceny wylosowanych przez autora podzbiorów wszystkich

24 Wystąpienia wielowyrazowych lematów były rozpoznawane w korpusie za pomocą odpowiednio napisanych wyrażeń JOSKIPI, opisujących strukturę form wyrazowych odpowiadających danemu lematowi. Wyrażenia te zostały automatycznie wydobyte z połączonego korpusu (por. Broda i in., 2008a).

par wydobytych z poszczególnych korpusów. Wielkość podzbiorów została dobrana tak, aby były one statystycznie reprezentatywne dla całości wydobytych danych (por. Piasecki i in., 2009c). Jako poprawnie wydobyte były traktowane wszystkie pary połączone relacją synonimii lub hiperonimii, nawet jeżeli w Słowosieci lematy z danej pary były połączone relacją hiperonimii jedynie pośrednio, tzn. dla wydobytej pary $\langle x, y \rangle$ istniała przynajmniej jedna jednostka leksykalna z występująca w ciągu hiperonimicznym pomiędzy x i y. Przypadki, w których x i y są synonimiczne w Słowosieci (tzn. należą do tego samego synsetu) były traktowane również jako poprawnie wydobyte. Ocena oparta była na definicjach synonimii i hiperonimii przyjętych w ramach projektu Słowosieci (Derwojedowa i in., 2007b).

Tabela 1. Rezultaty automatycznego wydobywania par hiperonimicznych za pomocą wzorców leksykalno-morfo-syntaktycznych skonstruowanych ręcznie (Dok. to dokładność oceniona ręcznie na podstawie reprezentatywnej próbki wydobytych par według zasad podanych w tej pracy)

Korp. IPI PAN		Korp. z Internetu		Korp. „Rzeczypospolitej"		Połączony korp.	
Licz. par	Dok.	Licz. par	Dok.	Licz. par	Dok.	Licz. par	Dok.
14611	30,06%	5983	32,52%	6682	33,16%	24437	30,69%

Wydobyta lista par nie nadaje się do bezpośredniego rozszerzania wordnetu lub słownika — dokładność jest bowiem zbyt niska, aczkolwiek jest porównywalna z wynikami prezentowanymi w literaturze dla języka angielskiego. Lista ta również nie jest dobrym narzędziem wspierającym pracę lingwisty, ponieważ poprawnych par jest dużo mniej niż 50%, a połowa podpowiedzi trafnych wydaje się być granicą, przy której narzędzie staje się akceptowalne przez lingwistę (por. Piasecki i in., 2009c).

Jeżeli wyznaczymy zbiór takich par lematów, że każda para występuje przynajmniej w dwóch zbiorach z trzech: zbiorze par wydobytych za pomocą złożonego wzorca (omawianego powyżej) i dwóch zbiorów par wydobytych za pomocą pozostałych dwóch zastosowanych wzorców, to uzyskany w ten sposób zbiór wykazuje dokładność 41,05% (Piasecki i in., 2009c). Dokładność wzrasta za cenę znacznie mniejszej liczby wydobytych par. Jeszcze większą dokładność 74,03% można uzyskać, wyznaczając część wspólną wszystkich trzech zbiorów, ale wtedy liczba par spada do 620 i niestety są to dość trywialne przypadki.

Za pomocą ręcznie skonstruowanych skomplikowanych wzorców leksykalno-syntaktycznych można wydobyć pary hiperonimiczne z relatywnie dużą dokładnością, ale kosztem ograniczonej liczby wydobytych par i znacznej ilości czasu poświęconego na konstrukcję wzorców i ich dopracowywanie w oparciu o analizę ich działania na korpusie. Z uwagi na swoją dużą siłę ekspresji tak szczegółowe wzorce są trudne do automatycznego pozyskania z korpusu. Potencjalne metody automatyczne wymagałyby zastosowania zaawansowanych technik maszynowego uczenia się.

Pantel i Pennacchiotti (2006) zaprezentowali algorytm Espresso, który pozyskuje automatycznie z korpusu zbiór prostych, ogólnych wzorców leksykalno-syntaktycznych, a następnie stosuje go w celu wydobycia instancji relacji semantycznej zadanej na starcie algorytmu. Poszukiwana relacja jest określana za pomocą zbioru przykładowych instancji. Zmodyfikowaną wersję Espresso i zaadaptowaną do cech języka

polskiego przedstawili pod nazwą Estratto Kurc i Piasecki (2008). Dla relacji hiperonimii uzyskano dokładność większą niż za pomocą jakiegokolwiek pojedynczego wzorca skonstruowanego ręcznie. Dodatkowo wydobyto kilka razy większą liczbę par lematów.

Tabela 2. Przykłady par lematów wydobytych z połączonego korpusu za pomocą złożonego wzorca łączącego wzorce leksykalno-morfo-syntaktyczne oparte na wyznacznikach leksykalnych: *i inny, taki jak* oraz *w tym*

Instancje hiperonimii		Niehiperonimiczne powiązania	
koncesja	decyzja	przepis	kwestia
kapłan	człowiek	silnik	jednostka
maj	okres	człowiek	drzewo
kwestia	problem	program	działanie
sowa	ptak	muzyka	dźwięk
klient	osoba	istota	nic
pielęgniarka	osoba	wojsko	organizacja
profesor	człowiek	stowarzyszenie	instytucja
galeria	miejsce	cień	wróg
matematyka	przedmiot	książka	materiał
matka	kobieta	słońce	czynnik
helikopter	maszyna		
droga	szlak		
zespół	grupa		
mecz	spotkanie		
restrukturyzacja	zmiana		
konsument	osoba		
tenis	sport		
festiwal	impreza		
dziennik	dokument		
medycyna	nauka		
anioł	istota		
spółka	firma		
szczur	szkodnik		
skorpion	znak		
rak	choroba		
nagroda	wyróżnienie		

Algorytmy Espresso i Estratto mogą zostać zastosowane do wydobywania dowolnej relacji semantycznej, która objawia się w korpusie poprzez pewne wyznaczniki leksykalno-syntaktyczne, np. Espresso był z powodzeniem stosowany do wydobywania hiperonimii, ale również innych relacji, takich jak *całość–część*, *reagowanie* (w sensie chemicznym) czy też *produkowanie* (producent–produkt).

Obydwa algorytmy działają według tego samego ogólnego schematu przedstawionego poniżej.

1. Do algorytmu zostaje dostarczony zbiór przykładowych instancji (par lematów, por. przypis 11) relacji, która będzie wydobywana przez algorytm.
2. Wszystkie bliskie współwystąpienia lematów należących do jednej instancji (np. współwystąpienia w obrębie kilku słów lub zdania) zostają odnalezione w korpusie i na ich podstawie generowane są wzorce w formie uogólnionych opisów[25] sekwencji tokenów występujących pomiędzy parami lematów.

25 Tzn. z zaobserwowanych w tekście sekwencji tokenów wyodrębniane są elementy wspólne dla wielu wystąpień par.

3. Dla każdego wzorca zostaje obliczona ocena jakości jego działania nazwana miarą niezawodności (ang. *reliability*). Miara ta jest obliczana na podstawie instancji relacji wydobytych[26] przez dany wzór oraz niezawodności tych instancji (niezawodność przykładowych instancji jest ustawiona początkowo na 1).
4. Podzbiór wzorców o najwyższej niezawodności zostaje zachowany i następnie zastosowany do wydobycia nowego zbioru instancji relacji, pozostałe wzorce są usuwane.
5. Na koniec niezawodność wydobytych instancji zostaje obliczona w analogiczny sposób do niezawodności wzorców[27]; podzbiór instancji o największej niezawodności zostaje zachowany i algorytm powraca do etapu 2, tylko tym razem z zachowanym podzbiorem wydobytych i ocenionych instancji, a nie zadanym *a priori* zbiorem przykładowym.

Naprzemienne wydobywanie wzorców i instancji relacji jest w algorytmach Estratto/Espresso powtarzane, aż do uzyskania instancji o założonej z góry niezawodności lub do osiągnięcia określonej z góry liczby przebiegów algorytmu.

W przypadku zaprezentowanej wcześniej metody opartej na rozbudowanych wzorcach konstruowanych ręcznie, każda instancja relacji była wydobywana na podstawie jednego przykładu z korpusu. W przypadku Espresso/Estratto niezawodność każdej instancji jest oceniana na podstawie wszystkich ogólnych wzorców, które ją wydobyły z różnych przykładów. Miara niezawodności odwołuje się do statystycznej oceny siły powiązania pomiędzy wystąpieniami w korpusie miejsc dopasowanych do wzorca a wystąpieniami konkretnej instancji relacji (określonej pary lematów). Czyli każda instancja relacji jest wydobywana na podstawie wielu przesłanek, tj. wielu użyć danej pary lematów jako wyrażeń powiązanych określonymi konstrukcjami językowymi w korpusie. Metoda tego typu łączy w sobie elementy statystycznej oceny dystrybucji z automatyczną analizą sposobu powiązania wystąpień obu lematów w korpusie.

W przeprowadzonym eksperymencie algorytm Estratto został zainicjowany listą par hiperonimicznych pobranych ze Słowosieci i zastosowany do Korpusu IPI PAN. Wydobytych zostało automatycznie 25 361 par hiperonimicznych z dokładnością 41%. Dokładność została oceniona w sposób identyczny jak w przypadku wzorców ręcznych, tzn. została wylosowana reprezentatywna próbka i każda para była oceniona przez jednego z autorów pracy (Kurc, Piasecki, 2008) pod kątem reprezentacji synonimii lub bliskiej hiperonimii zgodnie z zasadmi przyjętymi w projekcie Słowosieci (Derwojedowa i in., 2007b). Uzyskany rezultat jest statystycznie znacząco lepszy niż rezultat uzyskany przez złożony wzorzec ręczny omówiony wcześniej. Wstępne rezultaty zastosowania Estratto do wydobywania par meronimicznych i przymiotnikowych par antonimicznych są obiecujące, mimo iż osiągnięta dokładność około 30% jest dużo gorsza od osiągniętej dla hiperonimii.

W tabeli 3 przedstawione zostały przykłady par lematów wydobytych za pomocą Estratto z połączonego korpusu. Algorytm został zainicjowany listą par hiperoni-

26 Instancje wydobywane przez dany wzorzec to te, których konteksty wystąpień w tekście odpowiadają danemu wzorcowi.

27 Tzn. dla danej instancji brane są pod uwagę wszystkie wzorce, które ją wydobywają z korpusu i ich niezawodność.

micznych pozyskanych z wcześniejszej wersji Słowosieci, tzw. jądra Słowosieci (Derwojedowa i in., 2007a).

Tabela 3. Przykłady par lematów wydobytych z Korpusu IPI PAN za pomocą algorytmu Estratto

Instancje początkowe		Instancje wydobyte	
senator	mówca	szkoła	instytucja
nazwa	oznaczenie	maszyna	urządzenie
Polska	kraj	wychowawca	pracownik
Polska	państwo	kombatant	osoba
wynagrodzenie	świadczenie	bank	instytucja
agencja	jednostka	pociąg	pojazd
akademia	uczelnia	telewizja	medium
alkohol	substancja	prasa	medium
pożar	zdarzenie	szpital	placówka
należność	zobowiązanie	czynsz	opłata
protokół	dokument	grunt	nieruchomość

Słabością metod opartych na rozbudowanych, szczegółowych wzorcach (zwykle konstruowanych ręcznie, ze względu na stopień ich skomplikowania) jest wydobywanie instancji relacji z pojedynczych kontekstów wspólnych wystąpień danej pary lematów w korpusie. Każde takie wystąpienie może być akcydentalne. Z drugiej strony bogata siła ekspresji wzorców umożliwia pełniejsze wykorzystanie informacji zawartej w strukturach językowych i dla pewnej klasy tekstów, np. encyklopedii lub podręczników, metoda taka może przynieść bardzo dobre rezultaty. Podejmowane próby rozpoznawania w tekście definicji terminów i ich wydobywania były jednak również oparte w dużej mierze na wzorcach konstruowanych ręcznie (por. Przepiórkowski i in., 2007). Ponadto ta silna zależność od każdego wystąpienia dopasowanego do wzorca może być zaletą z punktu widzenia lingwisty. Możemy bowiem dla każdej wydobytej pary jednoznacznie wskazać konteksty w korpusie, które są jej źródłem, *de facto* wzorce konstruowane ręcznie generują rodzaj konkordancji w oparciu o język wyszukiwania o dużej sile ekspresji. Można stosować wzorce ręczne do wyszukiwania i analizy kontekstowej par potencjalnie hiperonimicznych, które zostały wydobyte z korpusu, a nie występują w tezaurusie. W przypadku par lematów wydobywanych przez ogólne wzorce pozyskiwane automatycznie sytuacja jest nieco odmienna. Wydobycie każdej pary jest wynikiem analizy wielu kontekstów jej wystąpienia, gdzie każdy z kontekstów w różnym stopniu przyczynia się do decyzji podjętej przez algorytm. Mimo iż możemy odnaleźć wszystkie rozpatrywane konteksty w korpusie, ich interpretacja nie jest już tak bezpośrednia dla człowieka.

6. Semantyka dystrybucyjna

Zgodnie z interpretacją hipotezy dystrybucyjnej przedstawioną w sekcji 3.2, porównanie statystycznych modeli dystrybucji określonych lematów może stanowić podstawę do oceny tego, jak blisko są z sobą powiązane znaczenia obydwu lematów. W większości znanych metod semantyki dystrybucyjnej model dystrybucji jest bu-

dowany dla lematu na podstawie wszystkich jego wystąpień w korpusie (np. Widdows, 2004). Nie są więc rozróżniane wystąpienia odpowiadające różnym jednostkom leksykalnym reprezentowanym przez lemat, który jest polisemiczny. Nie są też rozróżniane homonimy, jeżeli rozróżnienie takie nie jest możliwe na poziomie charakterystyki morfo-syntaktycznej. Model dystrybucji polisemicznego lematu jest wypadkową wpływu znaczeń poszczególnych jednostek leksykalnych. Niestety, wpływ najczęstszych jednostek w większości metod dystrybucyjnych jest dominujący.

6.1. Miara semantycznego powiązania

Podstawowym rezultatem zastosowania metody dystrybucyjnej jest miara powiązania znaczeniowego (dalej MPZ). MPZ jest funkcją, która dla pary lematów zwraca liczbę opisującą siłę związku semantycznego pomiędzy nimi, tzn.:

$$MPZ: L \times L \rightarrow R,$$

gdzie L jest zbiorem lematów, a R jest zbiorem liczb rzeczywistych.

MPZ jest często nazywana miarą podobieństwa znaczeniowego, jednak pary lematów otrzymujące wysoką wartość miary w większości metod nie zawsze są blisko siebie zlokalizowane w ramach hierarchii hiperonimii. Miara zwraca wysokie wartości dla par lematów powiązanych różnorodnymi relacjami semantycznymi. Na przykład, dla lematu *komputer* MPZ wygenerowana na podstawie połączonego korpusu (sekcja 5) i algorytmu GRWF (Broda i in., 2009; Piasecki i in., 2009c) zwraca jako lematy z wysoką wartością miary między innymi: *drukarka* (wartość miary: 0,187) — powiązanie sytuacyjne, łączne użycie, *maszyna* (0,177) — daleki hiperonim, *procesor* (0,167) — meronim czy też *terminal* (0,142) — hiponim (por. Piasecki, Broda, 2009). Mając na uwadze powyższe, bardziej ogólne pojęcie powiązania semantycznego wydaje się być bardziej zasadne (por. Mohammad, Hirst, 2006).

Zaproponowano wiele metod wydobywania MPZ, jednak w większość z nich można odnaleźć mniej lub bardziej wyraźnie wydzielone cztery podstawowe etapy:

1. Przetwarzanie wstępne korpusu — realizowane typowo na poziomie analizy morfologicznej i płytkiej analizy składniowej (por. Piasecki, 2008).
2. Konstrukcja macierzy koincydencji — w której wiersze odpowiadają opisywanym lematom, a kolumny cechom (por. sekcja 3.2); każda komórka $M[x_i, c_j]$ przechowuje częstość wystąpień lematu x_i w kontekście opisywanym cechą c_j[28] (lub kontekście opisywanym parą cecha–wartość).
3. Transformacja macierzy — potencjalna redukcja rozmiaru macierzy[29] i/lub kombinacja transformacji wartości komórek (tzw. ważenia wartości cech względem

[28] Na przykład w wyniku przetwarzania Korpusu IPI PAN, lemat x_i =*abonent* wystąpił w kontekście opisywanym cechą c_j=`modyfikowany przez telefoniczny` $\mathbf{M}[x_i, c_j]$=85 razy. Częstość wystąpień została ustalona na podstawie zastosowania skonstruowanego ograniczenia w języku JOSKIPI — ograniczenie to ma wysoką dokładność (por. Piasecki, Radziszewski, 2009), ale w niektórych wypadkach rozpoznania są błędne, pewna liczba wystąpień przypadku modyfikacji może też być pominięta.

[29] Redukcji dokonuje się za pomocą matematycznego przekształcenia zastępującego dotychczasowe cechy nowym zbiorem o mniejszej liczbie cech, które w pewien sposób kondensują informację, np.

lematów, tak aby wartość cechy odzwierciedlała jej istotność dla opisu danego lematu) i selekcji[30] cech.

4. Obliczenie semantycznego powiązania — opisy lematów (wiersze transformowanej macierzy) są porównywane poprzez zastosowanie przyjętej miary podobieństwa wierszowych wektorów.

Podczas etapu transformacji macierz koincydencji jest poddawana różnorodnym przekształceniom matematycznym mającym na celu uwypuklić istotne prawidłowości w zgromadzonych danych, a zniwelować szum przypadkowych częstości. Stosowane są miary wywodzące się ze statystyki, teorii informacji lub przekształcenia skonstruowane eksperymentalnie (por. Piasecki i in., 2009c, rozdz. 3.3.4).

Celem realizowanym w ramach metod opartych na statystyce jest ocena tego, jak silnie wystąpienia cechy są skorelowane z występowaniem poszczególnych lematów. Ocena ta jest wyrażona wartością liczbową dla danej cechy i danego lematu i zostaje użyta jako wartość cechy zamiast pierwotnej częstości.

Z kolei w ramach metod opartych na teorii informacji wylicza się nową wartość cechy jako wartość liczbową opisującą, ile informacji o poszczególnych lematach wnosi wystąpienie cechy.

Wyliczona nowa wartość cechy jest w większości podejść do transformacji zależna pośrednio od częstości występowania cechy i poszczególnych lematów. Poziom wartości pewnych cech może być dla niektórych lematów sztucznie zawyżony poprzez przypadkowo podwyższone częstości ich współwystępowania (np. poprzez często powtarzaną frazę w korpusie). Wady tej pozbawione jest podejście do transformacji nazwane uogólnionym ważeniem rangowym (ang. *generalised rank based weighting*) (Broda i in., 2009; Piasecki i in., 2009c) (dalej GRWF), w którym po zastosowaniu dowolnej funkcji transformacji wybiera się dla danego lematu k najlepszych cech (np. 1000) według ich transformowanych wartości i następnie przypisuje się im, jako nowe wartości, liczby naturalne odpowiadające ich pozycjom w rankingu. Najlepsza cecha otrzymuje wartość równą k, następna $k - 1$ itd. Przekształcenie to opiera się na założeniu, że dla człowieka przy porównywaniu dwóch lematów pod względem ich znaczenia nie są istotne subtelne oceny wartości cech, a jedynie lista cech charakterystycznych dla obu lematów, np. lista przymiotników, którymi możemy określać dwa rzeczownikowe lematy, lista uporządkowana od najbardziej do najmniej charakterystycznych. Proste przekształcenie GRWF przyniosło znaczącą poprawę jakości wydobywanych z jego pomocą MPZ (por. Broda i in., 2009; Piasecki i in., 2009c).

W macierzy koincydencji dystrybucja każdego lematu jest reprezentowana jako wektor liczb (sekwencja liczb) w postaci wiersza macierzy. Do porównania dystrybucji dwóch lematów można zastosować dowolną miarę podobieństwa wektorów. Bardzo często z dobrym skutkiem stosowana jest tzw. miara kosinusowa, która polega na obliczeniu kosinusa kąta, jaki tworzą dwa wektory, w tym przypadku wiersze macierzy odpowiadające dwóm lematom, w wielowymiarowej przestrzeni (o liczbie wy-

metoda zastosowana w technice *Latent Semantic Analysis* (analizie semantyki ukrytej) (Landauer, Dumais, 1997).

30 Selekcja polega na odrzuceniu cech, które nie wnoszą istotnej informacji o opisywanych lematach lub na wyborze tych, które wnoszą najwięcej informacji, np. lemat *abonent* w Korpusie IPI PAN był modyfikowany 128 razy przez przymiotnika *nowy*, który występuje często z tak wieloma lematami, że wnosi niewiele do ich opisu.

miarów określonej przez liczbę cech). Otrzymana w wyniku obliczenia wartość jest traktowana jako miara powiązania znaczeniowego między oboma lematami.

Mohammad i Hirst (2006) zauważyli, że w zależności od typu użytych cech możemy otrzymać miarę zbliżoną do podobieństwa semantycznego lub wskazującą na bardziej ogólne powiązania semantyczne. Semantyczne powiązanie otrzymuje się w oparciu o cechy definiowane jako wystąpienie lematu w określonym dokumencie (typ 1 na stronie 152) lub współwystąpienie opisywanego lematu z innym lematem w określonym fragmencie tekstu (typ 2 na stronie 152).

Możemy zaobserwować w działaniu miary tendencję do wyrażania podobieństwa semantycznego wtedy, gdy użyjemy cech typu 3 ze strony 152, tzn. cech zdefiniowanych w oparciu o leksykalno-syntaktyczne relacje wewnątrz kontekstu, np.:

x występuje jako `podmiot(`*`określony czasownik`*`)`
lub *x* jest `modyfikowany _ przez(`*`określony przymiotnik`*`)`.

W naszych eksperymentach przeprowadzonych na Korpusie IPI PAN (por. Piasecki, Broda, 2007) można było zaobserwować, że semantyczne powiązanie obejmuje szeroki zakres związków semantycznych pomiędzy lematami wynikających często ze współwystępowania obu lematów w opisie tej samej sytuacji. Zgodnie z badaniami Mohammada i Hirsta (2006) pary lematów otrzymujących wysokie wartości miary podobieństwa semantycznego powinny reprezentować instancję jednej z semantycznych relacji leksykalnych opisywanych w tezaurusach, np. synonimii, hiperonimii lub meronimii. Automatyczne wydobycie miary podobieństwa semantycznego z korpusu jest jednak bardzo trudne, jeżeli w ogóle możliwe, i we wszystkich przypadkach znanych z literatury mamy do czynienia z formami pośrednimi miar, tzn. dla podanego lematu najwyższe wartości miary uzyskują zarówno lematy powiązane leksykalnymi relacjami semantycznymi, jak i lematy powiązane w bardziej ogólny, mniej regularny sposób. Ponadto, w każdej MPZ pomiędzy lematami o najwyższej wartości miary dla danego lematu można odnaleźć lematy powiązane jedną z semantycznych relacji leksykalnych. Na charakter miary można wpływać, dobierając odpowiedni zestaw cech, tj. łącząc cechy oparte na współwystępowaniu lematów w szerszym kontekście (typy 1 i 2 na stronie 152) z cechami wykorzystującymi relacje leksykalno-syntaktyczne.

Miara semantycznego powiązania może być wykorzystana do grupowania lematów w pola semantyczne, aby przejść od semantycznego powiązania do automatycznej identyfikacji synonimii, a nawet hiperonimii (por. Broda i in., 2008b).

6.2. Eksperymenty i rezultaty

W tabeli 4 zaprezentowany został przykład trzech list 20 lematów najbardziej semantycznie powiązanych z zadanym lematem (wytłuszczonym w tabeli). Wyniki uzyskane dla kilkunastu tysięcy lematów są dostępne na stronie WWW http://plwordnet.pwr.wroc.pl/browser/automatic.jsp (Piasecki, Broda, 2009). Listy zostały wygenerowane przez MPZ opartą na algorytmie transformacji GRWF (omówionym w poprzedniej sekcji) i opisującą 13 285 nominalnych lematów ze zbioru utworzonego na potrzeby eksperymentów z ręcznymi wzorcami leksykalno-morfo-syntaktycznymi (por. sekcja 5) (zbiór ten był podstawą do półautomatycznego rozszerzania wstępnej wersji

Słowosieci). Zastosowana MPZ została wydobyta z połączonego korpusu (por. sekcja 5) za pomocą następującego zestawu cech opartych na relacjach leksykalno-morfosyntaktycznych:

— modyfikacja opisywanego lematu *x* przez *określony przymiotnik* lub *określony imiesłów przymiotnikowy* (41 619 cech, wartością cechy jest liczba odnalezionych przypadków modyfikacji),
— koordynacja (współrzędne złożenie) *x* z *określonym rzeczownikiem* (115 604 cech, wartości jw.),
— modyfikacja *x* przez *określony rzeczownik* w przypadku dopełniacza (115 604),
— wystąpienie *określonego czasownika*, dla którego opisywany lemat może pełnić rolę podmiotu w danym kontekście zdaniowym (19 665).

Tabela 4. Lista 20 najbardziej semantycznie powiązanych lematów z wyróżnionym lematem według miary powiązania znaczeniowego opartej na algorytmie GRWF (por. Piasecki i in., 2009c).

gaz ziemny		samochód		rozwaga	
gaz	0,258	pojazd	0,368	rozsądek	0,253
węgiel kamienny	0,207	auto	0,350	roztropność	0,222
węgiel brunatny	0,197	ciężarówka	0,297	ostrożność	0,212
ropa	0,193	wóz	0,290	umiar	0,197
olej opałowy	0,164	autobus	0,262	cierpliwość	0,191
paliwo	0,161	furgonetka	0,230	wnikliwość	0,186
wodór	0,160	limuzyna	0,220	staranność	0,175
kopalina	0,160	taksówka	0,218	odwaga	0,168
węgiel	0,143	autokar	0,216	wrażliwość	0,167
olej napędowy	0,140	motocykl	0,206	powściągliwość	0,166
gaz płynny	0,140	radiowóz	0,203	wyczucie	0,166
koks	0,127	ciągnik	0,196	dyskrecja	0,163
ołów	0,119	pociąg	0,191	obiektywizm	0,162
azot	0,119	karetka	0,185	wytrwałość	0,161
tlen	0,116	tir	0,185	pieczołowitość	0,158
uran	0,116	samolot	0,184	pracowitość	0,157
biokomponent	0,115	półciężarówka	0,175	stanowczość	0,157
cynk	0,114	mikrobus	0,175	mądrość	0,154
łupek palny	0,113	rower	0,172	uczciwość	0,153
benzyna	0,110	opel	0,170	umiarkowanie	0,153

Podane powyżej liczby odnoszą się do cech potencjalnie istniejących, tzn. opisują liczbę lematów określonej klasy, które mogły wystąpić jako elementy zleksykalizowanej cechy. Jako że nie zaobserwowano wystąpień wielu potencjalnie możliwych związków, jedynie 167 834 aktywnych cech zostało wykorzystanych do opisu lematów nominalnych[31]. Podana liczba opisuje ponadto stan po transformacji wartości cech i na-

31 Część potencjalnych cech mogła też nie zostać automatycznie zaobserwowana, ponieważ ograniczenia morfo-syntaktyczne zastosowane do zdefiniowania cech były napisane z myślą o maksymalizacji dokładności rozpoznania, kosztem nawet zmniejszonej kompletności, rozumianej jako stosunek liczby odnalezionych wystąpień cech do wszystkich rzeczywiście istniejących w korpusie.

stępującej później ich selekcji — część cech pominięto ze względu na zbyt małą liczbę wystąpień w korpusie, a co za tym idzie, zbyt małą wiarygodność statystyczną.

Analizując listę wygenerowaną dla lematu *gaz ziemny* przedstawioną w tabeli 4, możemy zauważyć, że zawiera ona jego hiperonimy, np. *gaz* i *kopalina*; kohiponimy, np. *węgiel kamienny* i *ropa*; luźno powiązanych „kuzynów" z szerzej rozpatrywanej hiperonimicznej struktury zawierającej *gaz ziemny*, np. *azot* i *cynk* oraz również lematy podobne do rozważanego ze względu na podobieństwo zastosowania odpowiednich substancji, np. *biokomponent* (jako dodatek do paliwa samochodowego), podczas gdy *gaz ziemny* może być samodzielnie użyty jako paliwo.

6.3. Problem oceny

Ocena jakości wydobytej MPZ jest notorycznym problem dla wszystkich istniejących podejść do jej wydobywania. Manualna ocena może być bardzo zwodnicza, ponieważ zawsze można znaleźć dobre i złe przykłady na liście lematów najbardziej powiązanych znaczeniowo z danym. Co gorsza, proste obliczenie pewnego rodzaju dokładności na podstawie analizy lematów pojawiających się na takiej liście pod kątem reprezentowania leksykalnych relacji semantycznych daje tylko częściowy obraz. MPZ jest funkcją przypisującą siłę semantycznego powiązania do każdej pary lematów. Przypisanie wyższych wartości do par reprezentujących pewne relacje semantyczne jest oczekiwane, ale nie wyczerpuje potencjału MPZ. Ręczna ocena wartości przypisywanych przez MPZ w praktyce nie jest możliwa do przeprowadzenia. Człowiek nie jest w stanie rozstrzygnąć czy dana wartość powinna być jedną setną mniejsza czy też większa. Nie istnieją ręcznie zbudowane zasoby językowe (zbiory danych), z którymi można by było porównać MPZ.

Spośród kilku potencjalnych metod oceny MPZ (por. Zesch, Gurevych, 2006; Piasecki i in., 2007) zastosowanie MPZ jako źródła wiedzy przy automatycznym rozwiązywaniu testu synonimii przyniosło wartościowe rezultaty w szeregu eksperymentów. Test synonimii jest rodzajem testu znajomości znaczeń leksykalnych oryginalnie używanym w odniesieniu do ludzi uczących się języka obcego. Przykładem może być jedno z zadań w obrębie testu *Test of English as a Foreign Language* (dalej TOEFL), stosowanego do oceny znajomości języka angielskiego. W tego typu zadaniu osoba rozwiązująca test ma rozpoznać wśród czterech słów odpowiedzi te, które jest synonimiczne z podanym słowem problemowym. TOEFL został wykorzystany po raz pierwszy przez Landauer i Dumais (1997) do oceny MPZ wygenerowanej za pomocą zaproponowanego przez nich algorytmu *Latent Semantic Analysis* (pol. analizy semantyki ukrytej) (dalej LSA, MPZ wygenerowaną za pomocą LSA będziemy oznaczać MPZ_{LSA}). MPZ_{LSA} została zastosowana do określenia siły powiązania pomiędzy słowem problemowym a każdą z czterech możliwych odpowiedzi. Jako właściwa wybierana była ta, dla której MPZ_{LSA} zwracała najwyższą wartość. TOEFL zawiera ograniczoną liczbę pytań i został już praktycznie rozwiązany metodami automatycznymi, np. Turney i in. (2003) osiągnęli dokładność 97,5%, budując MPZ w oparciu o dużą ilość danych tekstowych zebranych z Internetu. Jednak Freitag i in. (2005) zaproponowali nową wersję testu synonimii generowanego automatycznie na wzór TOEFL, na podstawie Princeton WordNetu. W teście tym nazwanym *WordNet-Based Synonymy Test* (test synonimii oparty na wordnecie) słowa prob-

lemowe wybierane są z wordnetu, prawidłowe odpowiedzi ze zbiorów synonimów (tzw. synsetów), do których należą słowa problemowe, a odpowiedzi niepoprawne są wybierane losowo spoza zbiorów synonimów słowa problemowego. Test generowany automatycznie może składać się nawet z kilkunastu tysięcy pytań i jest o wiele trudniejszy niż TOEFL, np. najlepszy wynik osiągnięty przez MPZ dla lematów nominalnych w pracy Freitag i in. (2005) to 75,8%.

Rodzina testów synonimii wzorowanych na pracy Freitag i in. (2005) dla języka polskiego, generowanych w oparciu o Słowosieć, została zaprezentowana między innymi w pracach Piaseckiego, Szpakowicza i Brody (2007; 2009c). Przykładowe pytania wygenerowane ze Słowosieci 1.0 w podstawowym wariancie testu zaprezentowane są poniżej:

— słowo problemowe: *diabeł*
  odpowiedzi: biolog, gęganie, *szatan*, wydech
— słowo problemowe: *pojazd kosmiczny*
  odpowiedzi: gałąź, *prom kosmiczny*, regulamin, znak rozpoznawczy

Omawiana wcześniej MPZ, oparta na algorytmie GRWF, wygenerowana na podstawie połączonego korpusu, osiągnęła dokładność 88,14% dla nominalnych lematów. Test zawierał 9486 pytań. Ta sama MPZ testowana jedynie dla lematów pojawiających się częściej niż 1000 razy w połączonym korpusie osiągnęła dokładność 92,28% (por. Piasecki i in., 2009c). Bardzo zbliżony test (wygenerowany na podstawie starszej wersji Słowosieci, dla wszystkich lematów, nie tylko częstych) został zastosowany w eksperymencie z udziałem grupy 29 studentów informatyki. Ich zadaniem było wskazanie wśród słów odpowiedzi słowa najbardziej zbliżonego znaczeniowo do słowa problemowego. Badani osiągnęli średnią dokładność 86,64% (maksymalna 96,24%). Oznacza to, że MPZ_{GRWF} wygenerowana na podstawie wielkiego korpusu tekstowego wykazuje zdolność do odróżniania bliskich i dalekich powiązań semantycznych pomiędzy lematami na poziomie człowieka, który nie ma wykształcenia w dziedzinie lingwistyki.

7. Metody wielokryterialne wydobywania relacji

Metody oparte na wzorcach oraz semantyce dystrybucyjnej są zasadniczo odmienne w przesłankach, na których się opierają, jak i w charakterze wydobywanych powiązań. Żadna z metod nie zapewnia również dokładności zbliżonej do 100%. Różnica co do charakteru wykorzystywanych przesłanek stwarza szansę, że błędy popełniane przez metody różnych typów nie będą się powielać. Z kolei różnica w charakterze wydobywanych danych daje szansę na opis relacji pomiędzy tymi samymi lematami z różnych perspektyw. Zarysowana powyżej idea łączenia metod o różnym charakterze, szeroko rozpowszechniona w sztucznej inteligencji, została zrealizowana w przynajmniej kilku podejściach do wydobywania relacji semantycznych. Możemy je podzielić na dwie główne grupy:

— metody klasyfikacji par lematów (np. Snowi in., 2005; Piasecki i in., 2008),
— wielokryterialne metody rozszerzania wordnetu (np. Caraballo, 1999; Snow i in., 2006; Piasecki i in., 2009d, b).

W przypadku metod pierwszej grupy celem jest skonstruowanie algorytmu, nazywanego klasyfikatorem, który dla podanej pary lematów na wejściu rozstrzygnie,

czy reprezentuje ona instancję jednej z określonych wcześniej semantycznych relacji leksykalnych. Decyzje podejmowane przez algorytm dzielą zbiór par lematów podawanych na wejściu na dwie klasy (podzbiory): instancji interesujących nas relacji i pozostałych par.

Snow, Jurafsky i Ng (2005) do konstrukcji klasyfikatora par lematów dzielącego je na bliskie hiperonimy i pozostałe zastosowali metodę maszynowego uczenia się na podstawie znanych przykładów par bliskich hiperonimów. Przykłady zostały pobrane automatycznie z Princeton WordNetu. Wygenerowano też przykłady negatywne par lematów, które nie są połączone relacją bliskiej hiperonimii. Korpus tekstów w języku angielskim został przetworzony relatywnie prostym parserem zależnościowym MiniPar (Lin, 1993). Następnie zidentyfikowane zostały wszystkie wystąpienia par przykładowych (zarówno pozytywnych, jak i negatywnych), w których elementy pary są powiązane ścieżką zależności syntaktycznych. Zbiory wydobytych ścieżek zależności dla przykładów pozytywnych i negatywnych zostały dostarczone do algorytmu maszynowego uczenia się, który wygenerował probabilistyczne reguły sterujące pracą klasyfikatora.

Podejście z pracy Snowa, Jurafsky'ego i Ng (2005) jest bliskie w swojej istocie algorytmowi Estratto (por. sekcja 5) — wzorce leksykalno-syntaktyczne opisujące konteksty współwystępowania par hiperonimicznych i niehiperonimicznych są łączone z informacją statystyczną, aby utworzyć reguły wyznaczające zbiór par hiperonimicznych.

W podejściu zaprezentowanym w pracy Piaseckiego, Szpakowicza, Marcińczuka i Brody (2008), podobnie jak u Snowa, Jurafsky'ego i Ng (2005), przykładowe pary zostały pobrane z wordnetu, tym razem ze Słowosieci. Ze względu na planowane zastosowanie klasyfikatora w ramach procesu rozszerzania Słowosieci, do klasy pozytywnych przykładów zostały włączone zarówno przypadki bliskiej hiperonimii i synonimii, jak i przypadki bezpośredniej meronimii. W rezultacie klasa pozytywnych przykładów obejmowała semantyczne relacje leksykalne wyznaczające bezpośrednio pozycję jednostki leksykalnej w ramach nominalnej części wordnetu. Do klasy negatywnych przykładów włączone zostały zarówno pary niepowiązane żadna relacją, jak i pary powiązane daleką hiperonimią. W odróżnieniu od Snowa, Jurafsky'ego i Ng (2005) wykorzystano 17 różnorodnych cech do opisu przykładów uczących na potrzeby maszynowego uczenia się klasyfikatora. Zbiór cech obejmował np. wartość MPZ pomiędzy lematami pary, częstość powiązań obu lematów za pomocą syntaktycznej koordynacji i modyfikacji dopełniaczowej, zdolność do opisu jednego lematu poprzez cechy leksykalno-syntaktyczne opisujące lemat drugi (z rozbiciem na typy cech), liczbę przymiotników powiązanych w sposób istotnie statystyczny z jednym i drugim lematem oraz współwystępowanie obu lematów w szerszym oknie tekstowym. Zastosowanych zostało kilka algorytmów maszynowego uczenia się, uzyskując zbliżone rezultaty (por. Piasecki i in., 2008; Piasecki i in., 2009c).

Eksperymenty zostały przeprowadzone dla tej samej listy 13 285 nominalnych lematów, która była podstawą do eksperymentów z ręcznie konstruowanymi wzorcami (por. sekcja 5). Przykłady pozytywne i negatywne generowane były na podstawie tzw. jądra Słowosieci, czyli wstępnej wersji z czerwca 2008 roku. Wartości 17 cech opisujących przykłady uczące były wydobywane z połączonego korpusu (por. sekcja 5). Początkowo przyjęty został stosunek liczby przykładów pozytywnych do negatywnych 1:1. Przy takim doborze danych uczących osiągnięte zostały bardzo do-

bre wyniki działania klasyfikatora w testach na wybranej losowo części przykładów, które nie były wykorzystane w trakcie uczenia. Dla najlepszego klasyfikatora dokładność[32] wyniosła 81,7%, natomiast kompletność[33] 78,4%. Ponieważ przykłady testowe, różne od treningowych, pochodziły ze Słowosieci, to ocena była przeprowadzana w odniesieniu do opisu relacji zawartego w niej. Niestety, tak dobre wyniki testowe nie przekładały się w pełni na praktyczne działanie klasyfikatora. W praktyce mamy zawsze o wiele więcej par negatywnych niż pozytywnych, np. jeżeli stosujemy klasyfikator do par lematów osiągających wysokie wartości MPZ w celu wydzielenia tych, które pozostają w jednej z wordnetowych relacji semantycznych. Klasyfikator trenowany na danych testowych o podziale 1:1 wykazywał w praktyce silną tendencję do klasyfikacji pozytywnych par lematów. Dlatego też powtórzono eksperymenty z podziałem danych uczących 1:10 na korzyść przykładów negatywnych. Osiągnięta dokładność 60,7% oraz kompletność 39,9% była dużo gorsza niż w przypadku podziału 1:1, ale analiza działania klasyfikatora w praktyce pokazała, że tym razem jego ocena testowa była bliższa rzeczywistej jakości w praktycznym działaniu. Klasyfikator taki nie rozwiązuje problemu wydzielenia z par lematów o wysokiej wartości MPZ tych, które są instancjami relacji wordnetowej, ale mimo to dostarcza cennej informacji, wydzielając część par jako będące prawdopodobnie w relacji. Jak zobaczymy dalej, wiedza dostarczona przez taki klasyfikator jest bardzo pomocna do budowy wielokryterialnego algorytmu rozszerzania istniejącego wordnetu.

Przykłady par lematów zaklasyfikowanych przez klasyfikator jako pozytywne i negatywne są przedstawione poniżej (wszystkie pary pochodzą z danych testowych wygenerowanych na podstawie Słowosieci):

— pary, które zostały poprawnie[34] zaklasyfikowane jako instancje bliskiej hiperonimii lub synonimii: *akt–ustawa*, *bank–firma*, *emocja–smutek*, *intelekt–przymiot*, *licencja–zezwolenie*, *pragnienie–ochota*, *terytorium–kolonia*, *warzywo–kartofel*,
— pary, które zostały niepoprawnie zaklasyfikowane jako instancje bliskiej hiperonimii lub synonimii: *celnik–policja*, *czynsz–oprocentowanie*, *dochód–dotacja*, *nonszalancja–rozrzutność*, *odpad–produkt*, *problem–rodzina*, *temat–dostarczyciel*, *zachwyt–zdumienie*,
— pary, które zostały prawidłowo odrzucone jako nienależące do rozpatrywanych relacji: *człowieczeństwo–prorok*, *licencja–zarządzenie*, *opis–hipoteza*, *ślub–kochanek*, *tempo–sport*, *trybunał–sejm*,
— pary, które zostały nieprawidłowo odrzucone jako nienależące do rozpatrywanych relacji: *linia–ogonek*, *konstrukcja–twierdza*, *nieprzychylność–emocja* (w sensie uczucia), *podpora–kula*, *zakochanie–emocja*.

Poza podejściami opartymi na konstrukcji klasyfikatora zbiory różnorodnych przesłanek (kryteriów) są wykorzystywane w programach rozszerzających automatycznie istniejący wordnet. Czytelnym przykładem takich rozwiązań jest praca Caraballo (1999). Autorka najpierw zbudowała za pomocą automatycznej metody grupowania

32 Dokładność to stosunek liczby par poprawnie zaklasyfikowanych jako pozytywne do liczby wszystkich decyzji na tak.

33 Kompletność to stosunek liczb par zaklasyfikowanych jako pozytywne do wszystkich par pozytywnych występujących w danych testowych.

34 Poprawnie w stosunku do stanu, jaki jest w Słowosieci, tzn. każda taka para jest albo bezpośrednio parą synonimiczną lub hiperonimiczną, albo występują tylko 1–2 jednostki leksykalne pośredniczące w łańcuchu hiperonimicznym je łączącym.

hierarchię grup lematów w oparciu o skonstruowaną wcześniej MPZ. Grupy zawierające lematy o bardziej szczegółowym zakresie znalazły się na dole hierarchii, natomiast grupy bardziej ogólne w wyższych obszarach hierarchii. Grupy otrzymywane w sposób automatyczny bardzo często nie zapewniają czytelnego, z punktu widzenia człowieka, kryterium przynależności do grupy ani zależności pomiędzy grupami. Dlatego Caraballo poddała uzyskaną hierarchię grup dalszej interpretacji w oparciu o pary lematów wydobyte za pomocą wzorców leksykalno-syntaktycznych Hearst (1992), ukierunkowanych na wydobycie par hiperonimicznych (por. sekcja 5). Na podstawie informacji uzyskanej ze wzorców i struktury hierarchii grup zidentyfikowane zostały pary hiperonimicznych lematów. Ręczna ocena uzyskanych wyników dla niewielkiej próbki pokazała dokładność wyznaczenia par hiperonimicznych na poziomie 33%, co jest wynikiem porównywalnym z uzyskanym przez nas za pomocą wzorców (por. sekcja 5). Caraballo jednak akceptowała podczas oceny jedynie bliższe powiązania hiperonimiczne. Wadą podejścia Caraballo jest silna zależność od dokładności wyznaczenia par hiperonimicznych za pomocą wzorców. Te z kolei wymagają zastosowania dobrej jakości parsera.

Snow, Jurafsky i Ng (2006) połączyli dwa różne źródła wiedzy o charakterze probabilistycznym, opisujące prawdopodobieństwo wystąpienia określonych relacji pomiędzy parą lematów:

— swój własny algorytm klasyfikacji (Snow i in., 2005), generujący decyzję w postaci prawdopodobieństwa pozostawania w relacji hiperonimii przez parę lematów (podejście to zostało omówione wcześniej w tej sekcji),
— oraz nowo skonstruowany algorytm, określający dla pary lematów prawdopodobieństwo pozostawania w relacji kohiponimii w określonej odległości od wspólnego przodka hiperonimicznego.

Probabilistyczne przesłanki generowane przez obydwa algorytmy zostały połączone w jeden łączny model wyznaczający najbardziej prawdopodobny kształt nowego wordnetu rozszerzonego o nowe lematy. Dokładność działania algorytmu została oceniona ręcznie przez kilka osób oceniających, ale na bardzo małej próbce 100 par lematów jako instancji hiperonimii. Dopuszczano również dalekie powiązania hiperonimiczne. Uzyskano 84% dokładności przy 10 000 dodanych do wordnetu par. Niezależnie od pewnych słabości procesu oceniania jest to wynik bardzo dobry.

Podejście z pracy Snowa, Jurafsky'ego i Ng (2006) jest ograniczone jedynie do przesłanek o charakterze probabilistycznym, wydobytych z bardzo dużej ilości danych. W ramach tego podejścia nie ma, np. możliwości wykorzystania informacji uzyskiwanej przez zastosowanie ręcznie konstruowanych wzorców leksykalno-syntaktycznych (por. sekcja 5). Wady tej pozbawiony jest algorytm dołączania poprzez regiony aktywacji (dalej DRA) oraz oparty na nim system wspomagający rozszerzanie wordnetu o nazwie WordNet Weaver, przedstawione w pracach Piaseckiego i in. (2009d, b, c). W budowie algorytmu DRA przyjęto podstawowe założenie, że każda z metod wydobywania popełnia błędy i może powiązać dany lemat nominalny z niewłaściwym miejscem w strukturze hiperonimii. Jednak błędne powiązania wypadające blisko punktu optymalnego są bardziej prawdopodobne niż umiejscowiane w większej odległości. Założenie to oparte jest na analizie dokładności poszczególnych metod i na rodzajach popełnianych błędów, a szczególnie na braku zdolności wszystkich metod do rozróżnienia bezpośredniego powiązania hiperonimicznego od hiperonimii bliskiej, ale nie bezpośredniej.

Struktura hiperonimiczna łączy w Słowosieci (analogicznie jak w innych wordnetach) zbiory synonimów, tzw. synsety[35]. W algorytmie DRA rozpoznanie potencjalnych miejsc dołączenia nowego lematu *x* jest realizowane w czterech etapach:

1. Określane jest dopasowanie semantyczne *x* do każdego lematu *y* obecnego już w wordnecie.
2. Na podstawie dopasowania *x* do poszczególnych lematów ustalane jest dopasowanie *x* do każdego synsetu *y*:
   — brane jest pod uwagę dopasowanie *x* do lematów zawartych[36] w synsecie *y*,
   — oraz dopasowanie *x* do synsetów powiązanych z *y* relacją hiperonimii lub hiponimii (w ramach pewnej odległości liczonej jako liczba pośredniczących powiązań hiperonimicznych),
   — dopasowanie do synsetów w hiperonimicznym otoczeniu *y* zostaje uwzględnione z pewną wagą (mniejszą niż 1) w wyliczeniu łącznego dopasowania do *y* (waga ta zmniejsza się proporcjonalnie wraz z odległością liczoną po hiperonimicznych powiązaniach).
3. Na podstawie wyliczonego dopasowania do poszczególnych synsetów odnajdywane są w strukturze hiperonimii obszary (podgrafy), nazywane obszarami aktywacji, takie, że wszystkie synsety zawarte w danym obszarze są dopasowane do *x*. Dla każdego obszaru aktywacji zapamiętywana jest jako miara jego oceny maksymalna miara jakości dopasowania *x* do synsetów danego obszaru.
4. Zapamiętane oceny obszarów aktywacji dla danego nowego lematu *x* wykorzystywane są następnie do wybrania kilku takich obszarów, które zostają zaprezentowane lingwiście w ramach systemu WordNet Weaver jako proponowane miejsca dołączenia *x*. Regiony te będą dalej nazywane regionami dołączania[37].

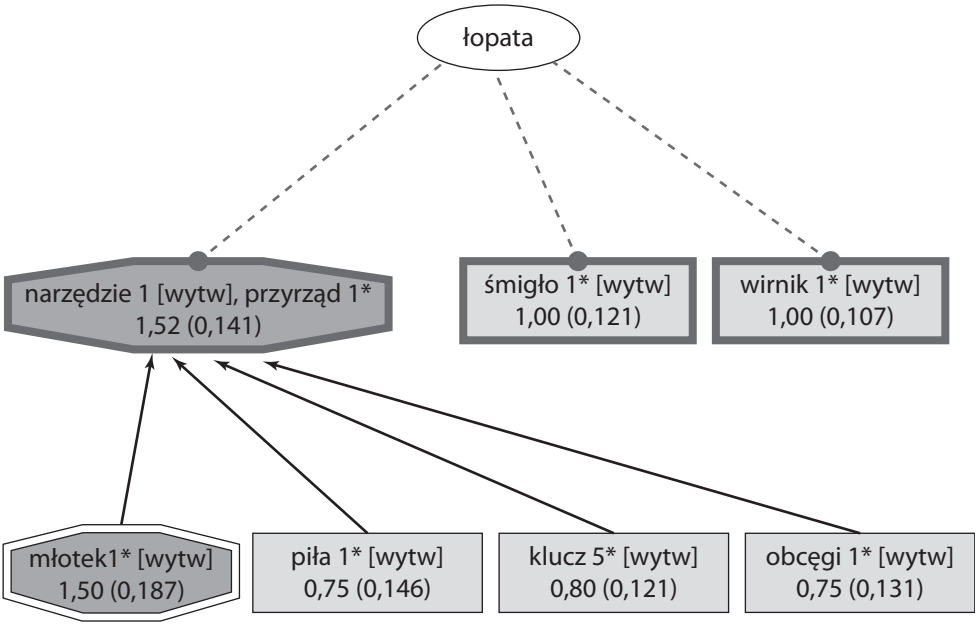
Dopasowanie semantyczne lematu do lematu (etap 1) jest wyznaczane na podstawie wszystkich dostępnych kryteriów, zgodnie z często stosowanym w sztucznej inteligencji schematem głosowania ważonego. Z każdym rodzajem przesłanki wiązany jest głos o określonej liczbowo sile. Dla każdej pary nowy lemat *x* i lemat *y* ze struktury wordnetu sumowane są głosy wynikające z poszczególnych przesłanek. Jako przesłanki wykorzystane zostały wyniki działania wszystkich metod automatycznego wydobywania relacji omówione wcześniej w tej pracy, tj. $\mathrm{MPZ}_{\mathrm{GRWF}}$, wzorce konstruowane ręcznie, klasyfikator par lematów oraz wyniki działania Estratto. Jeżeli sumaryczna liczba głosów przekroczy pewien założony próg, to zostaje stwierdzony fakt istnienia dopasowania. Dodatkowo dopasowania podzielone są na silne — wynikające z więcej niż jednej przesłanki oraz słabe — wynikające jedynie z dużej wartości MPZ dla danej pary lematów. Szczegółowo algorytm wyliczenia dopasowania oraz cały algorytm DRA został omówiony w pracy Piaseckiego i in., (2009c, rozdz. 4.5.3–

35 Synset może być również zbiorem jednoelementowym, co się często zdarza (por. Piasecki i in., 2009c). Relacja hiperonimii pomiędzy synsetami jest wywiedziona z relacji hiperonimii łączącej jednostki leksykalne należące do poszczególnych synsetów, tzn. jeżeli dwa synsety są połączone hiperonimią (na poziomie całych synsetów), to również ich elementy są odpowiednio parami w relacji hiperonimii.

36 Stosujemy tu dla uproszczenia opisu pewien skrót myślowy: do synsetu należy oczywiście określona jednostka leksykalna, która jest reprezentowana przez określony lemat.

37 Regiony dołączania można interpretować jako przybliżony opis jednostek leksykalnych, które zostały zaproponowane automatycznie dla nowego lematu *x*. Każdy region przybliża w pewien sposób własności semantyczne proponowanej jednostki leksykalnej.

4.5.4)Wyniki działania algorytmu DRA prezentowane są lingwiście w ramach systemu WordNet Weaver (Piasecki i in., 2009b, c), który stanowi rozszerzenie wcześniejszego edytora wordnetu (Piasecki i in., 2009c, rozdz. 2.4). Każdy sugerowany region dołączania jest prezentowany lingwiście w postaci osobnego podgrafu. Przykład takiego podgrafu został przedstawiony na rysunku 1. Na jednej ze podstron (Piasecki i in., 2009a) strony WWW projektu budowy Słowosieci można odnaleźć bogatą listę 3709 dalszych przykładów z wczesnego etapu rozszerzania Słowosieci (można tam też zobaczyć kolorową wersję prezentowanego tu przykładu). Owalny kształt reprezentuje nowy lemat, natomiast prostokąty i ośmiokąty to węzły grafu hiperonimii, które reprezentują synsety. Połączenia pomiędzy synsetami reprezentują instancje hiperonimii (hiperonim wskazuje strzałka). Połączenia pomiędzy nowym lematem (owal) a synsetami wskazują na centralne synsety (o najwyższym dopasowaniu) poszczególnych regionów dołączania. Dodatkowo każdy synset centralny jest wyróżniony pogrubioną niebieską ramką (tu szarą). Początkowo pokazane są tylko elementy regionów dołączania — dalsze elementy struktury hiperonimii mogą zostać rozwinięte/zwinięte poprzez przyciski ze znakiem trójkąta. Kształt węzła wskazuje na rodzaj dopasowania nowego lematu do niego: ośmiokąt oznacza dopasowanie silne, prostokąt słabe. Dodatkowo kolor węzła reprezentuje liczbową miarę oceny dopasowania: najniższe wartości to jasny szary, wyższe — ciemny żółty do czerwonego i fioletu (tu od jasnej szarości do ciemnej).



Rysunek 1. Regiony dołączenia nowego lematu *łopata* sugerowane przez system WordNet Weaver względem wersji Słowosieci z czerwca 2008 roku

Lingwista może zatwierdzić proponowane dołączenie, klikając w synset centralny i wybierając rodzaj relacji: synonimia, hiperonimia lub hiponimia, ale może wybrać

też dowolny inny synset, przechodząc swobodnie po strukturze. Dodatkowo może przełączyć się do innych ekranów edytora wordnetu, dołączając nowy synset w dowolne miejsce za pomocą dowolnej relacji. Ponadto lingwiści oznaczali wszystkie błędne sugestie algorytmu. Pozwoliło to na zebranie wyczerpujących danych o jakości sugestii prezentowanych w postaci regionów dołączania generowanych przez algorytm DRA. Szczegółowe wyniki oceny zostały przedstawione w pracach Piaseckiego i in. (2009b, c), tutaj ograniczymy się jedynie do zbiorczego podsumowania całego wielomiesięcznego procesu rozszerzania Słowosieci przy zastosowaniu algorytmu DRA i systemu WordNet Weaver.

Dokładność algorytmu DRA została zmierzona względem decyzji podjętych przez lingwistów w ciągu całego procesu rozszerzania jądra Słowosieci do Słowosieci w wersji 1.0 (charakterystyka ilościowa tego procesu została przedstawiona w następnej sekcji). W najbardziej ogólnym ujęciu zliczyliśmy procent nowych lematów, dla których chociaż jedna z sugestii była przydatna lingwiście, tzn. przynajmniej w przypadku jednego sugerowanego regionu dołączania nowa jednostka leksykalna dla rozpatrywanego nowego lematu została dodana w obrębie tego regionu. Przy tak zdefiniowanej dokładności otrzymaliśmy następujące wyniki:

— 63,76% — dla wszystkich nowych lematów nominalnych,
— 61,43% — dla nowych lematów opisanych jako odsłowniki (gerundia) lub niejednoznacznych pomiędzy odsłownikami i rzeczownikami,
— 64,12% — dla nowych lematów, które zostały rozpoznane jednoznacznie przez analizator morfologiczny Morfeusz (Woliński, 2006) jako rzeczowniki.

Zbiorcze rezultaty całości procesu półautomatycznego rozszerzania są znacznie niższe niż analogiczne rezultaty dla dużej próby testowej, tj. 80,36% (por. Piasecki i in., 2009b). Przyczyną jest wystąpienie w toku prac nad rozszerzaniem znaczniej większej liczby mniej częstych lematów niż to miało miejsce na początku prac, kiedy była losowana próba testowa. Automatyczny opis tego typu lematów zwykle jest gorszy pod względem dokładności. Algorytm DRA wykazuje niższą dokładność dla odsłowników, chociaż powyżej różnica nie jest znacząca. Spowodowane jest to większą średnio liczbą sugestii, które są generowane dla odsłowników niż dla rzeczowników. Niemniej, wartościowe sugestie, tzn. takie, które zostały wykorzystane przez lingwistów, zostały wygenerowane dla większości nowych lematów, włączając w to nawet te relatywnie rzadkie.

8. Wydobywanie opisu semantycznego struktur

Jak już wspominaliśmy, automatyczne wydobywanie z korpusu semantycznej reprezentacji poszczególnych zdań, a tym bardziej całego tekstu, jest bardzo trudne. Większość metod automatycznych działających na poziomie struktur semantycznych koncentruje się na wydobyciu semantycznego opisu podstawowych struktur argumentowo-predykatowych. Predykatem jest w tym ujęciu każdy element struktury językowej, któremu możemy przypisać wymaganie wystąpienia określonej liczby argumentów. Pod względem składniowym wymagane pozycje argumentowe mogą być scharakteryzowane poprzez podanie ich części mowy i wymaganych wartości wybranych kategorii gramatycznych (głównie przypadka). Na poziomie semantycznym często można określić dla każdej pozycji argumentowej zbiór kategorii semantycz-

nych, do których należą wyrażenia pojawiające się na niej najczęściej. Ze względu na mało wyraźne rozgraniczenie pomiędzy różnymi jednostkami leksykalnymi oraz ze względu na użycia metaforyczne, bardzo trudno byłoby sformułować precyzyjne listy kategorii semantycznych wyrażeń, które mogą pojawić się na poszczególnych pozycjach argumentowych. Dlatego też wygodniej jest mówić o kategoriach preferowanych dla danej pozycji, inaczej o preferencjach selektywnych (ang. *selectional preferences*) (por. Manning, Schütze, 2001) lub rolach semantycznych (por. Hajnicz, 2007). Zbiór kategorii semantycznych jest definiowany w relacji do pewnego zasobu zbudowanego ręcznie. Kategorie semantyczne wyrażeń argumentowych mogą odpowiadać kodom semantycznym przypisywanym jednostkom leksykalnym w słownikach (np. Bullon i in., 2003) lub też mogą być wyznaczone na podstawie struktury istniejącego zasobu opisującego semantykę leksykalną, np. na podstawie hierarchii hiperonimii Princeton WordNet (Fellbaum, 1998; Miller i in., 2006) lub Słowosieci (Derwojedowa i in., 2008; Piasecki i in., 2009c) (por. Hajnicz, 2007). W tym drugim przypadku wybrane ogólne hiperonimy traktowane są jako wyznaczniki kategorii semantycznych, do których należą wszystkie ich bezpośrednie i pośrednie hiponimy. W pracach Hajnicz i Wiecha (2008) oraz Hajnicz (2009a) wykorzystany został wewnętrzny, organizacyjny podział nominalnych jednostek leksykalnych Słowosieci na grupy tematyczne do wyznaczenia kategorii najwyższego poziomu. Podejście takie było uzasadnione dużym rozczłonkowaniem hiperonimii w ówczesnej ograniczonej wersji Słowosieci. Warto jednak zauważyć, że grupy tematyczne nie są częścią struktury Słowosieci, a jedynie narzędziem do podziału pracy pomiędzy lingwistów. Przydział do grup został zdefiniowany ręcznie na samym początku pracy nad Słowosiecią i jest w dużym stopniu nieadekwatny do poprawnego opisu jednostek leksykalnych. W przypadku pełnej wersji Słowosieci, tj. 1.0, struktura relacji jest jedynym narzędziem opisu semantycznego jednostek leksykalnych. Podejście przedstawione w pracy Hajnicz (2009b) opiera się już bezpośrednio na strukturze Słowosieci w opisie semantycznym pozycji argumentowych.

Proces wydobywania kategorii preferowanych rozpoczyna się zwykle od określenia dla każdego wystąpienia wyrażenia predykatywnego w tekście ramy subkategoryzacji syntaktycznej, która jest reprezentowana przez to wystąpienie (czyli liczby i charakterystyki argumentów wymaganych pod względem syntaktycznym). Jest to dość złożony problem, ponieważ wymaga przeprowadzenia automatycznej analizy struktury syntaktycznej tekstu. Eksperymenty przeprowadzone dla języka polskiego zostały przedstawione między innymi w pracy Przepiórkowskiego (2009). Następnie określane są kategorie semantyczne, które są w danym kontekście reprezentowane przez rozpoznane argumenty. Napotykamy tu dwa problemy:

— argument jest często wyrażeniem złożonym, gdy tymczasem kategorie semantyczne są w zasobach językowych przypisywane do jednostek leksykalnych,
— bardzo często również argumenty są niejednoznaczne pod względem semantycznym.

W pierwszym przypadku najczęściej przyjmuje się uproszczenie, w wyniku którego za zbiór kategorii całego argumentu przyjmuje się zbiór kategorii jego głównego elementu (ang. *head*). Problem drugi można rozwiązać na dwa sposoby. Po pierwsze można zastosować wcześniej jeden z algorytmów ujednoznacznia semantycznego (często określany jako algorytm ujednoznacznia sensów słów, ang. *word sense disambiguation*) (por. Agirre, Edmonds, 2006), w wyniku czego każdemu wystąpieniu niejednoznacznego lematu zostanie przypisana pewna jednostka leksykalna. Niestety,

wszystkie znane algorytmy ujednoznaczniania semantycznego wykazują istotny poziom błędu w rozstrzyganiu niejednoznaczności oraz implementacje tych algorytmów są zwykle ograniczone do pewnego podzbioru jednostek leksykalnych (związane jest to z zastosowanym w ich konstrukcji procesem maszynowego uczenia się).

Można jednak próbować uniknąć konieczności jednoznacznego przypisania jednostki leksykalnej do wystąpienia lematu, przyjmując dla każdego wystąpienia zbiór wszystkich jednostek odpowiadających danemu lematowi jako możliwych. Przypisanie zbioru jednostek pociąga za sobą przypisanie zbioru kategorii semantycznych. Rozważmy to na przykładzie lematu *komunikacja*, który reprezentuje trzy jednostki leksykalne: `komunikacja 1` (proces porozumiewania się), `komunikacja 2` (konstrukcja, np. część budynku) oraz `komunikacja 3` (dziedzina gospodarki). Każda z trzech jednostek należy do innej dziedziny semantycznej, którą można wyznaczyć na podstawie hierarchii hiperonimii Słowosieci, np. jako nazwy dziedzin można przyjąć jednostki leksykalne, które są zlokalizowane na górnych poziomach hiperonimii, odpowiednio: `porozumiewanie się 1`, `konstrukcja 4` oraz `komunikacja 3`[38].

Przypisanie zbioru kategorii semantycznych do wystąpienia lematu skutkuje wprowadzeniem pewnego rodzaju szumu informacyjnego, gdyż częstość występowania wielu kategorii semantycznych może być zawyżona w wyniku przypisania całych zbiorów kategorii zamiast wyłącznie właściwych. W podejściach takich zakłada się jednak, że następujący później etap wyznaczenia statystycznie istotnych zależności pomiędzy kategoriami a pozycjami argumentowymi ram poszczególnych wyrażeń predykatywnych wyeliminuje przypadkowe powiązania.

Po identyfikacji ram subkategoryzacji oraz kategorii lub zbiorów kategorii dla poszczególnych pozycji argumentowych, zlicza się dla całego korpusu dystrybucję kategorii. Budowana jest statystyka częstości wystąpień poszczególnych kategorii semantycznych na poszczególnych pozycjach argumentowych w obrębie konkretnych ram konkretnych lematów. Inaczej ujmując, celem jest określenie zbiorów kategorii preferowanych na poszczególnych pozycjach argumentowych dla każdego lematu predykatywnego i każdej jego ramy. Po zebraniu częstości stosuje się różnorodne miary oceny statystycznej siły powiązania kategoria semantyczna–pozycja argumentowa zleksykalizowanej ramy subkategoryzacji (por. Manning, Schütze, Hajnicz, 2007).

9. Zastosowania metod automatycznych

Jak już było wspomniane, mechanizm wzorców leksykalno-syntaktycznych (por. sekcja 5) może być postrzegany jako rozszerzony sposób generowania konkordancji w oparciu o złożony język opisu kontekstów użyć wydobytych par lematów. Wzorce stosowane w ramach rozszerzania istniejącego tezaurusa można łatwo połączyć ze śledzeniem miejsc dopasowanych do wzorca.

Ogólne wzorce pozyskiwane automatycznie (por. sekcja 5), poza automatycznym wydobywaniem hiperonimii, można stosować z powodzeniem do wydobywania bardzo specjalizowanych relacji o charakterze ontologicznym.

38 Jednostka `komunikacja 3` nie ma już hiperonimu w Słowosieci 1.0 jako jednostka dość ogólna znaczeniowo. Nie jest wykluczone, że hiperonim pojawi się w kolejnej wersji Słowosieci 2.0 budowanej w ramach projektu N N516 068637.

Miara powiązania znaczeniowego (dalej MPZ) jest narzędziem stosowanym już od dłuższego czasu w psycholingwistyce, w tym również w odniesieniu do języka polskiego (np. Kruszyński, Rączaszek-Leonardi, 2006).

Za pomocą wyznaczenia listy lematów najbardziej powiązanych semantycznie ze wskazanym lub grupowania lematów silnie powiązanych semantycznie z sobą nawzajem można wygenerować na podstawie MPZ pola semantyczne różnego rodzaju. Pola takie są wyznaczone na podstawie określonego korpusu obiektywną metodą opierającą się na ściśle zdefiniowanych metodach analizy tekstu.

Wydobyte automatycznie pola semantyczne dają wgląd w rozkład znaczeń w ramach określonego korpusu oraz pozwalają na wykrycie nietypowych powiązań znaczeniowych, np. bardzo specyficznych dla danej dziedziny. Umożliwiają zobrazowanie powiązań znaczeniowych obecnych w pewnej zbiorowości, która wytworzyła korpus oraz pozwalają na śledzenie zmian zachodzących w czasie w przypadku korpusu chronologicznego. Zapewniają również wsparcie dla analizy porównawczej różnych korpusów tekstu pod względem struktury w obszarze semantyki leksykalnej.

Metody automatycznego wydobywania relacji semantycznych mogą być cennym narzędziem przyspieszającym prace nad konstrukcją tezaurusów. W odniesieniu do języka polskiego algorytm rozszerzania poprzez regiony aktywacji i oparte na nim narzędzie lingwisty o nazwie WordNet Weaver (Piasecki i in., 2009b, c) zostały z powodzeniem zastosowane do półautomatycznego rozszerzania Słowosieci ze stanu tzw. jądra do pełnej wersji 1.0. Proces rozszerzania przebiegał w przedstawionych poniżej głównych etapach:

1. Zgromadzony został duży korpus tekstu (określany jako połączony korpus w sekcji 5), który został następnie przetworzony na poziomie morfo-syntaktycznym.
2. Z korpusu zostały wydobyte zbiory par powiązanych semantycznie lematów z zastosowaniem wszystkich omówionych metod wydobywania: ręcznych wzorców leksykalno-syntaktycznych (sekcja 5), wzorców wydobytych algorytmem Estratto (sekcja 5) oraz MPZ wygenerowanej z zastosowaniem uogólnionego algorytmu ważenia rangowego (por. Piasecki i in., 2009c; Broda i in., 2009).
3. Klasyfikator oznaczający pary lematów jako instancje semantycznych relacji leksykalnych (por. sekcja 7) został wytrenowany w oparciu o pary lematów z ówczesnej wersji Słowosieci i dane pozyskane w etapie poprzednim.
4. Nowe lematy (tj. nieopisane jeszcze w Słowosieci) zostały pogrupowane na podstawie wartości MPZ z wykorzystaniem gotowego programu do grupowania danych o nazwie Cluto (Karypis, 2002).
5. Wybrane grupy nowych lematów były ładowane do systemu WordNet Weaver, a algorytm rozszerzania poprzez regiony aktywacji był następnie uruchamiany w celu wygenerowania sugerowanych miejsc dołączenia nowych lematów.
6. Lingwiści mogli pracować swobodnie na wprowadzonych grupach lematów — mogli przeglądać propozycje w dowolnej kolejności i edytować strukturę wordnetu.
7. W dowolnym momencie lingwista mógł uruchomić ponownie algorytm generujący sugestie, aby potencjalnie otrzymać lepsze wyniki dla tych nowych lematów, które jeszcze nie zostały poddane edycji.
8. Po ukończeniu edycji określonych grup lingwiści informowali koordynatora, który mógł przeanalizować i skorygować rezultaty, używając dokładnie tego samego sytemu WordNet Weaver. System zapewnia swobodny dostęp wszystkim członkom zespołu poprzez Internet.

Proces rozszerzania Słowosieci, ocena działania poszczególnych narzędzi oraz osiągnięte wyniki zostały szczegółowo przedstawione w pracy Piaseckiego i in. (2009c). Tutaj warto jednak odnotować, że w rezultacie Słowosieć została rozszerzona o 8316 nowych lematów, 10 537 nowych jednostek leksykalnych, 8729 synsetów (tj. zbiorów synonimów — podstawowych elementów struktury wordnetu) i 11 063 instancji leksykalnych relacji semantycznych. Cały proces został zrealizowany w 3–4 osobomiesiące pracy lingwistów. Jako efekt uboczny wykryto szereg błędów w strukturze jądra Słowosieci, które zostały następnie poprawione w toku rozszerzania. Zastosowanie narzędzi automatycznych przyniosło znaczące, szacowane na 5–6-krotne, skrócenie czasu pracy. Niestety, nie została przeprowadzona wnikliwa analiza porównawcza z zastosowaniem metody wyłącznie ręcznej[39]. Konieczne są dalsze eksperymenty i wnikliwa analiza, ale kierunek jest bardzo obiecujący.

Fleksyjny i niepozycyjny charakter języka polskiego sprawiają, że nie jest możliwe bezpośrednie zastosowanie wielu metod opracowanych dla języka angielskiego. Stan ten zmienia się i jest już dostępny pakiet narzędzi SuperMatrix (Broda, Piasecki, 2008) pozwalający na przeprowadzenie różnorodnych eksperymentów z metodami statystycznymi dla języka polskiego.

Bibliografia

Agirre, E., Edmonds, P., (ed.), 2006: *Word Sense Disambiguation: Algorithms and Applications*, Springer.

Apresjan, J. D., 2000: *Semantyka leksykalna. Synonimiczne środki języka*, tł. Z. Kozłowska, A. Markowski,Warszawa.

Bańko, M., 2001: *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*, Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego.

Baldwin, T., 2007: *Scalable deep linguistic processing: Mind the lexical gap*, [in:] *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*, p. 3–12, Seoul, Korea (http://www.cs.mu.oz.au/ tim/pubs/paclic2007.pdf).

Berland, M., Charniak, E., 1999: *Finding parts in very large corpora*, [in:] *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, p. 57–64, Morristown, NJ, USA, Association for Computational Linguistics.

Broda, B., Piasecki, M., 2008: *SuperMatrix: a general tool for lexical semantic knowledge acquisition*, [in:] G. Demenko, K. Jassem, S. Szpakowicz (ed.) *Speech and Language Technology*, volume 11, p. 239–254, Polish Phonetics Assocation.

Broda, B., Derwojedowa, M., Piasecki, M., 2008a: *Recognition of structured collocations in an inflective language*, 'Systems Science', 34(4), 27–36.

Broda, B., Piasecki, M., Szpakowicz, S., 2008b: *Sense-based clustering of Polish nouns in the extraction of semantic relatedness*, [in:] *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, p. 83–89.

Broda, B., Piasecki, M., Szpakowicz, S., 2009: *Rank-Based Transformation in measuring semantic relatedness*, [in:] Gao, Japkowicz, 2009, p. 187–190; Bullon, S., Fox, C., Manning, E., Murphy, M., Urbom, R., Marwick, K. C. (ed.), 2003.

Longman Dictionary of Contemporary English, Pearson Education Limited, Fith impression.

39 Ponieważ rozbudowa Słowosieci była jednym z istotnych celów prac projektowych, nie było środków finansowych na dublowanie prac.

Calzolari, N., Cardie, C., Isabelle, P. (ed.), 2006: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Associationfor Computational Linguistics*, The Association for Computer Linguistics.

Caraballo, S. A., 1999: *Automatic construction of a hypernym-labeled noun hierarchy from text*, [in:] *Proceedings of ACL-99*, p. 120–126, Baltimore, MD (http://acl.ldc.upenn.edu//P/P99/P99-1016.pdf).

Ceglarek, D., Rutkowski, W., 2006: *Automated acquisition of semantic relations for information retrieval systems*, [in:] W. Abramowicz, H. C. Mayr (ed.), *Business Information Systems 10th International Conference, BIS 2006, Proceedings*, volume 85 of Lecture Notes in Informatics, p. 329–341, Gesellschaft für Informatik.Crouch, C. J., Yang, B., 1992: *Experiments in automatic statistical thesaurus construction*, [in:] *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 77–88, New York, NY, USA, ACM Press.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., 2007a: *PolishWordNet on a shoestring*, [in:] *Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, April 11–13 2007, p. 169–178, Universität Tübingen.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., 2007b: *Relacje w polskim WordNecie. Raport 1*, Politechnika Wrocławska, Instytut Informatyki Stosowanej.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., Broda, B., 2008: *Words, concepts and relations in the construction of Polish WordNet*, [in:] Tancs et al., p. 162–177.

Derwojedowa, M., Głabska, M., Piasecki, M., Rabiega-Wisniewska, J., Szpakowicz, S., Zawisławska, M., 2009: *Słowosieć 1.0 — wordnet języka polskiego*,. (www.plwordnet.pwr.wroc.pl, strona umożliwiająca elektroniczny dostęp do wordnetu języka polskiego o nazwie Słowosieć 1.0, ang. plWordNet).

Dowty, D. R., 1979: *Word Meaning and Montague Grammar, volume 7 of Synthese Language Library*, D. Reidel Publishing Company, Dordrecht: Holland/Boston: USA/London: England.

Dowty, D. R., Wall, R. E., Peters, S., 1981: *Introduction to Montague Semantics, volume 11 of Synthese Language Library*, D. Reidel Publishing Company, Dordrecht: Holland/Boston: USA/London: England.

Fellbaum, C. (ed.), 1998: *WordNet — An Electronic Lexical Database*, The MIT Press.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z., 2005: *New experiments in distributional representations of synonymy*, [in:] *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, p. 25–32, Ann Arbor, Michigan, Association for Computational Linguistics.

Fundacja Wikimedia (2009). Polska Wikipedia. http://pl.wikipedia.org (polskojęzyczna, autonomiczna edycja Wikipedii — otwartej, internetowej encyklopedii).

Gao, Y., Japkowicz, N. (ed.), 2009: *Advances in Artificial Intelligence*, 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009, Kelowna, Canada, May 25–27, 2009, Proceedings, volume 5549 of Lecture Notes in Computer Science, Springer.

Girju, R., Badulescu, A., Moldovan, D., 2006: *Automatic discovery of part-whole relations*, 'Computational Linguistics', 32(1), 83–135.

Grefenstette, G., 1993: *Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches*, [in:] *Proceedings of The Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, SIGLEX'93. ACL.

Grochowski, M., 1993: *Obiekty, cele i metody definiowania a rodzaje definicji. Zarys problematyki*, [w:] J. Bartmiński, R. Tokarski (red.), *O definicjach i definiowaniu*, s. 35–45, Wyd. Uniwersytetu Marii Curie-Skłodowskiej, Lublin.

Hajnicz, E., 2007: *Dobór czasowników do badań przy tworzeniu słownika semantycznego czasowników polskich*, Raport IPI PAN 1003, IPI PAN (http://nlp.ipipan.waw.pl/ hajnicz/Verbs-for-Diathesis.pdf).

Hajnicz, E., 2009a: *Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm*, [in:] Mykowiecka, Marciniak, p. 163–190, Bolc Festschrift.

Hajnicz, E., 2009b: *Similarity measure between frames for Polish semantic valence dictionary*, [in:] Z. Vetulani (ed.), *Proceedings of the 3rd Language & Technology Conference November 6–8, 2009 Poznan, Poland*, Wydawnictwo Poznańskie Sp. z o.o.

Hajnicz, E., Wiech, M., 2008: *Applying grade methods to detect similarity of semantic categories of nouns for semantic valence dictionary creation*, [in:] Kłopotek et al., p. 259–268.

Harris, Z. S., 1968: *Mathematical Structures of Language*, Interscience Publishers, New York.

Hearst, M. A., 1992: *Automatic acquisition of hyponyms from large text corpora*, [in:] *Proceeedings of COLING-92*, p. 539–545, Nantes, France, The Association for Computer Linguistics.

Hearst, M. A., 1998: *Automated Discovery of WordNet Relations*, p. 131–151, Volume 1 of Fellbaum.

Jacquemin, C., 2001: *Spotting and Discovering Terms through Natural Language Processing*, The MIT Press.

Karypis, G., 2002: *CLUTO a clustering toolkit. Technical Report 02–017*, Department of Computer Science, University of Minnesota (http://www.cs.umn.edu/ cluto).

Kłopotek, M. A., Wierzchoń, S. T., Trojanowski, K. (ed.), 2006: *Intelligent Information Processing and Web Mining — Proceedings of the International IIS: IIPWM'06 Conference held in Wisła, Poland, June, 2006*, Advances in Soft Computing, Springer, Berlin.

Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., Trojanowski, K. (ed.), 2008: *Intelligent Information Systems XVI. Proceedings of the International IIS'08 Conference held in Zakopane, Poland, June, 2008*, Advances in Soft Computing, Academic Publishing House EXIT, Warsaw.

Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., Trojanowski, K. (ed.), 2009: *Recent Advances in Intelligent Information Systems*, Academic Publishing House EXIT, Warszawa (http://iis.ipipan.waw.pl/2009/proceedings.html).

Kruszynski, B., Rączaszek-Leonardi, J., 2006: *Między strukturalistyczną a psychologiczną reprezentacją znaczenia: wielowymiarowa przestrzeń semantyczna (HAL)*, [w:] Stalmaszczyk, s. 282–295.

Kurc, R., Piasecki, M., 2008: *Automatic acquisition of WordNet relations by the morphosyntactic patterns extracted from the corpora in Polish*, [in:] *Proceedings of the International Multiconference on Computer Science and Information Technology — 3nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, p. 181–188.

Landauer, T. K., Dumais, S. T., 1997: *A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition*, 'Psychological Review', 104(2), 211–240.

Lewandowska-Tomaszczyk, B. (red.), 2005: *Podstawy językoznawstwa korpusowego*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Lin, D., 1993: *Principle-based parsing without overgeneration*, [in:] *Proceedings of ACL-93*, Columbus, Ohio.

Manning, C. D., Schütze, H., 2001: *Foundations of Statistical Natural Language Processing*, The MIT Press.

Martinek, J., 1997: *Pozyskiwanie informacji semantycznej ze słowników jednojęzycznych*, [w:] Z. Bubnicki, A. Grzech (red.), *Inżynieria wiedzy i systemy ekspertowe (materiały konferencji)*, tom 1, s. 326–333.

McEnery, T., 2003: *Corpus Linguistics*, rozdz. 24, p. 448–463, [in:] Mitkov.

Mel'čuk, I. A., 1988: *Dependency Syntax: Theory and Practice*, State University of New York Press.

Miller, G. A., Fellbaum, C., Tengi, R., Wolff, S., Wakefield, P., Langone, H., Haskell, B., 2006: *WordNet — a lexical database for the English language* (http://wordnet.princeton.edu/ — strona WWW projektu).

Mitkov, R. (ed.), 2003: *The Oxford Handbook of Computational Linguistics*, Oxford University Press.

Mohammad, S., Hirst, G., 2006: *Distributional measures as proxies for semantic relatedness* (http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf. — submitted for publication in Kluwer).

Mykowiecka, A., Marciniak, M. (ed.), 2009: *Aspects of Natural Language Processing*, 'LNCS' 5070, Springer. Bolc Festschrift.

Pantel, P., Pennacchiotti, M., 2006: *Esspresso: Leveraging generic patterns for automatically harvesting semantic relations*, [in:] Calzolari et al., p. 113–120 (http://www.aclweb.org/anthology/P/P06/P06-1015).

Piasecki, M., 2004: *Selektywne wprowadzenie do semantyki formalnej*, [w:] Szymanik, Zajenkowski, s. 113–155.

Piasecki, M., 2006: *Handmade and automatic rules for Polish tagger*, [in:] Sojka et al.

Piasecki, M., 2007: *Polish tagger TaKIPI: Rule based construction and optimisation*, 'Task Quarterly', 11(1–2), 151–167 (http://www.task.gda.pl/files/quart/TQ2007/01-02/tq111t-g.pdf).

Piasecki, M., 2008: *Cele i zadania lingwistyki informatycznej*, [w:] P. Stalmaszczyk (red.), *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, s. 252–290, Lexis, Kraków (http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/MetJezII-piasecki-ostateczna.pdf).

Piasecki, M., Broda, B., 2007: *Semantic similarity measure of Polish nouns based on linguistic features*, [in:] Abramowicz, W. (ed.), *Business Information Systems 10th International Conference, BIS 2007, Poznan, Poland*, April 25–27, 2007, Proceedings, volume 4439 of LNCS. Springer.

Piasecki, M., Broda, B., 2009: *Przykładowe wyniki działania miary powiązania znaczeniowego* (http://plwordnet.pwr.wroc.pl/browser/automatic.jsp — strona WWW prezentująca wyniki działania miary powiązania znaczeniowego).

Piasecki, M., Radziszewski, A., 2009: *Morphosyntactic constraints in acquisition of linguistic knowledge for Polish*, [in:] Mykowiecka, Marciniak, p. 163–190, Bolc Festschrift.

Piasecki, M., Szpakowicz, S., Broda, B., 2007: *Extended similarity test for the evaluation of semantic similarity functions*, [in:] Z. Vetulani (ed.), *Proceedings of the 3rd Language and Technology Conference, October 5–7, 2007, Poznań, Poland*, p. 104–108, Poznań, Wydawnictwo Poznańskie Sp. z o.o.

Piasecki, M., Szpakowicz S., Marcińczuk M., Broda, B., 2008: *Classification-based filtering of semantic relatedness in hypernymy extraction*, [in:] B. Nordström, A. Ranta (ed.), *Advances in Natural Language Processing, 6th International Conference*, *GoTAL 2008, Gothenburg, Sweden, August 25–27, 2008, Proceedings*, 'LNCS' 5221, p. 393–404, Springer.

Piasecki, M., Broda, B., Marcinczuk, M., 2009a: *Przykłady działania systemu WordNet Weaver do półautomatycznego rozszerzania Słowosieci* (http://plwordnet.pwr.wroc.pl/browser/graphs.jsp — strona WWW prezentujaca przykładowe wyniki rozszerzania wersji Słowosieci z czerwca 2008 roku).

Piasecki, M., Broda, B., Głabska, M., Marcinczuk, M., Szpakowicz, S., 2009b: *Semiautomatic expansion of Polish WordNet based on Activation-Area Attachment*, [in:] *Recent Advances in Intelligent Information Systems*, p. 247–260, Academic Publishing House EXIT (http://iis.ipipan.waw.pl/2009/proceedings/iis09-25.pdf).

Piasecki, M., Szpakowicz, S., Broda, B., 2009c: *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej (http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf).

Piasecki, M., Broda, B., Marcinczuk, M., Szpakowicz, S., 2009d: *The WordNet Weaver: Multi-criteria voting for semi-automatic extension of a wordnet*, [in:] Gao, Japkowicz, p. 237–240.

Piotrowski, T., 1994: *Z zagadnień leksykografii*, Wydawnictwo Naukowe PWN, Warszawa.
Piotrowski, T., Saloni, Z., 1999: *Kieszonkowy słownik angielsko-polski i polsko-angielski*, Wyd. Wilga, Warszawa.
Polanski, K. (red.), 1993: *Encyklopedia językoznawstwa ogólnego*, Ossolineum.
Przepiórkowski, A., 2004: *Korpus IPI PAN. Wersja wstępna*, Instytut Podstaw Informatyki PAN.
Przepiórkowski, A., 2008: *Powierzchniowe przetwarzanie języka polskiego*, Akademicka Oficyna Wydawnicza EXIT.
Przepiórkowski, A., 2009: *Towards the automatic acquisition of a valence dictionary for Polish*, [in:] Mykowiecka, Marciniak, p. 191–210 (http://adam.przepiorkowski.pl/Papers/2009-festschrift. Bolc Festschrift).
Przepiórkowski, A., Degórski, Ł., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V., Wójtowicz, B., 2007: *Towards the automatic extraction of definitions in Slavic*, [in:] J. Piskorski, B. Pouliquen, R. Steinberger, H. Tanev (ed.), *Proceedings of the Balto-Slavonic NLP workshop at ACL 2007*, p. 43–50, Prague, ACL.
Pustejovsky, J., 1991: *Generative Lexicon*, 'Computational Linguistics', 17(4), 409–441.
Wydawnictwo Naukowe PWN, 2007: *Słownik języka polskiego* (http://sjp.pwn.pl/).
Roget, P. M., 1856: *Thesaurus of English Words and Phrases*, Longman, Brown, Green, and Longmans, London, Fourth edition, enlarged and improved.
Rosenzweig, J., Mihalcea, R., Csomai, A., 2007: *WordNet Bibliography* (http://lit.csci.unt.edu/%7Ewordnet — web page: a bibliography referring to research involving the WordNet lexical database).
„Rzeczpospolita", 2008: Korpus Rzeczpospolitej [on-line]: www.cs.put.poznan.pl/dweiss/rzeczpospolita (korpus artykułów prasowych pobranych z internetowego serwisu gazety „Rzeczpospolita").
Sahlgren, M., 2001: *Vector-based semantic analysis: Representing word meanings based on random labels*, [in:] *Proceedings of the Semantic Knowledge Acquisition and Categorisation Workshop, ESSLLI 2001*, Helsinki, Finland.
Saint-Dizier, P., Viegas, E. (ed.), 1995a: *Computational Lexical Semantics*, Cambridge University Press.
Saint-Dizier, P., Viegas, E., 1995b: *An Introduction to Lexical Semantics from a Linguistic and a Psychological Perspective*, rozdz. 1, p. 1–30, [in:] Saint-Dizier, Viegas.
Siegel, N., Goolsbey, K., Kahlert, R., Matthews, G., 2004: *The Cyc® system: Notes on architecture*, Whitepaper, Cycorp, Inc., 3721 Executive Center Drive, Suite 100, Austin, Texas, USA (http://cyc.com/cyc/technology/whitepapers_dir/Cyc_Architecture_and_API.pdf).
Snow, R., Jurafsky, D., Ng, A. Y., 2005: *Learning syntactic patterns for automatic hypernym discovery*, [in:] L. K. Saul, Y. Weiss, L. Bottou (ed.), *Advances in Neural Information Processing Systems 17*, p. 1297–1304, Cambridge, MA, MIT Press (http://www.stanford.edu/ jurafsky/paper887.pdf).
Snow, R., Jurafsky, D., Ng, A. Y., 2006: *Semantic taxonomy induction from heterogenous evidence*, [in:] Calzolari et al. (http://www.stanford.edu/ jurafsky/COLACL101.pdf).
Sojka, P., Kopecek, I., Pala, K. (ed.), 2006: *Proceedings of the Text, Speech and Dialog 2006 Conference*, LNAI, Springer.
Stalmaszczyk, P. (red.), 2006: *Metodologie językoznawstwa. Podstawy teoretyczne*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
Szymanik, J., Zajenkowski, M. (red.), 2004: *O umyśle umyślnie i nieumyślnie*, Koło Filozoficzne przy MISH, Uniwersytet Warszawski, Warszawa.
Tancs, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (ed.), 2008: *Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, Hungary, January 22–25, 2008*, University of Szeged, Department of Informatics.

Tufiş, D., Cristea, D., Stamou, S., 2004: *BalkaNet: Aims, methods, results and perspectives. A general overview*, 'Romanian Journal of Information Science and Technology' 7(1–2), 9–43 (http://www.ceid.upatras.gr/Balkanet/journal/7_Overview.pdf Special Issue).

Turney, P. D., Littman, M. L., Bigham, J., Shnayder, V., 2003: *Combining independent modules to solve multiple-choice synonym and analogy problems*, [in:] *Proceedings International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, p. 482–489, Borovets, Bulgaria.

Vossen, P., 2002: *EuroWordNet general document version 3. Technical report*, University of Amsterdam.

Widdows, D., 2004: *Geometry and Meaning*, CSLI Publications.

Wierzbicka, A., 2006: *Semantyka. Jednostki elementarne i uniwersalne*, UMCS, Lublin.

Wolinski, M., 2006: *Morfeusz — a practical tool for the morphological analysis of Polish*, [in:] Kłopotek et al., p. 511–520.

Zesch, T., Gurevych, I., 2006: *Automatically creating datasets for measures of semantic relatedness*, [in:] *Proceedings of the Workshop on Linguistic Distances*, p. 16–24, Sydney, Australia, Association for Computational Linguistics (http://www.aclweb.org/anthology/W/W06/W06-1104).