Maciej Piasecki

Autoreferat

Data urodzenia: 28 stycznia 1970
Miejsce urodzenia: Wrocław
Obywatelstwo: polskie

Miejsce zatrudnienia: Katedra Inteligencji Obliczeniowej,
Wydział Informatyki i Zarządzania, Politechnika Wrocławska

Adres służbowy: Katedra Inteligencji Obliczeniowej, Politechnika Wrocławska
ul. Wybrzeże Wyspiańskiego 27
50-370 Wrocław

E-mail: maciej.piasecki@pwr.edu.pl

2. Wykształcenie – posiadane dyplomy

06.1988 **14 Liceum Ogólnokształcące im. Polonii Belgijskiej** we Wrocławiu
Świadectwo maturalne.

09.1993 **Politechnika Wrocławska**, Wydział Elektroniki
Dyplom ukończenia studiów magisterskich, kierunek **Informatyka**, specjalność **Systemy mikroprocesorowe i mikrokomputerowe**, z wynikiem bardzo dobrym z wyróżnieniem.
Praca magisterska pt. *Modelowanie semantyki zdań twierdzących języka polskiego za pomocą formuł rachunku intensjonalnego* napisana pod kierunkiem dr inż. Artura Rozwadowskiego dotyczyła podstaw automatycznej analizy znaczenia tekstu, w dorocznym ogólnopolskim konkursie Polskiego Towarzystwa Informatycznego na najlepszą pracę magisterską z dziedziny informatyki moja praca uzyskała drugą nagrodę.

06.2003 **Politechnika Wrocławska**, Wydział Informatyki i Zarządzania
Dyplom stopnia doktora nauk technicznych o specjalności informatyka uzyskany na podstawie rozprawy doktorskiej *Język modelowania znaczenia polskiej frazy nominalnej* napisanej pod kierunkiem prof. dr hab. inż. Zbigniewa Huzara. Praca otrzymała wyróżnienie. W 2003 otrzymałem nagrodę Rektora PWr przyznawaną dorocznie autorom najlepszych prac doktorskich.

3. Zatrudnienie

— **Bogart zo.o.** – firma z siedzibą we Wrocławiu
  — 1 IX 1993 r. – 31 I 1994 r. stanowisko projektanta-programisty w wymiarze pełnego etatu; brałem udział w budowie multimedialnych systemów informacyjnych oraz opracowałem koncepcję, projekt i kierowałem pracami nad implementacją systemu zarządzania obiektową bazą danych *OB.! OK.!*.
— **Wydział Informatyki i Zarządzania Politechniki Wrocławskiej**
  — 1 II 1994 r. – 30 IX 2003 r. stanowisko asystenta w wymiarze pełnego etatu,
  — 1 X 2003 do dnia dzisiejszego stanowisko adiunkta w wymiarze pełnego etatu.
— **Wyższa Szkoła Zarządzania w Częstochowie**: 02.2004 – 09.2004 oraz 10.2005–09.2006, stanowisko profesora WSZ w wymiarze pełnego etatu,
— **Wyższa Szkoła Zawodowa Kolegium Karkonoskie w Jeleniej Górze** 10.2004–09.2005 stanowisko wykładowca w wymiarze pełnego etatu.
— **Techland** (`https://techland.pl/`) w latach 1999–2001 współpraca jako konsultant w zakresie lingwistyki informatycznej przy projekcie budowy systemu automatycznego tłumaczenia English-Translator (do wersji 2.0 włącznie).
— **Solidne Oprogramowania spółka z.o.o.**, 06.2009-12.2009, konsultacje z dziedziny projektowania interfejsu użytkownika w ramach realizacji projektu „Stworzenie specjalistycznego portalu pośrednictwa pracy tymczasowej w sektorze IT", Innowacyjna Gospodarka, UDA-POIG.08.01.00-02-116/08-00.

Moje dokonania naukowe w skrócie

193 opublikowane recenzowane prace naukowe (w tym **31** przed doktoratem),

1 170 punktów oceny publikacji wyliczonych wg skali MNiSW obowiązującej we wrześniu 2018 (w tym **50** przed doktoratem),

659 ręcznie sprawdzone cytowania moich publikacji w pracach naukowych (bez autocytowań), zob. sekcja H, (w tym **2** przed doktoratem),

51 864 046,88 zł łącznego budżetu w ramach koordynowanych i kierowanych przeze mnie projektów badawczych, ich części lub wydzielonych dużych zadań badawczych, w realizacji w tych prac wzięły udział setki pracowników badawczych z kilku jednostek naukowych (przed doktoratem nie kierowałem projektami lub zadaniami badawczymi, byłem jedynie wykonawcą).

Najważniejsze aspekty osiągnięcia naukowego (omówione szczegółowo w sekcji 4):
— opracowanie lingwistycznie motywowanego modelu wordnetu (sekcja 4.2),
— opracowanie algorytmu ujednoznaczniania morfosyntaktycznego opartego na warstwowym procesie i wykorzystaniu ograniczeń leksykalno-morfosyntaktycznych do wzmocnienia siły ekspresji drzew decyzyjnych (sekcja 4.4.1) oraz opartego na nim pierwszego tagera morfosyntaktycznego dla języka polskiego o wielu zastosowaniach badawczych (sekcja 5.1),
— opracowanie metody półautomatycznej budowy wordnetu w oparciu o duże zbiory tekstów (sekcja 4.3),
— opracowanie kompleksowego zbioru heterogenicznych metod do wydobywania relacji leksykalno-semantycznych z korpusów polskich tekstów (sekcja 4.4),
— opracowanie algorytmu automatycznego rozszerzania wordnetu w oparciu propagację aktywacji w sieci wordnetu z wykorzystaniem heterogenicznych źródeł wiedzy wydobytych z dużych korpusów tekstu (sekcja 4.6),
— zbudowanie bardzo dużego wordnetu języka polskiego o nazwie *Słowosieć* (ang. *plWordNet*), największego tego typu zasobu językowego na świecie o wielu unikatowych cechach, jednocześnie wielkiego relacyjnego słownika języka polskiego (jednego z największych współczesnych słowników języka polskiego) o setkach zastosowań badawczych w Polsce i na świecie (sekcja 5.2).

4. Osiągnięcie naukowe

Jako osiągnięcie naukowe pt. *Kompleksowa metoda do półautomatyczna konstrukcji dużej leksykalnej sieci semantycznej dla języka polskiego w oparciu o korpus tekstów* przedkładam wymienione poniżej dwanaście prac.

Większość przedstawionych poniżej prac powstała w ramach realizacji projektów naukowych, dotyczy często kilku zadań badawczych i dlatego jest sygnowana przez kilku członków zespołów badawczych. Deklaracje współautorów określające ich wkład w poszczególne publikacje stanowią Załącznik 1 do niniejszego wniosku o przeprowadzenie postępowania habilitacyjnego. Mój własny wkład jest scharakteryzowany krótko po każdej pracy.

Uwaga: wszystkie poniższe prace można pobrać pod adresem:
`https://nextcloud.clarin-pl.eu/index.php/s/v3sLzWOvEURze5F`
Większość moich publikacji (zob. Sekcja B) można pobrać ze strony Research Gate:
`https://www.researchgate.net/profile/Maciej_Piasecki/contributions`

1. Maciej Piasecki and Grzegorz Godlewski. "Effective Architecture of the Polish Tagger". In: *Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic, September*

11-15, 2006, Proceedings. Ed. by Petr Sojka, Ivan Kopeček, and Karel Pala. Vol. 4188. LNCS. **13 pkt IF (lista czasopism w Web of Science) Lista A MNiSW w 2006 r.** Springer, 2006, pp. 213–220
— *mój wkład:* [85%] jestem autorem koncepcji rozwiązania, zaproponowanych metod oraz tekstu pracy; Pan mgr inż. Grzegorz Godlewski był ówcześnie moim dyplomantem, wykonał wszystkie prace programistyczne oraz przeprowadzał eksperymenty pod moim nadzorem;
— *charakter publikacji:* konferencja TSD to najważniejsza konferencja z dziedziny inżynierii języka naturalnego w regionie i prawdopodobnie najważniejsza dla prac dotyczących języków słowiańskich;
— *cytowania:* 9, zob. sekcja H.

2. Maciej Piasecki. "Hand-Written and Automatic Rules for Polish Tagger". In: *Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic, September 11-15, 2006, Proceedings*. Ed. by Petr Sojka, Ivan Kopeček, and Karel Pala. Vol. 4188. LNCS. **13 pkt IF (lista czasopism w Web of Science Lista A MNiSW w 2006 r.)** Springer, 2006, pp. 205–212
— *cytowania:* 10.

3. Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. "Automatic Selection of Heterogeneous Syntactic Features in Semantic Similarity of Polish Nouns". In: *Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, Proceedings*. Ed. by Václav Matousek and Pavel Mautner. Vol. 4629. LNCS. **10 pkt (Lista A** czasopism MNiSW za rok 2007 opublikowana 28 XI 2008 r.) **(Web of Science)**. Springer, 2007, pp. 99–106
— *mój wkład:* zaproponowałem kluczową dla pracy metodę opisu kontekstów tekstowych za pomocą ograniczeń leksykalno-morfosyntaktycznych oraz opracowałem zbiór ograniczeń i adaptacje miar podobieństwa, w znacznej części napisałem treść pracy;
— *cytowania:* 6.

4. Maciej Piasecki, Michał Marcińczuk annd Stanisław Szpakowicz, and Bartosz Broda. "Classification-based Filtering of Semantic Relatedness in Hypernymy Extraction". In: *Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Sweden, August 25-27, 2008, Proceedings*. Ed. by Bengt Nordström and Aarne Ranta. Vol. 5221. LNCS. **10 pkt (Lista A** czasopism MNiSW za rok 2008) **(Web of Science)**. Springer, 2008, pp. 393–404
— *mój wkład:* [60%] zapronowałem reprezentację znaczeń lematów przy pomocy zestawu heterogenicznych cech (odnoszących się do koincydencji lematów, lingwistycznych własności kontekstów i pochodnych miar podobieństwa) oraz wykorzystania jej w rozpoznawaniu relacji semantycznych za pomocą maszynowego uczenia, w znacznej części napisałem treść pracy;
— *cytowania:* 2.

5. Bartosz Broda, Magdalena Derwojedowa, Maciej Piasecki, and Stanisław Szpakowicz. "Corpus-based Semantic Relatedness for the Construction of Polish WordNet". In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. **10 pkt (Web of Science) ACL Anthology**. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/459_paper.pdf`
— *mój wkład:* [70%] opracowałem metodę opisu znaczeń przymiotników za pomocą ograniczeń, zaproponowałem zbiór ograniczeń, zaplanowałem proces walidacji, wykazałem nieprawdziwość znanej hipotezy o niewystępowaniu quasi-synonimicznych przymiotników w bliskich kontekstach tekstowych,w znacznej części napisałem treść pracy;
— *charakter publikacji:* konferencja LREC to jedna z najważniejszych światowych konferencji w dziedzinie lingwistyki informatycznej, artykuł został zakwalifikowany do wygłoszenia,

— *cytowania:* 6.

6. Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up.* **Monografia, 25 pkt, MNiSW 2018: 25 pkt.** Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2009. URL: `www.dbc.wroc.pl/Content/4220/Piasecki_Wordnet.pdf`
   — *mój wkład:* [70%] opracowałem koncepcję i strukturę pracy, w przeważającej części napisałem treść pracy, z wyjątkiem rozdziału 3.5 byłem głównym autorem zaproponowanych metod i treści rozdziałów;
   — *charakter publikacji:* recenzowana monografia o unikatowym charakterze w skali światowej – wg najlepszej mojej wiedzy jedyna monografia przedstawiająca oryginalne metody i proces półautomatycznej budowy wordnetu w oparciu o korpus tekstów,
   — *cytowania:* 91.

7. Bartosz Broda, Maciej Piasecki, and Stanisław Szpakowicz. "Extraction of Polish Noun Senses from Large Corpora by Means of Clustering". In: *Control and Cybernatics* 39.2 (2010). **13 pkt (Lista A czasopism MNiSW za rok 2010** opublikowana 10 XII 2010r., indeks JRC **IF=0,3 MNiSW 2018: 14 pkt) Web of Science**, pp. 401–420. URL: `http://matwbn.icm.edu.pl/ksiazki/cc/cc39/cc3926.pdf`
   — *mój wkład:* [40%] opracowałem metodę reprezentacji kontekstów, wniosłem istotny wkład w opracowanie szczegółowych rozwiązań w ramach zaproponowanej metody (na przykład łączenie indukowanych znaczeń z wordnetem), w przeważającej części napisałem treść pracy;

8. Roman Kurc, Maciej Piasecki, and Stan Szpakowicz. "Automatic Acquisition of Wordnet Relations by Distributionally Supported Morphological Patterns Extracted from Polish Corpora". In: *Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings.* Ed. by Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala. Vol. 6231. Lecture Notes in Computer Science. **13 pkt (Lista A** czasopism MNiSW za rok 2010 opublikowana 10 XII 2010r.) **(Web of Science)**. 2010, pp. 133–141. ISBN: 978-3-642-15759-2
   — *mój wkład:* [50%] opracowałem podstawowe elementy metody, w przeważającej części napisałem treść pracy,
   — *cytowania:* 2.

9. Maciej Piasecki, Roman Kurc, Radosław Ramocki, and Bartosz Broda. "Lexical Activation Area Attachment Algorithm for Wordnet Expansion". In: *Proceedings of the 15th International Conference on Artificial Intelligence: Methodology, Systems, Applications.* Ed. by Allan Ramsay and Gennady Agre. Vol. 7557. Lecture Notes in Computer Science. **Best Paper Award.** Varna, Bulgaria: Springer, 2012, pp. 23–31. ISBN: 978-3-642-33184-8
   — *mój wkład:* [70%] opracowałem kluczowy element pracy – unikatową metodę automatycznego rozszerzania wordnetu w oparciu o heterogeniczne źródła wiedzy i schemat propagacji aktywacji w grafie, byłem głównym autorem treści pracy;
   — *charakter publikacji:* AIMSA to bardzo dobra konferencja z dziedziny sztucznej inteligencji, z długą tradycją i silnym zespołem znanych badawczy z obszaru inżynierii języka naturalnego w komitecie programowym.

10. Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. "The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations". In: *Langauge Resources and Evaluation* 47.3 (2013). **IF=0,518 15pkt, MNiSW 2018: 20pkt**, pp. 769–796. URL: `http://link.springer.com/article/10.1007%2Fs10579-012-9209-9`

— *mój wkład:* [40%] opracowałem fundamentalną dla pracy koncepcję grupowania jednostek leksykalnych w synsety w oparciu o tzw. konstytutywne relacje leksykalno-semantyczne (wywiera ona coraz większy wpływ na prace w dziedzinie wordnetów i słowników), byłem wiodącym autorem treści pracy,
— *cytowania:* 5.

11. Maciej Piasecki, Radosław Ramocki, and Michal Kaliński. "Information Spreading in Expanding Wordnet Hypernymy Structure". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013.* Ed. by Ruslan Mitkov, Galia Angelova, and Kalina Boncheva. **ACL Anthology**. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sept. 2013, pp. 553–561. URL: `http://aclweb.org/anthology/R13-1073`
— *mój wkład:* [70%] opracowałem kluczowy element pracy – unikatową metodę automatycznego rozszerzania opartą na propagacji aktywacji w grafie, która została uogólniona do połączonego grafu wordnetowego, napisałem większość treści pracy, Pan mgr inż. Michał Kaliński był ówcześnie moim dyplomantem, wykonał wszystkie prace programistyczne oraz przeprowadzał eksperymenty pod moim nadzorem; wkład Pana mgr. inż. Radosława Ramockiego, ówczesnego doktoranta, został przedstawiony w jego oświadczeniu;
— *charakter publikacji:* konferencja RANLP to jedna z najlepszych na świecie konferencji z dziedziny inżynierii języka naturalnego (często szacowana na czwartą w rankingu), z kilkunastoprocentowym współczynnikiem akceptacji dla długich artykułów, takich jak nasz, indeksowana w prestiżowej ACL Anthology.

12. Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. "plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource". In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan.* Ed. by Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad. **ACL Anthology**. ACL. ACL, 2016, pp. 2259–2268. URL: `http://www.aclweb.org/anthology/C16-1213`
— *mój wkład:* [30%] praca przedstawia wyniki wieloletnich interdyscyplinarnych badań nad budową polskiego wordnetu, którymi kierowałem, które były oparte w dużej mierze na moich koncepcjach i na które wywarłem decydujący wpływ, opracowałem strukturę pracy i byłem wiodącym autorem;
— *charakter publikacji:* konferencja COLING to najstarsza światowa konferencja z dziedziny lingwistyki informatycznej, druga lub trzecia na świecie pod względem istotności, z około 30% ogólnym współczynnikiem akceptacji, bardzo trudno jest uzyskać akceptację dla pracy o zasobach językowych, indeksowana w prestiżowej ACL Anthology, ponadto nasz artykuł był zakwalifikowany do ustnego wygłoszenia,
— *cytowania:* 6.

Łączna liczba punktów MNiSW[1]: **149**

Powyższe prace przedstawiają kluczowe elementy zaplanowanego przeze mnie i realizowanego konsekwentnie pod moim kierownictwem w latach 2005–2016 długofalowego programu badawczego. Jego celem była budowa sformalizowanego opisu semantyki leksykalnej języka polskiego zapewniającego dobre odwzorowanie zjawisk językowych i bardzo obszerne pokrycie materiału leksykalnego. Aby to osiągnąć musieliśmy rozwiązać szereg problemów badawczych z dziedziny budowy zasobów językowych

[1] Ponieważ zasady punktacji i rozpiętość skali zmieniały się w ciągu lat, szczególnie w odniesieniu do prac spoza listy A, we wszystkich wyliczeniach punktów przyjąłem zasady parametryzacji z roku 2017 i listy czasopism z roku 2017 obowiązujące w roku 2018.

oraz metod automatycznego wydobywania wiedzy lingwistycznej z dużych zbiorów tekstów. Podjęte wyzwania badawcze i zaproponowane przez nas rozwiązania zostały opisane w pracach wskazanych jako osiągnięcie naukowe oraz krótko omówione w dalszych sekcjach niniejszego referatu.

W wyniku realizacji postawionych celów badawczych powstała, między innymi, dwujęzyczna leksykalna sieć semantyczna o nazwie *Słowosieć* (angielska nazwa *plWordnet*)[2] obecnie dostępna w wersji 4.0. *Słowosieć* można również opisać jako wielki relacyjny słownik języka polskiego. Jako podstawowy zasób językowy dla języka polskiego *Słowosieć* sama w sobie jest osiągnięciem naukowym: to największy tego typu zasób na świecie i jeden z najbogatszych opisów systemu leksykalnego języka polskiego oraz bardzo istotne narzędzie badawcze dla inżynierii języka naturalnego i lingwistyki informatycznej, szczególnie w odniesieniu do języka polskiego, zob. Sekcja 5.2. *Słowosieć* jest również bardzo istotnym zasobem dla polskiej kultury i społeczeństwa. W tym drugim przypadku jako podstawa wielu systemów analizy treści wypowiedzi w języku polskim.

Całość mojego wieloletniego programu badawczego była realizowana w ramach siedmiu dużych i bardzo dużych projektów, z których pięcioma **kierowałem**, a w dwóch pozostałych **kierowałem** dużymi zadaniami częściami projektów wykonywanymi przez Politechnikę Wrocławską, zob. Sekcja J.

4.1. Rola podstawowych zasobów językowych

W ciągu ostatnich kilkunastu lat można było zaobserwować ogromny postęp w rozwoju algorytmów analizy tekstu opartych na metodach statystycznych i automatycznym wydobywaniu modeli opisujących wybrane aspekty języka naturalnego z bardzo dużych zbiorów danych tekstowych. W dziedzinie opisu znaczeń leksykalnych *semantyka dystrybucyjna*, oparta na statystycznej analizie dużych zbiorów tekstów, jest obecnie wręcz dominującym paradygmatem, a technika wektorowej reprezentacji znaczeń słów o angielskiej nazwie *word embeddings* budowanej za pomocą algorytmu *word2vec* [47, 48][3] stała się niemal symbolem analizy semantycznej tekstu. Modele wektorowe i statystyczne modele językowe budowane są przeważnie bezpośrednio z dużych zbiorów tekstów, podzielonych w bardzo uproszczony sposób na wyrazy (ściślej pseudo-wyrazy), ograniczone białymi znakami. Tekst nie jest poddawany żadnej wstępnej analizie językowej. Oczekuje się, że statystyczny algorytm uczący się wydobędzie automatycznie z danych model często bardzo złożonych zależności językowych.

W kontekście sukcesów podejść opartych na statycznym uczeniu się rola ręcznie konstruowanych *zasobów językowych* – czyli baz wiedzy opisujących język naturalny – jest często minimalizowana. Szczególnie jest to widoczne w ramach maszynowego tłumaczenia i wyszukiwania informacji, jak również rozpoznawania pisma i mowy, które wykorzystują modele językowe. Jednak wizja, w której wielki zbiór tekstów jest jedynym i wystarczającym źródłem wiedzy jest kusząca, ale myląca, por. uwagi w [32]. Przede wszystkim metody statystyczne wymagają odpowiednio dużej ilości danych. Jeżeli liczba obserwacji wystąpień pewnej zmiennej losowej spada, to ilość informacji, w oparciu o którą budowany jest model, może okazać się nie wystarczającą, na przykład na podstawie kilku wystąpień słowa x w zbiorze tekstów o rozmiarze setek milionów wyrazów, praktycznie nie jest możliwe zbudowanie modelu opisującego podobieństwo semantyczne pomiędzy x a dziesiątkami tysięcy innych słów, wśród których ogromna większość występuje podobnie rzadko jak x. Na przykład, w pracy [63] pokazaliśmy, jak dokładność reprezentacji relacji semantycznych przez modele wektorowe (tzw. *wektory osadzeń słów*, ang. *word embeddings*) spada bardzo znacząco wraz ze spadkiem częstości występowania opisywanych słów do 30 wystąpień na 4 miliardy wyrazów.

Warto również zauważyć, że we wszystkich wymienionych przykładowych zastosowaniach jest wykorzystywana wiedza pochodząca od ludzi, na przykład w maszynowym tłumaczeniu są to tłumaczenia konkretnych tekstów w postaci tzw. *korpusów równoległych* (ang. parallel corpora). Ponadto wiele zastosowań *inżynierii języka naturalnego*, opartych na analizie treści, nie poddaje się tak łatwo metodom

[2] `http://plwordnet.pwr.edu.pl`

[3] Spis literatury cytowanej w niniejszej i następnej sekcji został umieszczony pomiędzy Sekcją 5 a B

czysto statystycznym, ponieważ sam zbiór tekstów tylko w ograniczonym stopniu zawiera informację wyjaśniającą znaczenie wypowiedzi językowych. Interpretacja wypowiedzi językowych jest zawsze osadzona w pozajęzykowym kontekście. Rosnąca popularność i znaczenie połączonych otwartych danych (POD) (ang. *Linked Open Data*, LOD), będących rodzajem sieci semantycznych, wskazuje na istotność zasobów wiedzy skonstruowanych ręcznie przez człowieka. POD są albo budowane w oparciu o wiedzę wydobywaną ze strukturalizowanych zasobów takich jak Wikipedia[4], albo konstruowane ręcznie, na przykład bazy terminologiczne. Na poziomie opisu morfologicznego ogromna większość rozwiązań jest oparta o sformalizowane opisy morfologii budowane ręcznie. Natomiast programy do analizy składniowej są konstruowane w oparciu o zbiory zdań (czasami też i całych tekstów), których struktura składniowa została ręcznie opisana w sposób sformalizowany za pomocą metadanych (tzw. *anotacji składniowej*).

Nie bez znaczenia jest również fakt, iż ogromna większość metod opartych na statystycznym uczeniu została opracowana pierwotnie dla języka angielskiego. Cechy języka angielskiego w dużym stopniu ułatwiają budowę modeli statystycznych ze względu na jego bardzo ograniczoną morfologię oraz mocno zdeterminowany porządek linearny w ramach wyrażeń językowych i zdań. Języki słowiańskie, w tym język polski, posiadają bogatą morfologię (na przykład ponad 100 możliwych form dla przymiotnika) oraz słabo ograniczony porządek linearny[5], co powoduje, że zakres zmienności, którą należałoby opisać w modelu statystycznym, jest o kilka rzędów większy niż w przypadku języka angielskiego.

Ze względu na ograniczenie modeli statystycznych wydobywanych automatycznie, podstawą dla konstrukcji wielu systemów przetwarzających język naturalny są *zasoby językowe*, czyli sformalizowane bazy wiedzy opisujące język naturalny na różnych poziomach, na przykład słowniki (poziomy morfologiczny i morfo-syntaktyczny), leksykony semantyczne, gramatyki czy też anotowane korpusy tekstów[6]. Zasoby językowe są nieodzowną podstawą[7] do konstrukcji *narzędzi językowych*, czyli programów do analizy języka naturalnego na określonych poziomach opisu, na przykład analizator morfologiczny, parser składniowy czy program do ujednoznaczniania znaczeń leksykalnych słów (ang. *Word Sense Disambiguation*, zwykle nazywany ang. skrótem WSD), lub w zakresie określonych aspektów przetwarzania, na przykład program do rozpoznawania i klasyfikacji użyć nazw własnych i jednostek identyfikacyjnych (ang. *Named Entity Recognition*). Poszczególne typy zasobów i narzędzi językowych przyjęły się[8] w toku rozwoju inżynierii języka naturalnego jako *podstawowe* komponenty technologii językowych[9]. Idea *podstawowego zbioru zasobów i narzędzi językowych*, wyrażona po raz pierwszy w postaci propozycji ich *quasi* standardu BLARK[10] [27, 28], przyjęła się w inżynierii języka naturalnego zgodnie z tendencją do przechodzenia do dojrzałej, komponentowej architektury systemów analizy języka naturalnego opartej o wyspecjalizowane komponenty określonych typów, wykorzystywane w wielu różnych zastosowaniach. Idea BLARK szybko stała się popularna i wraz ze wzrostem popularności architektury opartej na komponentach pojawiło się oczekiwanie, że podstawowe zasoby i narzędzia językowe (czyli ujęte w BLARK) powinny być dostępne dla każdego języka naturalnego, w tym języka polskiego. Co więcej, brak dostępności podstawowych zasobów i narzędzi językowych dla danego języka stał się czynnikiem istotnie ograniczającym rozwój systemów analizy dla danego języka. Bez podstawowych zasobów i

[4] `http://pl.wikipedia.org`

[5] Nieliczne konstrukcje, na przykład przyimkowe, narzucają mocne ograniczenia na porządek linearny, w większości przypadków każda permutacja wyrazów jest poprawna składniowo, choć nie tożsama znaczeniowo

[6] Korpus tekstów to zbiór tekstów dobranych ze względu na pewne kryteria badawcze, na przykład tekstów reprezentujących określoną dziedzinę lub sposób użycia języka. Anotacja to metadane przypisane do tekstu, które opisują wybrane własności językowe, na przykład własności gramatyczne poszczególnych wyrazów.

[7] Przynajmniej w postaci anotowanych ręcznie korpusów tekstów.

[8] Częściowo w rezultacie tradycyjnego podziału w językoznawstwie na poziomy opisu języka naturalnego, częściowo w wyniku wyodrębnia się popularnych zadań w ramach analizy wypowiedzi językowych.

[9] Technologia językowa obejmuje zasoby i narzędzia językowe oraz architektury przetwarzania opracowane w celu połączenia zasobów i narzędzi w systemy przetwarzające język naturalny.

[10] The Basic Language Resource Kit (BLARK).

narzędzi językowych o odpowiedniej dokładności i pokryciu możliwości konstrukcji systemów przetwarzających dany język naturalny są bardzo mocno ograniczone.

W roku 2004 zbiór dostępnych publicznie zasobów i narzędzi językowych dla języka polskiego był bardzo ubogi[11]. Stan ten praktycznie blokował większość możliwych zastosowań. W tym kontekście sformułowałem podstawy mojego długofalowego programu badań, którego ***głównym celem stało się zbudowanie otwartej podstawowej technologii językowej dla języka polskiego umożliwiającej szerokie zastosowania***. Wymagało to realizacji szeregu zadań badawczych oraz bardzo dużych nakładów roboczych na prace badawczo-rozwojowe. Realizacja tak szeroko zakrojonego planu całkowicie przekraczała możliwości pojedynczego badacza i stała się możliwa dzięki współpracy z zespołami naukowymi, którymi kierowałem w szeregu projektów, oraz dzięki nawiązaniu owocnej i długoterminowej współpracy z Zespołem Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN. Pozwoliło to na podzielenie się zakresami odpowiedzialności.

W ramach prac badawczych realizowanych bezpośrednio przeze mnie, skoncentrowałem się na budowie semantycznych zasobów językowych, a w szczególności na dużym *wordnecie języka polskiego*, czyli leksykalnej sieci semantycznej opisującej znaczenia leksykalne za pomocą określonego zbioru relacji leksykalno-semantycznych. Zbudowana w ten sposób sieć otrzymała polską nazwę *Słowosieć* oraz jednocześnie angielską *plWordNet*. Po kilkunastu latach nieprzerwanego rozwoju osiągnęła ona stan, w którym można ją określić również jako *wielki relacyjny słownik semantyczny języka polskiego* – jeden z największych współczesnych słowników języka polskiego.

Podstawowy zasób językowy powinien być dostatecznie duży, aby zapewnić wystarczająco bogate na potrzeby praktycznych zastosowań pokrycie słów występujących w tekstach. Szczególnie dotyczy to leksykonów semantycznych typu wordnet. W roku 2004 wymóg taki oznaczał ogrom pracy do wykonania, rzędu dziesiątek osobolat, tylko w odniesieniu do pojedynczego zasobu językowego – wordnetu! Wydawało się to mało realne do osiągnięcia. Dlatego też od samego początku jednym z celów moich badań stało się ***opracowanie i konsekwentne wdrożenie kompleksowej półautomatycznej metody budowy dużej leksykalnej sieci semantycznej dla języka polskiego***. Metoda ta miała początkowo przede wszystkim zmniejszyć nakłady robocze wymagane do konstrukcji dużego wordnetu, jednak zaowocowała również poprawą jego jakości poprzez bardziej bezpośrednie powiązanie modelu z danymi językowymi, pochodzącymi z dużych zbiorów tekstów. W efekcie zostało opracowane unikatowe kompleksowe rozwiązanie w dziedzinie zasobów językowych oraz elektronicznej leksykografii.

Opracowana metoda stała się również składową unikatowej *metody budowy wordnetu w oparciu o duży korpus tekstów* [40][12], która została wdrożona i konsekwentnie stosowana przez lata przy budowie *Słowosieci* (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016).

Aby osiągnąć zmierzony cel, konieczne było zrealizowanie kilku zdań badawczych dotyczących problemów nierozwiązanych do tej pory dla języka polskiego, a w wielu wypadkach również również na poziomie światowym. Warto podkreślić, że podjęte wyzwania badawcze miały charakter interdyscyplinarny, wymagały połączenia metod i wiedzy z dziedzin językoznawstwa, sztucznej inteligencji i inżynierii języka naturalnego oraz współpracy z interdyscyplinarnymi zespołami badawczymi, która to była niezbędna do rzetelnej oceny jakości rozwiązań na rzeczywistych danych w skali o praktycznej istotności[13]. Wspomniane powyżej wyzwania badawcze to:

[11] Nielicznymi wyjątkami były analizator morfologiczny *Morfeusz* [105] oraz powstający, przełomowy Korpus IPI PAN [92], pierwszy duży, anotowany, publicznie dostępny korpus języka polskiego.

[12] W opisie osiągnięć naukowych będą stosował odnośniki w formacie: autor(rzy) plus rok, tam gdzie odnoszą się one do prac przedstawionych jako podstawowe osiągnięcie. Pozostałe odnośniki wskazują na pozycje w spisie literatury zamieszczonym po Sekcji 5

[13] Bardzo wiele algorytmów w inżynierii języka naturalnego, a nawet szerzej w sztucznej inteligencji, jest rozwijane i testowane ciągle na tych samych, ograniczonych zbiorach danych. Może pojawić się swoisty efekt przeuczenia na podstawie transferu wiedzy o zbiorze danych poprzez publikacje i ich czytelników. Tymczasem prawdziwe problemy pojawiają się przy przejściu do rzeczywistych danych o dużych wolumenach i bardzo zróżnicowanej charakterystyce. Na przykład, większość metod do wydobywania podobieństwa i relacji znaczeniowych jest testowana na słowach występują-

1. Opracowanie lingwistycznego modelu wordnetu języka polskiego definiującego w sposób możliwie precyzyjny i spójny podstawowe elementy wordnetu i jego struktury.
2. Opracowanie zasad procesu budowy wordnetu w oparciu o duży korpus tekstów, zapewniającego efektywne połączenie pracy ręcznej lingwistów z wykorzystaniem narzędzi do wydobywania wiedzy lingwistycznej z tekstów i zwiększającego efektywność pracy i spójność jej rezultatów.
3. Opracowanie szeregu heterogenicznych metod do wydobywania relacji leksykalno-semantycznych z korpusów polskich tekstów – wykorzystujących w efektywny sposób specyficzne cechy języka polskiego oraz różne wyznaczniki językowe relacji leksykalno-semantycznych w tekstach.
4. Opracowanie algorytmów do rozpoznawania i klasyfikacji występowania relacji derywacyjnych języka polskiego pomiędzy poszczególnymi słowami jako realizacji wybranych relacji leksykalno-semantycznych.
5. Opracowanie metody automatycznego rozszerzania wordnetu w oparciu propagację aktywacji w sieci wordnetu z wykorzystaniem heterogenicznych źródeł wiedzy wydobytych z dużych korpusów tekstu.

4.2. Lingwistyczny model wordnetu języka polskiego

Princeton WordNet[14] [18] to bardzo duża leksykalna sieć semantyczna dla języka angielskiego, której budowa została zapoczątkowana w roku 1983. Stał się on największym leksykalno-semantycznym zasobem na świecie, uzyskał tysiące zastosowań badawczych i praktycznych oraz wyznaczył standard *de facto* dla zasobów leksykalno-semantycznych. *WordNet* jest bardzo często postrzegany jako istotny element zbioru BLARK dla angielskiego. Z tych względów dla języków, dla których sieć typu *WordNet* jeszcze nie istnieje, właściwym pytaniem nie jest czy ją budować, tylko jak. W dalszej części niniejszego autoreferatu będziemy leksykalną sieć semantyczną nawiązującą do struktury WordNetu określać mianem wordnetu. W roku 2002 byłem jednym z pomysłodawców budowy wordnetu dla języka polskiego i w roku 2005 stałem się kierownikiem projektu badawczego (3T11C01829, zob. Sekcja J.2) finansowanego przez MNiSW, którego celem było opracowanie metod wydobywania relacji leksykalno-semantycznych z tekstów oraz zbudowanie prototypowego wordnetu języka polskiego, później nazwanego *Słowosieć* (ang. nazwa *plWordNet*).

Na potrzeby realizacji projektu udało się zebrać wybitny interdyscyplinarny zespół specjalistów z zakresu zarówno informatyki, jak i językoznawstwa. Od samego początku Słowosieć miała być nie tylko zasobem językowych (zbiorem danych) na potrzeby przetwarzania, ale możliwie rzetelnym i wiernym obrazem polskiego systemu leksykalnego, co bardzo wyróżniało nasze przedsięwzięcie od większości wordnetów świata budowanych na zasadzie tłumaczenia *WordNetu*. Dlatego też interdyscyplinarność zespołu tworzącego ją oraz twórcze połączenie wysiłku lingwistów i informatyków stały się wyznacznikami działań na rzecz jej konstrukcji poprzez kolejne projekty badawczo-rozwojowe, realizowane bez przerwy po dzień dzisiejszy. W całym tym okresie byłem nieprzerwanie kierownikiem zespołu badawczego budującego i rozszerzającego *Słowosieć*, wyznaczając kierunki jej rozwoju. Miałem ogromny wpływ na jej kształt i zawartość, podejmowałem lub aprobowałem wszystkie kluczowe decyzje.

Pierwszym, z perspektywy czasu naiwnym pomysłem, było podążanie w konstrukcji *Słowosieci* za strukturą *WordNetu*. Szybko się jednak okazało, że byłaby to strategia niewłaściwa, choć stosowana w przypadku ogromnej większości wordnetów dla języków innych niż angielski[15]. Po wnikliwej analizie doszliśmy do wniosku, że definicje podstawowych pojęć dla *WordNetu* są sformułowane w sposób zbyt ogólnikowy, aby można było na ich podstawie sformułować jasne procedury pracy umożliwiających

cych przynajmniej 1 000 razy w korpusie, podczas gdy rzadsze słowa są nieporównywalnie trudniejsze do automatycznego opisu

[14] `http://wordnet.princeton.edu/`

[15] Praktycznie tylko wordnet niemiecki – GermaNet [22] – oraz wordnet duński – DanNet [53] – są oparte na modelach wykazujących unikatowe cechy. W obydwu tych przypadkach modele były opracowywane przy dużym udziale leksykografów i w obu przypadkach wordnety te były budowane od podstaw, a nie tłumaczone z WordNetu

spójne działania zespołu lingwistów nad edycją struktury wordnetu, w tym *Słowosieci*, co pokazaliśmy w [15] (Piasecki, Szpakowicz, and Broda, 2009) i (Maziarz, Piasecki, and Szpakowicz, 2013). Na przykład, pojęcie *synsetu*, który jest podstawowym elementem struktury wordnetu, było definiowane w literaturze jedynie przy pomocy krótkich, jedno lub dwu zdaniowych definicji, czego reprezentatywnym przykładem może być poniższa definicja:

> "a set of synonyms that serve as identifying definitions of lexicalised concepts" [49, p. 5] 'zbiór synonimów, który wykorzystywany jest jako definicja identyfikująca zleksykalizowane pojęcie'

W myśl tej definicji synset i jego znaczenie jest identyfikowane poprzez słowa dobierane ze względu na ich synonimię. Synset w oparciu o jego elementy – słowa – odnosi rodzimego użytkownika danego języka do pewnego zleksykalizowanego pojęcia, które jest współdzielone przez wszystkie elementy synsetu.

Synset bywał też definiowany jako zbiór słów reprezentujących to samo zleksykalizowane pojęcie, a relacje pomiędzy synsetami jako relacje pojęciowe, czyli pomiędzy pojęciami (ang. conceptual relations). [17, p. 210]

Tufiş, Cristea, and Stamou [102, p. 10] określają synsety jako węzły sieci leksykalno-semantycznej, które

> "represent sets of actual words of English sharing (in certain contexts) a common meaning." 'reprezentują zbiory rzeczywistych słów języka angielskiego, które współdzielą (w pewnych kontekstach) wspólne znaczenie'.

Jak można zauważyć, kluczowym elementem tego typu definicji jest pojęcie synonimii. Tymczasem sformułowanie operacyjnej definicja synonimii pozostaje nierozwiązanym, trudnym problemem językoznawstwa, a „zleksykalizowane pojęcie" nie zostało nigdy zdefiniowane w literaturze dotyczącej wordnetów. Ponadto w modelu *WordNetu* nie wprowadzono konsekwentnego rozróżnienia pomiędzy reprezentacją wiedzy (abstrakcją), a opisem semantyki języka naturalnego, na przykład kluczowe relacje pomiędzy znaczeniami leksykalnymi są określane jako „pojęciowe" (ang. „conceptual relations"), chociaż ich nazwy (na przykład hiponimia czy meronimia) oraz definicje jasno nawiązują do aparatu pojęciowego leksykografii.

Już na samym początku prac nad budową *Słowosieci* opracowaliśmy unikatowy, spójny model wordnetu, w ramach którego zaproponowaliśmy definicję synsetu zarówno opartą na tradycji lingwistycznej oraz posiadającą charakter operacyjnym, tj. ułatwiający jej spójne zastosowanie w trakcie budowy wordnetu przez leksykografów. Ponadto, model ten oparliśmy na spójnym systemie ograniczonej liczby dobrze zdefiniowanych pojęć, na przykład (Piasecki, Szpakowicz, and Broda, 2009) i (Maziarz, Piasecki, and Szpakowicz, 2013). Dzięki temu uniknęliśmy powiązania *Słowosieci* z jedną konkretną teorią znaczenia i uczyniliśmy ją bardziej otwartą na różnorodne zastosowania.

Wyszliśmy od założenia, że podstawowym elementem struktury jest *jednostka leksykalna*, a podstawowym narzędziem opisu znaczeń leksykalnych są relacje leksykalno-semantyczne. W ten sposób bezpośrednio nawiązaliśmy do paradygmatu relacyjnego znanego z leksykografii, jednocześnie zachowując podstawową ideę wordnetu, że sieć instancji relacji semantycznych (tj. konkretnych powiązań) definiuje znaczenia leksykalne. Jednostka leksykalna w naszym ujęciu jest rozumiana jako trójka: $\langle p, m, i \rangle$, gdzie p reprezentuje klasę gramatyczną, m – lemat[16], a i to identyfikator znaczenia.

[16] Lemat definiujemy jako arbitralnie wybraną podstawową formę morfologiczną reprezentującą zbiór form wyrazowych różniących się pod względem wartości kategorii gramatycznych, ale posiadających to samo znaczenie, na przykład dla rzeczowników lemat to forma mianownika liczby pojedynczej, a dla czasowników to bezokolicznik.

Synset został zdefiniowany jako zbiór jednostek leksykalnych, które współdzielą *relacje konstytutywne* oraz *cechy konstytutywne*[17]. Relacje konstytutywne to wybrane relacje leksykalno-semantyczne, które są:
— *współdzielone* pomiędzy jednostkami leksykalnymi[18],
— występują dość *często* w języku,
— są dobrze osadzone w *tradycji leksykograficznej*[19],
— *wykorzystywane* są w wordnetach[20],
— i mogą być zdefiniowane w sposób pozwalający na osiągnięcie *zgodności* w pracy lingwistów[21].

Przykładami relacji konstytutywnych mogą być: hiperonimia[22], hiponimia[23], holonimia[24], meronimia, kauzacja (dla czasowników) czy też gradacyjność (dla przymiotniki).

Definicje relacji leksykalno-semantycznych w modelu Słowosieci odwołują się do użycia jednostki w praktyce językowej[25] oraz wykorzystują między innymi takie narzędzia lingwistyczne jak *testy podstawieniowe*[26], dzięki czemu istotnie zmniejsza się rola intuicji językowej lingwistów w podejmowaniu decyzji edycyjnych.

Cechy konstytutywne to wybrane własności semantyczne jednostek leksykalnych, które występują często, są współdzielone pomiędzy jednostkami leksykalnymi oraz są dobrze opisane w tradycji leksykograficznej. Na potrzeby *Słowosieci* przyjęto zestaw czterech cech konstytutywnych: *rejestr stylistyczny*, *aspekt* (dla czasowników), a także *klasy semantyczne* czasowników oraz przymiotników. Każda z cech została zdefiniowana przy pomocy operacyjnej definicji, która odwołuje się do użycia w praktyce językowej oraz zawiera w sobie procedurę lingwistyczną określającą sposób podjęcia decyzji odnośnie wartości cechy. Procedura ta może przybierać dość skomplikowaną postać drzewa decyzyjnego zbudowanego w oparciu o wiele kryteriów szczegółowych. Struktura i kształt tych drzew zostały dopracowane pod kątem osiągnięcia jak największego i jednocześnie akceptowalnego poziomu zgodności między lingwistami [44].

Ponieważ w definicjach relacji konstytutywnych występują odwołania do cech konstytutywnych, na przykład większość relacji może wiązać tylko jednostki o zgodnych rejestrach stylistycznych, to relacje są ściśle powiązane z cechami. Pomimo iż cechy są nierelacyjnym narzędziem opisu, to jednak poprzez wpływ na kształt struktury relacji dobrze wpisują się w relacyjny paradygmat opisu znaczeń wordnetu.

W rezultacie zastosowania naszego modelu, struktura synsetu wynika z rozpoznanej sieci relacji leksykalno-semantycznych, a synset staje się *de facto* skrótem notacyjnym do zapisu faktu współdzielenia przez zbiór jednostek powiązań w sieci.

Zaproponowany model wordnetu ma charakter unikatowy w świecie i do dziś w literaturze światowej trudno znaleźć model o tak rygorystycznie sformułowanym aparacie pojęciowym, konsekwentnej konstrukcji i kompleksowym charakterze. Co więcej, zaproponowany przez nas model został konsekwentnie

[17] Ta kluczowa idea została zaproponowana przeze mnie już w roku 2007, jako podstawa do wyjścia z impasu pomiędzy dążeniem do konstruowania synsetów, a niedookreślonością operacyjną pojęcia synonimii. Idea relacji konstytutywnych pozwala na uniknięcie konieczności opierania się bezpośrednio na synonimii.

[18] Współdzielenie rozumiane jest tutaj jako występowanie dwóch lub więcej jednostek leksykalnych $x_1, \ldots, x_k$ w relacji R do tej samej jednostki leksykalnej y.

[19] Relacje wykorzystywane w leksykografii są lepiej opisane oraz rozumiane przez lingwistów, co powinno ułatwiać ich spójne stosowanie przy budowanie wordnetu.

[20] Wzmacnia to kompatybilność konstruowanego wordnetu z innymi zasobami tego typu.

[21] Czyli jest możliwe opracowanie operacyjnych definicji i procedur pracy opartych na nich

[22] W uproszczeniu, łącząca bardziej ogólną jednostkę leksykalną z bardziej szczegółową.

[23] Relacja odwrotna do hiperonimii.

[24] Złożona relacja z wieloma podtypami, w dużym uproszczeniu, relacja całości do części.

[25] To znaczy do analizy przykładów użycia jednostek leksykalnych, które miałyby być powiązane relacją lub rodzajów kontekstów w których są używane w tekstach. Podstawowym źródłem wiedzy staje się duży korpus tekstów.

[26] Test podstawieniowy zbudowany jest z jednego lub kilku wyrażeń językowych, które zawierają zmienne instancjonowane słowami lub wyrażeniami, które są testowane, na przykład na fakt ich powiązania relacją. Każde z wyrażeń językowych tworzących test jest opatrzone etykietą oczekiwanego stopnia poprawności językowej danego wyrażenia po jego ukonkretnieniu.

wdrożony w konstrukcji *Słowosieci* na niespotykaną wcześniej skalę ilościową, co można również uznać za unikatowy eksperyment naukowy w dziedzinie leksykografii i technologii językowej.

Rozwijana konsekwentnie przez lata struktura relacji *Słowosieci*, zob. na przykład (Piasecki, Szpakowicz, and Broda, 2009), (Maziarz, Piasecki, and Szpakowicz, 2013), (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016), obejmuje obecnie 53 główne typy oraz 107 podtypów na dwóch poziomach: relacji jednostek leksykalnych i relacji synsetów. W pracy [16] zaproponowaliśmy w ramach zbioru relacji dla czasowników wprowadzenie nowej kategorii relacji: *niekonstytutywnych relacji synsetów*, które są współdzielone pomiędzy jednostkami leksykalnymi w synsecie, charakteryzują istotne aspekty znaczeń czasownikowych, ale nie są koniecznymi warunkami w definicji synsetów. Dają one możliwość wzbogacenia wordnetu o cechy pomocne dla wielu aplikacji.

4.3. Proces budowy wordnetu w oparciu o duży korpus tekstów

W ramach projektu EuroWordNet [103] zaproponowano dwie podstawowe metody budowy wordnetu określane jako metody *łączenia* (ang. merge) oraz *transferu* (ang. transfer). Pierwsza z nich zakłada uprzednią dostępność elektronicznych semantycznych zasobów leksykalnych, na przykład elektronicznych słowników lub tezaurusów, które mogą być wykorzystane jako podstawa do budowy wordnetu.

Metoda transferu polega na przetłumaczeniu struktury wordnetu dla innego języka na strukturę wordnetu języka docelowego. Praktycznie we wszystkich dotychczasowych przypadkach jako źródłowym wykorzystywany był wordnet języka angielskiego, czyli *Princeton WordNet*. Transfer polega najczęściej na kopiowaniu struktury relacji i synsetów oraz na tłumaczeniu ich zawartości na język docelowy. W przypadku, gdy tłumaczenie danego synsetu źródłowego jest trudne lub niemożliwe (na przykład z powodu różnic w pokryciu leksykalnym pomiędzy oboma językami), pozostawiany jest pusty synset w strukturze języka docelowego. Już to ostatnie jest rozwiązaniem budzącym wątpliwości z punktu widzenia zasobu opisującego określony język naturalny. Co więcej metoda transferu nie prowadzi do konstrukcji wordnetu, który wiernie opisuje dany język. Struktura relacji, przynajmniej hiperonimii i hiponimii, jest zdeterminowana opisem systemu leksykalnego języka źródłowego, czyli zwykle angielskiego. W przypadku *Słowosieci*, od samego początku jednym z fundamentalnych założeń było zbudowanie jej jako *wiernego opisu języka polskiego*. Dlatego też, na samym początku, zdecydowanie odrzuciliśmy metodę transferu.

Metoda łączenia wymaga dostępności zasobów leksykalnych, które były całkowicie niedostępne w roku 2005, kiedy ruszały prace nad *Słowosiecią* (przynajmniej w jakiejkolwiek bardziej otwartej, naukowej postaci, a korzystanie z płatnych zasobów zapewne ograniczyłoby późniejsze zastosowania *Słowosieci*). Do dziś trudno znaleźć inne semantyczne zasoby leksykalne dla języka polskiego niż *Słowosieć*.

Budowa od podstaw dużego wordnetu jest procesem wymagającym bardzo dużych nakładów pracy. Dlatego od samego początku założyliśmy konieczność wsparcia pracy leksykografów poprzez opracowanie i zastosowanie półautomatycznych metod do wydobywania wiedzy o semantyce leksykalnej z korpusów tekstów. Ze względu na zamierzoną wysoką jakość *Słowosieci* oraz wierność opisu, metody automatyczne były pomyślane jedynie jako narzędzia wspierające pracę leksykografów i uruchamiane pod ich ścisłą kontrolą. Każda decyzja edycyjna była podejmowana zawsze przez leksykografa i każdy element *Słowosieci* jest wynikiem pracy człowieka.

W oparciu o zbudowane narzędzia informatyczne wypracowaliśmy unikatową w skali światowej *metodę korpusową budowy wordnetu*, w której możliwie duży korpus danego języka naturalnego jest podstawowym źródłem wiedzy lingwistycznej. Wraz ze wzrostem wielkości *Słowosieci*, aby zapewnić odpowiednią liczbę przykładów wystąpień coraz rzadszych słów w niej opisywanych, konieczne było powiększanie korpusu (nazwanego *Korpusem Słowosieci*) na którym opierała się jej budowa. W przypadku *Słowosieci 4.0* używany do jej budowy *Korpus Słowosieci 10.0* osiągnął wielkość ponad 4

miliardów wyrazów[27]. Pomimo tak dużego rozmiaru, bardzo wiele z opisanych rzeczowników wystąpiło w nim rzadziej niż 20 razy. Oznacza to, że *Słowosieć* wyróżnia się na tle wielu światowych zasobów leksykalnych bardzo wysokim stopniem pokrycia dla tekstów z wielu dziedzin.

Proces budowy wordnetu w oparciu o metodę korpusową jest realizowany w kolejnych *iteracjach* podzielonych na szereg kroków, przedstawionych poniżej, por. (Maziarz, Piasecki, Rudnicka, and Szpakowicz, 2013), [39], (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016).

1. Wydobycie z korpusu listy k najczęstszych lematów określonej części mowy, które nie są jeszcze opisane w danym wordnecie.
2. Ręczne usunięcie z listy wszystkich elementów, które nie są lematami danego języka[28].
3. Automatyczne wydobycie z korpusu źródeł wiedzy opisujących wybrane lematy, w tym *miarę powiązania znaczeniowego* oraz potencjalne *instancje relacji wordnetowych*.
4. Automatyczne wydobycie z korpusu przykładów ilustrujących potencjalne znaczenia wybranych lematów przy pomocy metody *LexCSD* [5].
5. Automatyczne pogrupowanie lematów z listy w tzw. *paczki* motywowane znaczeniowo w oparciu o miarę powiązania znaczeniowego (Piasecki, Szpakowicz, and Broda, 2009).
6. Automatyczne wygenerowanie propozycji jednostek leksykalnych dla poszczególnych lematów i ich opisu w postaci podgrafów wordnetu za pomocą algorytmu *Paintball* (Piasecki, Ramocki, and Kaliński, 2013).
7. Przydzielanie przez koordynatora paczek do poszczególnych leksykografów pracujących nad edycją wordnetu (krok realizowany stopniowo, w miarę postępu prac).
8. Podejmowanie decyzji edycyjnych przez leksykografów w oparciu o:
   a) jednostki leksykalne zaproponowane dla poszczególnych lematów w ramach systemu *WordnetWeaver* do półautomatycznego konstrukcji wordnetu (w oparciu o algorytm *Paintball*).
   b) automatycznie wybrane przykłady użycia (szczególnie istotne dla czasowników, przymiotników i przysłówków),
   c) listy najbardziej semantycznie bliskich lematów uzyskane na podstawie miary powiązania znaczeniowego,
   d) konkordancje dla edytowanych lematów wyszukane w *Korpusie Słowosieci* i innych dużych korpusach dla języka polskiego,
   e) konsultacje encyklopedii, tezaurusów, dostępnych słowników papierowych i elektronicznych,
   f) intuicję językową leksykografów popartą oceną wyników zastosowania testów podstawieniowych.
9. Edytowanie struktury wordnetu po zastosowaniu testów podstawieniowych właściwych dla wybranej relacji – testy są automatycznie prezentowane i uzupełniane wybranymi lematami w ramach systemu *WordnetLoom* [70, 50].
10. Weryfikacja decyzji leksykografów przez koordynatora zespołu przy wsparciu webowego systemu do monitorowania pracy, zob. [62].
11. Automatyczna, kompleksowa diagnostyka struktury wordnetu w odniesieniu do różnych aspektów [62].
12. Eksport danych z bazy danych produkcyjnej do bazy prezentacyjnej[29] i do kilku formatów dystrybucji wordnetu.

[27] *Korpus Słowosieci 10.0* (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016) składa się z Korpusu IPI PAN [92], pierwszego anotowanego korpusu języka polskiego jaki został zbudowany, danych z Narodowego Korpusu Języka Polskiego [93] (w ramach współpracy w CLARIN-PL), polskiej Wikipedii (od roku 2016), Korpusu *Rzeczpospolitej* [98] – korpusu elektronicznych wydań tej gazety z lat 1993-2003 oraz tekstów pozyskanych z Internetu. W tym ostatnim przypadku były włączane tylko teksty wykazujące ograniczony procent słów nie rozpoznanych przez analizator morfologiczny Morfeusz 2.0 [105] . Przy dołączaniu tekstów do korpusu duplikaty są wykrywane automatycznie i usuwane.

[28] Automatyczne przetwarzanie tekstów korpusu na poziomie morfosyntaktycznym (tj. własności morfologiczno-gramatycznych słów) powoduje, że wiele wyrazów zapisanych błędnie, obcych lub członów nazw własnych pojawia się na listach jako potencjalne lematy języka polskiego.

[29] Widocznej na oficjalnej stronie: `http://plwordnet.edu.pl`

W literaturze trudno znaleźć przykład wordnetu zbudowanego w oparciu o tak rozbudowaną, dopracowaną i systematyczną metodę, jak również trudno znaleźć przykłady innych metod budowy wordnetów, które tak konsekwentnie opierałyby się na danych językowych pozyskanych z korpusów tekstów.

W przeciwieństwie do wielu projektów wordnetowych całość edycji od samego początku odbywała się w ramach systemu *WordnetLoom* [69, 70], który od pierwszej wersji w roku 2005 wyposażony jest w interfejs graficzny i umożliwia zdalną pracę zespołu leksykografów na centralnej bazie danych. Dzięki temu został praktycznie wyeliminowany problem błędów formalnych przy tworzeniu wordnetu.

Proces budowy został zorganizowany w oparciu o wiedzę lingwistyczną wydobywaną z bardzo dużych korpusów tekstów: począwszy od listy lematów, a skończywszy na wiedzy przybliżającej opis znaczeń poszczególnych lematów. Ze względu na edycję coraz rzadszych znaczeń, z korpusu są wydobywane nie tylko zasoby o charakterze statystycznym (na przykład miary powiązania znaczeniowego), ale również potencjalne instancje relacji w postaci par lematów przy pomocy wzorców zbudowanych ręcznie, które wykorzystują nawet pojedyncze wystąpienia par lematów w korpusie. W celu połączenia tak heterogenicznych źródeł wiedzy opracowałem algorytm *Paintball*[30]. Został omówiony bliżej w sekcji 4.6. *Paintball* oparty jest na idei wykorzystania źródeł wiedzy do zainicjowania początkowych pobudzeń w grafie relacji w wordnecie i następnie ich dalszej propagacji w strukturze sieci. Umożliwia to połączenie probabilistycznych źródeł wiedzy o potencjalnych instancjach relacji leksykalno-semantycznych ze źródłami kategorialnymi otrzymanymi w wyniku zastosowania wzorców leksykalno-semantycznych, zob. sekcję 4.4.

Jednostki leksykalne zaproponowane dla poszczególnych lematów przez *Paintball* jako potencjalne ich znaczenia są prezentowane lingwistom za pomocą systemu *WordnetWeaver*, który został zbudowany w oparciu o koncepcję zaproponowaną przeze mnie, na przykład (Piasecki, Szpakowicz, and Broda, 2009). Został on skonstruowany jako rozszerzenie systemu edycyjnego *WordnetLoom*. Jednostki leksykalne proponowane dla nowego lematu są przedstawiane wizualnie jako podgrafy relacyjnej struktury wordnetu. Za pomocą kolorów jest wyrażana informacja o sile dopasowania znaczeniowego danego lematu do określonego miejsca w strukturze relacji wordnetu. Na tej podstawie, lingwista – edytor wordnetu – może bezpośrednio przejść do ewentualnego dodania nowego znaczenia oraz edycji struktury wordnetu.

Według najlepszej mojej wiedzy, *WordnetWeaver* to system unikatowy w skali światowej, a zastosowanie tego typu narzędzia w budowie słowników semantycznych nie ma precedensu. Ponieważ propozycje nowych jednostek leksykalnych są generowane na podstawie kilku różnych metod semantycznej eksploracji korpusów, to *WordnetWeaver* umożliwia wyjście poza tradycyjne przeszukiwanie korpusów na poziomie słów, ich własności morfosyntaktycznych i konkordancji.

System *WordnetWeaver* wykazał swoją dużą przydatność dla częstszych lematów (na przykład o częstości powyżej 100 wystąpień na 2 miliardy słów) oraz szczególnie dla rzeczowników. Przy spadającej częstości opisywanych lematów coraz trudniej było wydobywać automatycznie z korpusów dostatecznie wyraźną informację o ich własnościach semantycznych. Ponadto, bardzo często, wiele znaczeń lematów jest znacznie mniej częstych niż znaczenia dominujące i przez to są one również znaczniej słabiej widoczne w danych wydobytych automatycznie. Dla mniej częstych lematów lub ich znaczeń bardzo ważne stały się narzędzia, których wyniki pracy są bezpośrednio interpretowane i kontrolowane przez lingwistów na poziomie tekstu, takie jak konkordancje czy automatycznie wydobywane przykłady użycia. Na każdym etapie rozbudowy wordnetu konsultowane są również różnego rodzaju tradycyjne źródła wiedzy leksykograficznej jak słowniki czy encyklopedie. W naturalny sposób leksykografowie wykorzystują także swoją kompetencję językową oraz współpracę w grupie nadzorowanej przez koordynatora zespołu lingwistów.

Po zakończeniu całej iteracji (czyli procesu analizy i edycji określonej porcji k najczęstszych, brakujących lematów) proces budowy wordnetu według metody korpusowej może być kontynuowany w

[30] Algorytm ewoluował i zmieniał nazwy przy zachowaniu podstawowej idei działania: (Piasecki, Szpakowicz, and Broda, 2009), [60], [61], [68] i [75].

ramach kolejnej iteracji. Warto zauważyć, że metodę korpusową można łatwo połączyć z metodą łączenia wykorzystując istniejące już uprzednio zasoby leksykalne jako źródła wiedzy o wysokim stopniu pewności (parametr uwzględniany w algorytmie *Paintball*, por. sekcję 4.6).

Zaproponowany proces budowy okazał się bardzo efektywną podstawą pracy zespołu lingwistów. W trakcie rozwoju *Słowosieci* w ciągu ostatnich 13 lat zrealizowano kilkadziesiąt iteracji rozwoju w oparciu o *Korpus Słowosieci*, który został powiększony z początkowego rozmiaru ok. 500 mln. segmentów do ponad 4 miliardów segmentów. *Słowosieć 4.0* osiągnęła rozmiar 191 447 lematów i 288 074 jednostek leksykalnych stając się największym wordnetem świata i jednym z największych słowników języka polskiego w historii.

Kierowany przeze mnie zespół tworzący *Słowosieć* wraz z jej rzutowaniem na *Princeton WordNet* dla języka angielskiego w ciągu ostatnich kilku lat składał się z blisko 30 lingwistów. Łączne nakłady robocze na zbudowanie *Słowosieci* do wersji 4.0 włącznie przekroczyły 40 osobolat. Jest to długofalowy projekt badawczo-rozwojowy, wyjątkowy w skali nie tylko polskiej, ale też światowej nauki.

4.4. Wydobywanie relacji leksykalno-semantycznych z korpusów polskich tekstów

Jednostki leksykalne – znaczenia lematów – są opisane w wordnecie za pomocą relacji leksykalno-semantycznych. W korpusie tekstów nie można obserwować bezpośrednio jednostek leksykalnych, ale można zaobserwować użycia reprezentujących je lematów oraz różnorodne przesłanki wskazujące na występowanie relacji semantycznych pomiędzy lematami (dokładnie pomiędzy ich poszczególnymi znaczeniami). Na ich podstawie można wydobyć automatycznie pary lematów, które wydają się być powiązane określoną relacją, np. hiperonimią/hiponimią, synonimią, holonimią/meronimią i innymi. Przyjąłem założenie, że dla lematu x na podstawie powiązanych z nim lematów y_i oraz rodzajów relacji można spróbować określić liczbę jednostek leksykalnych (znaczeń leksykalnych) x oraz ich położenie w strukturze relacji wordnetu.

Metody wydobywania relacji semantycznych z korpusów tekstów można podzielić na trzy główne grupy metod, ze względu na ich pochodzenie. Są to więc metody oparte na:

1. *semantyce dystrybucyjnej*,
2. wzorcach leksykalno-syntaktycznych,
3. algorytmach maszynowego uczenia się.

Powyższe klasy nie są całkowicie rozłączne, stosowane też są podejścia łączące cechy więcej niż jednej grupy.

Metody żadnej z grup nie dają gwarancji objęcia opisem wszystkich lematów występujących w korpusie, ani też osiągnięcia bardzo wysokiej dokładności opisu. Na przykład metody oparte na semantyce dystrybucyjnej mogą wygenerować opis dla wszystkich lematów, które wystąpiły w korpusie, ale w przypadku lematów o małej liczbie wystąpień błędy w opisie mogą być znaczne. W oparciu o doświadczenie empiryczne zebrane w wielu moich badaniach, możemy oczekiwać, że lematy są dobrze opisane, gdy występują powyżej 200 razy na dwa miliardy słów, mają szansę być dobrze opisane przy liczbie wystąpień pomiędzy 100 i 200, natomiast przy liczbie wystąpień poniżej 100 tylko przy szczęśliwym zbiegu okoliczności (na przykład jednoznaczne i spójne konteksty użycia w tekstach) jest szansa na uzyskanie dobrego opisu. Dobry wgląd we własności metod dystrybucyjnych dają wyniki oceny tych metod przeprowadzone przeze mnie wraz z zespołem w oparciu o *Słowosieć*, na przykład [84, 63].

Rezultaty metod opartych na wzorcach zależą od zwykle mocno ograniczonej liczby wystąpień par lematów w tekstach korpusu. Metody oparte na maszynowym uczeniu w naturalny sposób zależą od doboru przypadków treningowych i ich liczby oraz pośrednio również od częstości wystąpień określonych cech w korpusach. Najczęściej mamy do czynienia z silnym niezbalansowaniem i bardzo znaczącą przewagą negatywnych przypadków nad pozytywnymi. Co gorsza, najczęściej nie ma pewności, czy wśród automatycznie generowanych przypadków negatywnych nie ma błędnych par, ponieważ

w zasobach leksykalnych opisuje się jedynie ustalone powiązania jednostek leksykalnych i nie określa się jawnie braku powiązania. Jeśli wziąć dodatkowo pod uwagę (w odniesieniu do wszystkich metod) błędy narzędzi językowych używanych do wstępnego przetwarzania, to należy oczekiwać, że metody oparte na maszynowym uczeniu generują rezultaty o ograniczonej kompletności i dokładności.

Mając na uwadze naszkicowane powyżej trudności i dążąc do jak najpełniejszego opisu w docelowej wersji Słowosieci, już w 2005 roku rozpocząłem prace nad rozwojem metod ze wszystkich trzech grup (na przykład zob. [56]). Celem było wychwycenie w korpusie możliwie wielu przesłanek opisujących relacje semantyczne pomiędzy lematami, a za ich pośrednictwem znaczenia tych lematów.

4.4.1. Przygotowanie warsztatu – podstawowych narzędzi dla języka polskiego

W przypadku metod wszystkich trzech grup, im bardziej szczegółowym opisem kontekstów użyć poszczególnych lematów dysponujemy, tym lepsze wyniki wydobywania relacji możemy uzyskać. Szczególnie jest to istotne przy mniejszej ilości danych tekstowych, co pokazały jedne z pierwszych naszych wyników testów dystrybucyjnych miar powiązania znaczeniowego dla języka polskiego, w których metody konstrukcji miar wykorzystujące ograniczoną analizę lingwistyczną tekstu osiągnęły znacznie lepsze wyniki [59].

W momencie rozpoczęcia prac w roku 2005 dla języka polskiego z zasobów językowych i narzędzi publicznie dostępne były tylko *Korpus IPI PAN* [92] – anotowany morfosyntaktycznie – oraz analizator morfologiczny *Morfeusz* [105]. Dlatego włączyłem się aktywnie w budowę podstawowych narzędzi dla języka polskiego począwszy od poziomu analizy morfosyntaktycznej. Byłem głównym autorem pierwszego publicznie dostępnego[31] i szeroko stosowanego tagera morfosyntaktycznego dla języka polskiego o nazwie *TaKIPI* (Piasecki and Godlewski, 2006), [57], zob. Sekcja 5.1. Tager *TaKIPI* został oparty na unikatowej kombinacji reguł budowanych ręcznie oraz systemu klasyfikatorów budowanych przy pomocy maszynowego uczenia. Na potrzeby zapisu reguł ujednoznaczniania opracowałem język ograniczeń leksykalno-morfosyntaktycznych o nazwie JOSKIPI (Piasecki, 2006). W języku tym zostały też zapisane ograniczenia, których wyniki działania (logiczne i symboliczne) były wykorzystywane w procesie uczenia i działania klasyfikatorów, jako cechy wysokiego poziomu. Ponieważ brakowało dla języka polskiego parsera składniowego o dobrej dokładności, wyrażenia zapisane w języku JOSKIPI (a później pochodnym od niego języku WCCL [95]), były wykorzystywane do rozpoznawania wybranych relacji składniowych z bardzo dobrą dokładnością, która nie mogłaby być uzyskana przy pomocy dostępnych ówcześnie parserów, por. [73], (Piasecki, Szpakowicz, and Broda, 2009), (Piasecki, Stanisław Szpakowicz, and Broda, 2008) czy (Broda, Derwojedowa, Piasecki, and Szpakowicz, 2008). Zakres działania tagera TaKIPI został później rozszerzony o program do predykcji opisów morfologicznych o nazwie *Odgadywacz* (byłem głównym autorem jego algorytmu i pracy [72]). Miałem też istotny wkład w rozwój narzędzia do rozpoznawania i klasyfikacji wystąpień nazw własnych *Liner* [33, 35, 34]. Ponieważ nazwy własne stanowią otwartą klasę poza leksykonem języka, ich rozpoznanie w tekście może znacząco poprawić jego analizę i wydobywanie wiedzy lingwistycznej z tekstu.

4.4.2. Dystrybucyjne miary powiązania znaczeniowego oparte na dużych korpusach tekstów

Semantyka dystrybucyjna to nurt lingwistyki informatycznej, którego korzenie wywodzą się z tzw. hipotezy dystrybucyjnej, na przykład [23], zgodnie z którą analiza częstości użycia słów i wyrażeń języka naturalnego w różnych kontekstach językowych może przybliżać nas do opisu znaczenia tych słów i wyrażeń. W semantyce dystrybucyjnej kontekst językowy został sprowadzony w większości przypadków do kontekstu tekstowego opisywanego na różnych poziomach analizy języka, np. morfosyntaktycznym lub syntaktycznym. W dalszej części tej sekcji skoncentrujemy uwagę na opisie słów, chociaż metody semantyki dystrybucyjnej mogą być również stosowane do opisu złożonych wyrażeń językowych.

[31] Krótki rys historii rozwoju tagerów dla języka polskiego przedstawiłem w pracy [85]

W semantyce dystrybucyjnej znaczenie słowa X jest reprezentowane za pomocą wektora $\mathbf{x}$, którego składowe odpowiadają cechom opisującym konteksty użyć X, gdzie cecha może być oparta na dowolnym obserwowalnym elemencie struktury językowej tekstu, na przykład formie wyrazowej, lemacie, lemacie opisanym tagiem morfosyntaktycznym, wystąpieniu relacji składniowej, relacji do określonego lematu, itd. Pierwotne wartości cech to zaobserwowane częstości wystąpienia cech w korpusie. Każde słowo ze słownika (lub jego wybranego podzbioru) jest opisywane przez wektor o tej samej długości i oparty na tych samych cechach. Często macierz wszystkich wektorów dla danego zbioru słów nazywa się *macierzą koincydencji* – czyli współwystępowania słów i cech. Ponieważ wektory cech reprezentują występowanie słów w różnych kontekstach, to w oparciu o hipotezę dystrybucyjną zakłada się, że w pewnym stopniu reprezentują one znaczenia leksykalne tych słów. Warto podkreślić, że wektory powstają w wyniku automatycznej (algorytmicznej) analizy korpusu tekstów, a więc w sposób niezależny od interpretacji człowieka. Dlatego też zaletą semantyki dystrybucyjnej jest możliwość wglądu w znaczenia leksykalne w sposób do pewnego stopnia obiektywny, tzn. uwarunkowany jedynie zawartością korpusu.

Jedną z głównych metod eksploracji wektorowej reprezentacji znaczeń leksykalnych jest analiza podobieństwa semantycznego poszczególnych słów (lub bliskości semantycznej). Najczęściej na podstawie wektorowej reprezentacji generowana jest *miara powiązania znaczeniowego* (ang. measure of semantic relatedness) słów (dalej MPZ). MPZ dla dwóch słów X i Y zwraca wartość numeryczną określającą poziom ich semantycznej bliskości i jest obliczana przy pomocy operacji na wektorach reprezentujących znaczenia X i Y. Ponieważ częstości jako wartości pierwotne cech są bardzo podatne na szumy i różnego rodzaju obciążenia wynikające z tekstów korpusu (na przykład różna długość dokumentów, powtórzenia, wyrażenia będące częścią struktury dokumentów itp.), to jeszcze przed obliczeniem podobieństwa wektorów stosowane są różne techniki transformacji wartości cech. Wykorzystuje się również techniki redukcji wymiarów całej macierzy koincydencji. Szczególnie skuteczny okazał się tu algorytm *word2vec* [47] wykorzystujący sieć neuronową, który buduje od razu wektory w mocno zredukowanej przestrzeni i który wykazuje się znacznie wyższą efektywnością działania niż wiele poprzednich metod opartych na redukcji wymiaru. Wektory generowane przez *word2vec* zostały nazwane *wektorami zagłębień słów* (ang. *word embeddings*). Nazwa ta zaczęła być mylnie używana jako synonim semantyki dystrybucyjnej.

Opis za pomocą MPZ może teoretycznie objąć wszystkie słowa występujące w korpusie użytym do jej budowy, co jest jej zaletą. Wartości generowane przez MPZ mogą być użyte na różne sposoby, na przykład bezpośrednio do oceny bliskości pary słów, generowania listy k najbardziej powiązanych znaczeniowo słów do danego słowa X, wyznaczania grup słów wzajemnie bliskich znaczeniowo za pomocą algorytmu grupowania. Wszystkie te sposoby eksploracji mogą być bardzo przydatne i były wykorzystywane w budowie Słowosieci.

Po moich pierwszych wstępnych eksperymentach z budową MPZ dla języka polskiego [56], które były jednymi z pierwszych opublikowanych dla tego języka, zauważyłem, że konieczne jest zarówno wykorzystanie większego korpusu tekstów, jak również wzbogacenie tekstu o wyniki przetwarzania językowego. W odróżnieniu od wielu prac z literatury światowej (w ogromnej mierze zrealizowanych dla języka angielskiego), moim celem było zbudowanie MPZ o możliwie dużym *skutecznym pokryciu* i możliwie wiernym oddaniu relacji leksykalno-semantycznych jako źródła wiedzy lingwistycznej wspierającego proces półautomatycznej budowy wordnetu. Poprzez skuteczne pokrycie rozumiem liczbę lematów opisanych ze skutecznością przydatną dla zastosowań praktycznych. Trudna kwestia oceny MPZ zostanie omówiona poniżej, można jednak powiedzieć, że intuicyjnie jednak oczekujemy, że lemat X jest dobrze opisany w MPZ jeżeli wśród jego k najbardziej powiązanych lematów Y_i znacząca część jest rzeczywiście uznawana przez człowieka za powiązane znaczeniowo z X. Mówimy o znaczącej części, ponieważ oczekiwanie, że przynajmniej połowa z, na przykład, $k = 20$ powiązań lematów nie jest realne w przypadku zastosowania MPZ na większą skalę dla dziesiątek tysięcy lematów. W literaturze wyniki podaje się zwykle dla eksperymentów przeprowadzanych dla najczęstszych lematów występujących więcej niż 1 000 razy, co mocno upraszcza problem, podnosi wyniki, ale jest całkowicie niepraktyczne z punktu widzenia budowy słownika. Tak częste lematy stanowią małą część leksykonu, który ma być

objęty opisem w wielkim słowniku. Informacja zebrana z wielu ich wystąpień oraz duża liczba ich powiązań znaczeniowych powoduje, że automatyczne generowanie dla nich wielu dobrych powiązań znaczeniowych jest relatywnie łatwe. Jednak w dużym wordnecie ogromna większość opisywanych lematów jest znacznie rzadsza, por. (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016). Lematy częste są zwykle opisywane jako pierwsze, co w *Słowosieci* zostało zrobione ręcznie. Dlatego też MPZ jako praktyczne źródła wiedzy ma znaczenie przy opisie lematów występujących znacznie rzadziej, w przeciwieństwie do typowych wyników testów MPZ prezentowanych w literaturze. Ponadto wykorzystując MPZ jako źródło wiedzy do budowy wordnetu oczekujemy, że wyższe wartości MPZ są generowane dla tych par lematów, które są powiązane relacjami leksykalno-semantycznymi wykorzystywanymi w wordnecie. Oznacza to, że wydobyta z tekstu MPZ ma zbliżyć się do idei *miary podobieństwa semantycznego*[32], co podnosi skalę trudności. Wyniki pierwszych eksperymentów pokazały, że jest to zadanie trudne, ponieważ, na przykład, na podstawie analizy samego tekstu bez dodatkowej informacji lingwistycznej uzyskujemy raczej miarę oddającą tematyczne powiązania lematów, niż miarę podobieństwa semantycznego skorelowaną z relacjami leksykalno-semantycznymi. To podobieństwo tematyczne wynika ze współwystępowania lematów w różnych dokumentach tekstowych reprezentujących różne pola lub wątki tematyczne.

Dążenie do wykorzystania większego korpusu oznaczało w roku 2005 konieczność jego budowy we własnym zakresie, ponieważ w tamtym czasie dla języka polskiego dostępny był jedynie *Korpus IPI PAN* [92] o rozmiarze około 270 mln segmentów. *Narodowy Korpus Języka Polskiego* [93], który powstał znacznie później, nigdy nie został udostępniony do badań nad wydobywaniem informacji lingwistycznej poza wąską grupą członków konsorcjum naukowego powołanego do jego budowy. Dlatego też pod moim kierownictwem zespół z Politechniki Wrocławskiej rozpoczął trwające do dziś prace nad budową wspomnianego już badawczego *Korpusu Słowosieci*, który w wersji 10 osiągnął wielkość ponad 4 miliardów segmentów.

Przetworzenie bardzo dużej ilości danych tekstowych w roku 2006 było poważnym wyzwaniem. Brakowało też publicznie dostępnych implementacji większości algorytmów do wydobywania MPZ. Aby temu zaradzić zaproponowałem ideę budowy uniwersalnego systemu do wydobywania modeli semantyki dystrybucyjnej (nie tylko MPZ) o nazwie *SuperMatrix* [7, 6], zob. Sekcja 5.5, którego byłem również przez lata jednym z głównych współautorów. Głównym wykonawcą, szczególnie od strony projektowej i implementacyjnej był ówczesny doktorant, obecnie dr inż. Bartosz Broda. *SuperMatrix* od samego początku jest dostępny na licencji otwartej. Implementuje dziesiątki algorytmów semantyki dystrybucyjnej (między innymi wiele sposobów transformacji macierzy koincydencji oraz wyliczenia podobieństwa wektorów na ich podstawie), umożliwia efektywne rozproszone przetwarzanie na sieci komputerów lub węzłów obliczeniowych oraz wspiera zaproponowaną przeze mnie unikatową metodą wzbogacania opisu zlematyzowanego tekstu o binarne relacje morfosyntaktyczne przy pomocy ograniczeń leksykalno-morfo-syntaktycznych. *SuperMatrix* został też od samego początku ukierunkowany na analizę języka polskiego. Nasza praca [1] dotycząca *SuperMatrix* zdobyła pierwszą nagrodę za najlepszą prezentację podczas międzynarowej konferencji *Computational Linguistics – Applications*.

W sytuacji braku odpowiedniego parsera dla języka polskiego, zaproponowałem w roku 2007 metodę wzbogacania opisu zlematyzowanego tekstu w oparciu o wykrywanie wybranych binarnych relacji syntaktycznych pomiędzy wyrazami tekstu za pomocą ograniczeń leksykalno-morfosyntaktycznych skonstruowanych ręcznie i zapisanych w języku JOSKIPI (tj. języku reguł i ograniczeń, który zaprojektowałem dla tagera TaKIPI). Udało się skonstruować ograniczenia wykrywające z dużą dokładnością takie relacje jak: modyfikacja rzeczownika przez przymiotnik, współrzędne złączenie dwóch rzeczowników lub przymiotników, powiązanie rzeczownika z czasownikiem jako podmiot i predykat (orzeczenie), powiązanie przyimka z rzeczownikiem itd., por. [73] i (Piasecki, Szpakowicz, and Broda, 2009).

[32] W literaturze autorzy najczęściej o miarach wydobywanych z korpusów piszą jako o miarach podobieństwa semantycznego, gdy tymczasem uzyskanie takiej miary jest bardzo trudne i zbliżenie się do tego ideału wymaga odpowiedniej konstrukcji i dostrojenia algorytmu.

Ograniczenia opierają swoją skuteczność na eksploracji kompatybilności form wyrazowych w zakresie wartości kategorii gramatycznych. Z mniejszą skutecznością, ale ciągle z dużą przydatnością działają w przypadku, gdy można się oprzeć jedynie na porządku linearnym wyrazów, na przykład modyfikacja czasownika przez przysłówek lub rzeczownika przez inny rzeczownik w dopełniaczu. W pracy (Piasecki, Szpakowicz, and Broda, 2007) pokazaliśmy znaczącą poprawę jakości wydobywanych MPZ dla polskich rzeczowników przy wykorzystaniu różnych zestawów ograniczeń. Później w pracy (Broda, Derwojedowa, Piasecki, and Szpakowicz, 2008) pokazaliśmy, że metoda ta może zostać rozszerzona na znacznie trudniejszy problem opisu przymiotników. W pracy tej też wykazaliśmy, że występujące w literaturze tezy o niewystępowaniu pewnych kombinacji cech w przypadku przymiotników synonimicznych były przyjęte zbyt pochopnie, a przynajmniej nasze badania w odniesieniu do polskich przymiotników ich nie potwierdziły. W pracy [84] opisaliśmy pozytywne wyniki osiągnięte dla mierzenia podobieństwa znaczeniowego czasowników polskich, co kontrastowało z typowymi podejściami wymagającymi do tego zadania wykorzystania dobrej jakości parserów. W pracy (Piasecki, Szpakowicz, and Broda, 2009) przedstawiliśmy szczegółową syntezę uzyskanych wyników w dziedzinie wydobywania MPZ jako źródeł wiedzy o relacjach leksykalno-semantycznych na potrzeby budowy relacyjnych słowników semantycznych.

W literaturze zostało zaproponowanych kilka sposobów mierzenia jakości MPZ. W pracy [84] dokonaliśmy ich przeglądu i pokazaliśmy, że większość z nich nie mogła być zastosowana w odniesieniu do języka polskiego ze względu na brak wymaganych zbiorów danych, służących do porównania z nimi wydobytych MPZ, lub narzędzi językowych i aplikacji do oceny MPZ w działaniu. Jako punkt wyjścia wybraliśmy *test synonymii oparty na wordnecie* (ang. *WordNet-based Synonymy Test*) (dalej WBST) zaproponowany przez [20]. W metodzie tej na podstawie wordnetu generowany jest test złożony z pytań testowych z jedną odpowiedzią, gdzie pytaniem jest lemat wybrany z wordnetu, a odpowiedzi to $n = 4$ inne lematy wybrane z wordnetu. Poprawną odpowiedzią jest lemat najbliższy znaczeniowo lematowi-pytaniu. Odpowiedzi niepoprawne to lematy niepowiązane z lematem-pytaniem, por. (Piasecki, Szpakowicz, and Broda, 2009). Freitag et al. [20] do wygenerowania par pytanie-odpowiedź użyli synsetów, natomiast odpowiedzi niepoprawne były losowane z pozostałej części wordnetu. W rezultacie możliwe jest wygenerowanie testów zawierających dziesiątki tysięcy pytań. Ponieważ procent synsetów jednoelementowych w *Słowosieci* jest znacznie wyższy niż w *Princeton WordNet*, zaproponowaliśmy modyfikację tego testu nazwaną H-WBST, w której jako odpowiedzi poprawne są traktowane lematy należące do hiperonimów synsetów jednoelementowych, por. [84] i (Piasecki, Szpakowicz, and Broda, 2009). Pozwoliło to na znaczne zwiększenie liczby pytań testowych i pokrycie słownictwa przez test.

Okazało się jednak, że wraz z udoskonaleniem naszych metod generowania MPZ test H-WBST stawał się zbyt łatwy – wyniki naszych miar były zbliżone do wyników osiąganych przez ludzi. Zaproponowałem ideę dalszego rozszerzenia tego testu polegającą na dobieraniu odpowiedzi negatywnych w taki sposób, że lematy stanowiące niepoprawne odpowiedzi, ale które są zbliżone znaczeniowo do poprawnej odpowiedzi, mają większą szansę na włączenie do pytań testowych niż niewłaściwe odpowiedzi odległe znaczeniowo od odpowiedzi poprawnej. Podobieństwo znaczeniowe zostało określone w oparciu o miarę podobieństwa semantycznego wykorzystującą wiedzę zawartą w wordnecie. Wartości podobieństwa stały się podstawą do wygenerowanie rozkładu prawdopodobieństwa na potrzeby generowania elementów pytań testowych. Naszą metodę oceny nazwaliśmy *Enhanced WBST*, czyli *rozszerzony test synonimii oparty na wordnecie*. Test taki okazał się znacznie trudniejszy zarówno dla ludzi jak i MPZ [84] i (Piasecki, Szpakowicz, and Broda, 2009). Pokazaliśmy ostatnio, że może być również z powodzeniem stosowany do oceny modeli dystrybucyjnych opartych na zagłębieniach słów (ang. word embeddings) [63]. Trudność EWBST może być łatwo regulowana poprzez zmianę metody obliczania podobieństwa lematów w oparciu o wordnet – na przykład w pracy [63] metodę tę nieznacznie zmieniliśmy i dostosowaliśmy do współczesnej *Słowosieci* – oraz sposobu generowania rozkładu prawdopodobieństwa wyboru lematu jako odpowiedzi negatywnej. Zaletą metod oceny z rodziny WBST jest to, że są to bardzo obszerne testy, obejmujące dziesiątki tysięcy lematów, a mimo to całość testu jest oparta na

decyzjach ludzi zapisanych w strukturze wordnetu. Testy takie również w naturalny sposób mierzą zdolność MPZ do rozróżniania pomiędzy znaczeniami opisanymi w wordnecie.

Ponieważ testy rodziny WBST są ukierunkowane na mierzenie zdolności do rozróżniania znaczeń, zaproponowałem również prosty *test dokładności obciętej odwzorowania wordnetu* (ang. *Wordnet-based Cut-off Rendering Test*) w pierwszej wersji w pracy [88] oraz w wersji rozszerzonej w pracy [63]. Test dokładności obciętej sprawdza dla każdego lematu X ile par na liście k najbardziej powiązanych z nim lematów pokrywa się z parami lematów, z którymi X jest w jednej z relacji wordnetowych (tj. relacji jednostek leksykalnych lub synsetów). W pracy [63] pokazaliśmy, że obydwa typy testów, tj. WBST oraz dokładności obciętej uzupełniają się. Ten drugi mierzy zdolność MPZ do reprezentowania relacji wordnetowych.

Pracując nad jak najlepszym dostrojeniem MPZ do reprezentowania relacji wordnetowych, zauważyłem, że największym problemem jest zbyt duża zależność od często przypadkowych różnić częstości występowania koincydencji poszczególnych lematów w korpusie, zob., na przykład (Piasecki, Szpakowicz, and Broda, 2009) i [88]. Szczególnie jest to uciążliwe w przypadku rzadszych lematów, gdzie liczba cech je opisujących jest często ograniczona do kilkunastu i pojawianie się wśród nich kilku cech przypadkowo częstszych lub przypadkowych ze swojej natury może generować powiązania z innymi lematami, które nie wynikają absolutnie z systemu leksykalnego danego języka. Oczywiście, prymarnym źródłem tego problemu jest niedostatek danych: zbyt mała liczba wystąpień określonych lematów oraz zbyt mała reprezentatywność korpusu. Poprawienie tego jednak jest bardzo trudne, bo lematy o niższej częstości dominują w *Słowosieci*, zob. [40], i będą dominowały w każdym większym wordnecie. Skoro nie można usunąć samych przyczyn problemu, zaproponowałem dwa różne podejścia łagodzące jego skutki. W pierwszym, nazwanym *transformacja rangowa* (ang. Rank-Based Transformation) [10], (Piasecki, Szpakowicz, and Broda, 2009), wartości cech są zastępowane odwrotnością rang ich istotności w rankingu obciętym dla danego lematu X. Ranking ma charakter porządku częściowego, w którym cechy o zbliżonych wartościach współdzielą pozycje rankingowe. W rezultacie cechy o różnych wartościach pierwotnych, które wynikają z różnic w częstościach w korpusie, otrzymują nowe wartości obrazujące ich istotność dla danego lematu. Metoda ta przyniosła poprawę dla lematów częstszych o większej liczbie cech, ale nie złagodziła problemu przypadkowych powiązań. W podejściu drugim, zaproponowałem ideę *częściowej MPZ* (ang. Partial MSR) [88], w której wartości miary są obliczane jedynie dla tych par lematów, których wektory współdzielą wystarczającą liczbę informatywnych cech. Dla pozostałych par lematów wartość MPZ pozostaje nieokreślona. Minimalna liczba współdzielonych cech oraz parametry algorytmu oceny ich informatywności mogą być ustalane eksperymentalnie na wydzielonym zbiorze lematów.

Dużą zaletą MPZ jest jej kompletność, tzn. miara jest w stanie przypisać wartość powiązania semantycznego dla każdej pary lematów jakie występują w korpusie. Jednak w przypadku rzadszych lematów oraz wartości obliczonej dla pary: częsty lemat – rzadki lemat wartości te mogą być mylące. Wartości MPZ nie mogą być również bezpośrednio interpretowane w kategoriach relacji leksykalnych-semantycznych. Pośród k lematów najbardziej powiązanych z lematem X, jedynie mała część tych powiązań reprezentuje określone relacje, co pokazaliśmy w wielu badaniach, na przykład czy [63]. Oznacza to, że wykorzystanie MPZ jako źródła wiedzy do rozbudowy wordnetu nie jest oczywiste.

W pracy [11] rozszerzyliśmy algorytm *Clustering by Committee* (Broda, Piasecki, and Szpakowicz, 2010). W jego nowej zaproponowanej przez nas wersji został on również dostosowany do specyficznych cech MPZ dla języka polskiego oraz systemu *SuperMatrix*. Algorytm w kilku kolejnych iteracjach wydobywa z MPZ grupy silnie semantycznie powiązanych lematów oraz stopniowo koryguje je i ogniskuje wokół potencjalnych znaczeń budując grupy lematów przypominające synsety. Ujawnia się tutaj kolejne silne ograniczenie MPZ i semantyki dystrybucyjnej w ogólności. Ponieważ modele semantyki dystrybucyjnej oparte są na analizie statystycznej, konteksty występujące częściej, a tym samym częstsze znaczenia, dominują w masie danych. W rezultacie wektory dla polisemicznych lematów reprezentują zwykle co najwyżej kilka, najczęściej 2-3, z wielu możliwych znaczeń. Widać to w analizie ręcznej

$k = 100$ lematów najbardziej powiązanych z danym lematem, na przykład (Piasecki, Szpakowicz, and Broda, 2009). Dotyczy to również obecnie popularnych modeli opartych na zagłębieniach słów, na przykład zob. wyniki testów w naszej pracy [63]. Ograniczenie to jest na tyle silne, że doszedłem do wniosku, iż należy poszukać jeszcze alternatywnych źródeł wiedzy o relacjach leksykalno-semantycznych, które potencjalnie uzupełniłyby niepełny obraz wyrażany przez MPZ.

W oparciu o nasze prace nad wydobywaniem znaczeń na podstawie reprezentacji dystrybucyjnej, opracowaliśmy w dalszej kolejności metodę wydobywania znaczeń leksykalnych bezpośrednio z danych korpusowych przy pomocy grupowania kontekstów. Następnie jądra grup były wykorzystane do konstrukcji zespołów klasyfikatorów do rozpoznawania tych znaczeń w tekstach [5]. Badania te stały się podstawą rozprawy doktorskiej jednego z kluczowych ówcześnie członków mojego zespołu badawczego dr inż. Bartosza Brody.

4.4.3. Wzorce leksykalno-syntaktyczne w wydobywaniu relacji

Hearst [24] pokazała, że przy pomocy stosunkowo prostych wzorców o sile ekspresji wyrażeń regularnych, odwołujących się do tekstu anotowanego na poziomie własności morfosyntaktycznych i prostego, płaskiego podziału na *całostki*[33] można wydobyć pary lematów, których znaczenia są w określonej relacji. Wzorce mogą być zapisane przy pomocy wyrażeń regularnych uruchamianych na symbolicznej reprezentacji anotowanego tekstu. Wzorce Hearst opierając się na identyfikacji w tekście wybranych elementów leksykalnych (takich jak czasowniki, partykuły i spójniki), wybranych własności gramatycznych (na przykład części mowy) oraz fraz wybranych typów (na przykład prostych fraz rzeczownikowych). W każdym wzorcu dwa wyróżnione elementy struktury zdania – rzeczowniki lub proste frazy rzeczownikowe – są identyfikowane jako połączone określoną relacją. Hearst przebadała eksperymentalnie jedynie siedem zaproponowanych przez siebie wzorców dla hiperonimii. Pokazała, że wzorce tego typu charakteryzują się dość wysoką dokładnością, ale ograniczoną kompletnością.

Jako uzupełnienie dla metod semantyki dystrybucyjnej przeanalizowałem kilkadziesiąt potencjalnych wyznaczników leksykalno-syntaktycznych relacji hiperonimii. Podobnie jak w przypadku metod dystrybucyjnych analiza struktury zdania została oparta jedynie na ograniczeniach zapisanych w języku *JOSKIPI* (z mojego tagera *TaKIPI*). Większość wyznaczników okazała się być mało produktywna, gdy została rozbudowana do wzorców o dobrej dokładności. Ostatecznie kilka wzorców zostało uznanych za godne uwagi jako źródła wiedzy. Wyniki zostały przedstawione częściowo w [54] oraz w pełnej formie w rozdz. 4 monografii (Piasecki, Szpakowicz, and Broda, 2009). Niestety wzorce tego typu zależą bardzo od występowania w tekście dość specyficznych zdań o charakterze definiującym. Można je znaleźć w tekstach informacyjnych, na przykład pochodzących z Wikipedii, ale w tekstach beletrystycznych występują rzadko i bardzo często w sposób mylący, na przykład odnoszą się do akcydentalnego znaczenia metonimicznego, wyrażając porównanie lub metaforę. Ze względu na ograniczoną częstość występowania par lematów dopasowanych do wzorców, próby filtrowania wyniku zastosowania wzorców poprzez częstość wydobycia określonej pary lematów lub jej potwierdzenie przez więcej niż jeden wzorzec przynoszą poprawę dokładności, ale za cenę bardzo dramatycznego spadku liczby wydobytych par lematów. W praktyce wzorce te dają cenne wyniki, ale gdy są zastosowane do tekstów określonego typu i brana jest pod uwagę każda wydobyta para. Po rozszerzeniu wzorców o wykorzystanie elementów struktury tekstu, na przykład artykułów z Wikipedii, otrzymałem również dobre rezultaty w wydobywaniu par meronimicznych z Wikipedii [66].

Wzorce leksykalno-semantyczne tworzone ręcznie osiągają stosunkowo wysoką dokładność dzięki swojej szczegółowości, a przez to niskiej kompletności. Kompletność można łatwo zwiększyć poprzez uproszczenie i uogólnienie wzorców, ale to z kolei spowoduje dramatyczny spadek dokładności. Rozwiązaniem może być użycie wielu prostszych, bardziej ogólnych wzorców i wydobycie par lematów potwierdzonych jedynie przez co najmniej kilka wzorców. Pantel and Pennacchiotti [51] zaproponowali

[33] Całostki odpowiadają głównym składnikom zdania, tzw. ang. *chunks*, czyli w pewnym uproszczeniu proste frazy (z wyłączeniem nieuzgodnionych fraz modyfikujących, np. fraz przyimkowych)

algorytm *Espresso* oparty na idei *zdalnego nadzoru* (ang. *remote supervision*), w którym na początku ręcznie definiowana jest poszukiwana relacja poprzez kilkadziesiąt przykładowych par lematów pozostających w tej relacji, a następnie algorytm iteracyjnie generuje i wydobywa uogólnione wzorce leksykalno-syntaktyczne pokrywające pary lematów pozostające w relacji. Później za pomocą wydobytych wzorców wydobywa kolejne pary i proces jest rekurencyjnie powtarzany. Zarówno wydobyte wzorce, jak i pary lematów są oceniane na podstawie częstości występowania w korpusie i siły powiązania ze wzorcami/parami oraz wartościami ich oceny. Zainspirowani tym algorytmem zaproponowaliśmy jego nową, rozszerzoną wersję o nazwie *Estratto* (Kurc, Piasecki, and Szpakowicz, 2010). Przede wszystkim uogólniliśmy algorytm, uniezależniając go od dostępu do wyszukiwarki internetowej o dużym pokryciu, ponieważ takie wyszukiwarki są w większości produktami komercyjnymi o ograniczonej dostępności dla badań. Ponadto uniezależniliśmy również algorytm od dostępności płytkiego parsera (tzw. *chunkera*), którego nie było dla języka polskiego. Wymagało to zmiany algorytmu wydobywania i uogólniania wzorców. Sama reprezentacja wzorców została zaprojektowana w *Estratto* tak, aby maksymalnie wykorzystać ograniczoną dostępną technologię dla języka polskiego, czyli głównie tager oraz pierwsze wersje Słowosieci (jako źródło wiedzy) oraz aby dostosować wzorce do specyficznych cech języka polskiego, tj. bogatej morfologii oraz słabo ograniczonego szyku zdania.

W zastosowaniach praktycznych proces *Espresso/Estratto* najczęściej nie osiąga zbieżności do stabilnych zbiorów wzorców i wydobytych par. Dlatego też konieczne było określenie z góry liczby iteracji algorytmu. *Estratto* wydobywa listę par lematów jako reprezentujących relację zadaną na początku poprzez pary przykładowe. Do każdej pary przypisana jest wartość jej oceny (z końcowej iteracji), co umożliwia zdefiniowanie rankingu i kryterium odcięcia na poziomie minimalnej wartości oceny par.

Espresso/Estratto okazał się podatny na zjawisko dryftu pojęciowego. W ciągu kolejnych iteracji wydobywane pary mogą odbiegać od relacji wyrażonej przez pary przykładowe. Odkryliśmy, że bardzo uważnie trzeba podejść do dobrania par przykładowych, ponieważ ukierunkowują one proces wydobywania oraz parametrów algorytmu, por. (Kurc, Piasecki, and Szpakowicz, 2010). Zarówno pary przykładowe jak i wartości parametrów (na przykład progi odcięcia) powinny być dobierane w odniesieniu do korpusu, na którym jest uruchamiany algorytm. Pomimo ujawnionych wad, *Estratto* wykazał się dobrym stosunkiem precyzji do kompletności i umożliwił wydobycie kilkudziesięciu tysięcy par z korpusu o wielkości ok. 800 mln. segmentów (Piasecki, Szpakowicz, and Broda, 2009).

4.4.4. Algorytmy maszynowego uczenia w rozpoznawaniu relacji

MPZ opisuje numerycznie bliskość semantyczną lematów. Udało nam się ukierunkować jej wartości na pozytywną korelację z relacjami leksykalno-semantycznymi, ale ciągle MPZ nie zapewnia wyraźnego kryterium w identyfikacji instancji relacji (tj. par lematów). Wzorce ręczne zapewniają dużą dokładność, ale małą kompletność, a pary wydobywane przez *Estratto* wykazują w pewien sposób niejednoznaczny opis danej relacji. Jednak w korpusie można zaobserwować cechy, które wskazują zarówno na cechy semantyczne zarówno lematu, jak i na rodzaj opozycji semantycznej pomiędzy dwoma lematami, na przykład liczbę i ogólność cech modyfikujących dany rzeczownik (wskaźnik ogólności), występowanie danego rzeczownika jako dopełniaczowy modyfikator innego (potencjalny sygnał składania się), wzajemne pokrywanie się cech dwóch rzeczowników itd., por. (Piasecki, Stanisław Szpakowicz, and Broda, 2008). Na tej podstawie zidentyfikowałem 17 cech opisujących zarówno pojedyncze lematy jak i pary lematów, które wyliczane są na podstawie dystrybucji tych lematów w korpusie. Część z nich pokrywa się z informacją wykorzystywaną w budowie MPZ, ale również w tych przypadkach cechy uwypuklają wybrane aspekty semantyczne.

Na podstawie zdefiniowanych cech zbudowaliśmy wektory opisujące pary lematów. Pary treningowo-testowe zostały pobrane z pierwszego prototypu *Słowosieci* (a więc zasobu o bardzo ograniczonym rozmiarze i pokryciu). Wektory opisujące zostały wyliczone na podstawie wczesnej wersji *Korpusu Słowosieci*, obejmującego wtedy ok. 700 mln. segmentów. Na tym zbiorze zastosowaliśmy wybrane algorytmy maszynowego uczenia się, w tym C4.5 i przetestowaliśmy je według schematu dziesię-

ciokrotnej walidacji krzyżowej. Pomimo szeregu ograniczeń: stosunkowo małego korpusu, wczesnego prototypu *Słowosieci* i niedużego zbioru treningowo-testowego osiągnęliśmy bardzo dobre wyniki w rozpoznawaniu lematów powiązanych bliską relacją w strukturach wordnetu, tj. hiperonimią lub meronimią w odległości do 2 łuków hiperonimicznych na poziomie 80% dokładności i 70% kompletności, por. (Piasecki, Stanisław Szpakowicz, and Broda, 2008). Relacja zdefiniowana w taki sposób może się wydawać zbyt ogólna, ale dla lingwistów jest to cenna wskazówka, ponieważ zawęża poszukiwania miejsca w wordnecie dla nowej jednostki leksykalnej dla nowego lematu[34],do istniejącego już lematu, z którym nowy ma być powiązany ścieżką o długości do dwóch łuków grafu hiperonimii lub jednego meronimii/holonimii jest bardzo cenną wskazówką dla lingwistów, bo bardzo zawęża obszar wordnetu, w którym należy poszukiwać miejsca dla nowej jednostki leksykalnej, por. (Piasecki, Szpakowicz, and Broda, 2009).

Duża liczba zdefiniowanych cech wyliczanych z coraz to większych wersji *Korpusu Słowosieci* dla coraz to większych zbiorów treningowo-testowych nastręczała szereg problemów obliczeniowych. Metody semantyki dystrybucyjnej dokonujące redukcji wymiarów przestrzeni reprezentacji znaczeń, na przykład *word2vec*, skutkują swoistą kondensacją reprezentacji semantycznej lematu w wektorze o ograniczonym rozmiarze. W szeregu prac pokazano, że za pomocą operacji arytmetycznych na wektorach można wyodrębnić z wektorów dla konkretnych lematów składniki odpowiadające reprezentacji określonych aspektów semantycznych opozycji pomiędzy lematami, na przykład żeńskości, jak w parze *król* i *królowa*. Wiele prac nad wykorzystaniem wektorów lematów jako podstawy opisu par treningowo-testowych w algorytmach rozpoznawania relacji semantycznych za pomocą maszynowego uczenia przynosiło sprzeczne rezultaty. Obok podejść uzyskujących bardzo zachęcające wyniki pojawiły się wpływowe prace sugerujące, że na podstawie par wektorów i wyników operacji na nich (na przykład różnicy) algorytm nie uczy się rozpoznawania relacji semantycznej, a jedynie faktu, że część lematów reprezentuje bardziej ogólne – prototypowe pojęcia. W pracy [12] pokazaliśmy w oparciu o modele wektorowe wydobyte z bardzo dużego *Korpusu Słowosieci 10.0* (obejmującego ponad 4 mld. segmentów) oraz pary treningowo-testowe wydobyte ze *Słowosieci 3.2*, że na podstawie wektorowej reprezentacji jest możliwe rozpoznanie relacji hiperonimii/hiponimii łączących pary lematów. W oparciu o pracę [21] treningowo-testowe pary lematów: $\langle X, Y \rangle$ są reprezentowane jako różnice wektorów dystrybucyjnych: $\mathbf{x} - \mathbf{y}$. Dodatkowo wektory różnicowe przykładów treningowych są grupowane do k grup (gdzie k jest parametrem algorytmu). Grupowanie ma na celu uwzględnienie cech wprowadzanych przez dziedziny semantyczne, do których przynależą lematy. Następnie dla każdej grupy z osobna jest konstruowany osobny klasyfikator. W trakcie testów para lematów jest uznawana za reprezentację wydobywanej relacji, jeżeli chociaż jeden z klasyfikatorów zwrócił pozytywny wynik. Na potrzeby eksperymentów zbiory treningowo-testowe zostały zbudowane w taki sposób, aby wykluczyć przypadek uczenia się przez klasyfikatory rozpoznawania prototypowych elementów, na przykład pary testowe nie przecinały się z treningowymi oraz do negatywnych par testowych zostały włączone pary skonstruowane w taki sposób, że oba elementy pochodzą z różnych par pozytywnych. W przeprowadzonych eksperymentach uzyskaliśmy bardzo pozytywne wyniki dla rozpoznawania hiperonimii, tj. miarę F na poziomie 0,8, zob. [12]. Ponieważ nasze eksperymenty zostały przeprowadzone w analogiczny sposób do wielu eksperymentów raportowanych w literaturze, ale na o wiele większych i bardziej złożonych danych treningowo-testowych (dzięki *Słowosieci* dysponowaliśmy dużo większym wordnetem o znacznie głębszej hierarchii hiperonimii, a także znacznie większym korpusem tekstów), to osiągnięte wyniki są bardzo silnym argumentem za tym, że wektory dystrybucyjnej reprezentacji znaczenia mogą być skutecznie użyte do opisu opozycji znaczeniowych odpowiadających relacjom leksykalno-semantycznym. Co więcej uzyskaliśmy również bardzo pozytywne wyniki dla znacznie trudniejszej relacji jaką jest meronimia.

[34] Nowego lematu czyli jeszcze nie opisanego w strukturze wordnetu.

4.5. Rozpoznawanie i klasyfikacja relacji derywacyjnych języka polskiego

Język polski podobnie jak inne języki słowiańskie jest językiem o bogatej fleksji i słowotwórstwie. Można wyróżnić kilkadziesiąt typów relacji słowotwórczych. Wiele z nich posiada jasną interpretację semantyczną. Stały się one podstawą do zdefiniowania szeregu cennych relacji relacji leksykalno-semantycznych, takich jak *żeńskość*, *augmentatywność* (np. zgrubienia i zdrobnienia), *mieszkaniec*, zob. [37], *pary aspektowe* czasowników, *zawieranie roli: agent, narzędzie, miejsce itp. multiplikatywność* (powtarzanie lub dystrybuowanie akcji), zob. [38, 16] czy *stan/cecha*, zob. [46]. Wzbogacają one opis znaczeń w Słowosieci oraz zwiększają gęstość grafu wordnetowego. Relacje motywowane słowotwórczo można znaleźć w większości wordnetów, jednak *Słowosieć* wyróżnia się pod względem ich liczby, różnorodności, pokrycia ich opisu (tzn. liczby instancji) oraz precyzji ich definicji i konsekwentnie semantycznego charakteru tych relacji (na przykład w niektórych wordnetach odnotowywane są relacje na zasadzie związku pomiędzy formami). Opis relacji motywowanych słowotwórczo ujęty w Słowosieci ma duży wpływ na badania i rozwój wordnetów w zakresie opisu powiązań morfosemantycznych, co przejawia się stosunkową dużą liczbą cytowań naszych prac z tego zakresu.

Wystąpienia relacje wordnetowych motywowanych słowotwórczo są sygnalizowane na poziom związków pomiędzy formami lematów poprzez prefiksy i sufiksy słowotwórcze oraz tzw. *wymiany wewnątrztematowe*. Jednak te morfologiczne środki wyrazu nie są jednoznaczne, tzn. wiele relacji słowotwórczych jest wyrażanych poprzez identyczne afiksy, np. prefiks *-ka* pojawia się zarówno dla form żeńskich (*aktor***ka**) i dla roli narzędzia *wiertar***ka**. Ponadto zdarzają się pozorne związki słowotwórcze pomiędzy lematami, na przykład *pierwiastka* nie jest żeńską formą od *pierwiastek*. Dlatego poprzestawanie na porównywaniu form lematów generuje potencjalne błędy. Co więcej relacje leksykalno-semantyczne motywowane słowotwórczo są zdefiniowane dla określonych znaczeń leksykalnych. Jeżeli dwa powiązane słowotwórczo lematy mają więcej niż jedno znaczenie, to nie znaczy, że odpowiednia relacja leksykalno-semantyczna zachodzi pomiędzy wszystkimi parami znaczeń leksykalnych. Aby odrzucić pozorne związki słowotwórcze, a także rozpoznać poprawne instancje relacji konieczna jest analiza semantyczna znaczeń par lematów.

Pierwszy krok, tzn. wykrycie lub wygenerowanie par lematów powiązanych słowotwórczo, okazał się być tylko pozornie prostym. Dla języka polskiego nie istniało bowiem żadne narzędzie językowe, ani też słownik, które opisywałyby dla danego lematu jego bazę derywacyjną lub derywaty danego lematu, jeżeli takowe istnieją. Dla wielu języków słowiańskich, na przykład [106] dla języka czeskiego, powstały analizatory związków słowotwórczych, które są oparte na skomplikowanych systemach reguł i wzorców tworzonych ręcznie. Osiągają one bardzo wysoką jakość, ale ich zbudowanie wymaga bardzo dużych nakładów roboczych. Dla języka polskiego podobny system lub słownik nie istniał w momencie rozpoczynania prac. Przeniesienie systemu reguł słowotwórczych z innego języka słowiańskiego na język polski nie jest operacją prostą i prawdopodobnie nie zmniejszyłoby nakładów pracy wymaganych do zbudowania analizatora słowotwórczego dla polszczyzny. Nie dysponując odpowiednimi nakładami roboczymi zdecydowałem się na opracowanie metody półautomatycznej budowy analizatora słowotwórczego. Podstawą był częściowy opis relacji leksykalno-semantycznych motywowanych słowotwórczo w *Słowosieci* wersji 1.5 oraz opracowany przeze mnie wcześnie *Odgadywacz* morfologiczny [72]. Celem było wykorzystanie mechanizmu uczenia się sufiksów przez drzewo *a tergo* wewnątrz *Odgadywacza*. Ponieważ reguły słowotwórcze często obejmują przekształcenia form wykraczające poza afiksy, w oparciu o zbiór potencjalnych wymian wewnątrztematowych zestawiony przez dr. Marka Maziarza [77], wprowadziłem mechanizm wstępnego dopasowywania derywatu i jego bazy derywacyjnej poprzez automatycznie ustalaną i stosowaną sekwencję wymian. Ostatecznie, dla każdego przykładu uczącego, tj. pary lematów ze Słowosieci, ustalana jest najpierw sekwencja wymian i kierunek najlepszego dopasowania: od początku lub końca, a następnie przetworzona para jest podawana na wejście mechanizmu uczenia *Odgadywacza*. W rezultacie automatycznie pozyskane reguły słowotwórcze składają się z: informacji o kierunku dopasowania, sekwencji wymian oraz afiksu ustalonego przez *Odgadywacz* [76]. Skonstruowany w sposób

automatyczny *Derywator* osiągnął dobrą dokładność i kompletność rozpoznawania powiązań słowotwórczych przy ograniczonych nakładach roboczych na jego konstrukcję. Zaproponowana przeze mnie półautomatyczna konstrukcja ułatwiła jego dalszy rozwój poprzez powiększanie bazy przykładów par lematów powiązanych słowotwórczo. Do poprawy jakości działania *Derywatora* można też zastosować filtrowanie wyników w oparciu o wzorce morfosyntaktyczne (specyficzne dla określonych związków słowotwórczych) – stosowane dla lematów rozpoznanych przez analizator morfologiczny *Morfeusz* [105] – oraz filtry semantyczne oparte na dziedzinach semantycznych i strukturze hiperonimicznej *Słowosieci* – stosowane dla lematów opisanych już w *Słowosieci*.

Ze względu na wspomnianą już niejednoznaczność wyrażania powiązań słowotwórczych na poziomie morfologii, zaproponowałem metodę semantycznej analizy potencjalnych instancji relacji generowanych przez *Derywator*, przy pomocy zespołu klasyfikatorów wytrenowanych w oparciu o przykładowe instancje relacji ze Słowosieci oraz wiedzę pozyskaną z dużych korpusów tekstów. Problem rozpoznania relacji semantycznej wiążącej dwa słowa jest jednym z dobrze znanych problemów w ramach lingwistyki informatycznej. Tutaj mamy jednak do czynienia z jego trudniejszą wersją, która jest rzadko podejmowana w literaturze. Derywat i jego baza derywacyjna niezwykle rzadko pojawiają się zarówno w tych samych zdaniach w tekście, jak również w prostszych konstrukcjach językowych. Oznacza to, że trudno jest wydobyć cechy lingwistyczne, które bezpośrednio charakteryzowałaby ich wzajemną relację semantyczną. W przypadku większości relacji relacji studiowanych w literaturze, powiązane nimi słowa występują w bliskich kontekstach w korpusie. W przypadku relacji motywowanych słowotwórczo konieczne jest zbieranie z korpusu cech charakteryzujących każdy lemat z osobna, a następnie ustalanie ich wzajemnej relacji w oparciu o analizę ich reprezentacji. Po wstępnych podejściach [76], w pracy [78] przyjęliśmy szeroki zbiór cech opisujących konteksty wystąpień tożsamy z cechami używanymi wcześniej do konstrukcji miar powiązania znaczeniowego, np. (Piasecki, Szpakowicz, and Broda, 2009). Następnie macierz koincydencji lematów i cech została poddana transformacji LDA redukującej wymiar. Dodatkowo każdy lemat został opisany przy pomocy sumy zbioru jego analiz morfologicznych, długości formy oraz odległości Levenshteina pomiędzy danym lematem a każdym z $k = 200$ najczęstszych sufiksów, wyznaczonych na podstawie analizy automatu wewnątrz *Derywatora*. Osiągnęliśmy bardzo dobre wyniki działania zespołu klasyfikatorów dla kilkunastu najczęstszych relacji motywowanych słowotwórczo: dokładność w przedziale $[70 - 95]\%$, kompletność $[62 - 98]\%$ oraz miarę F $[66 - 97]\%$.

Całość systemu, *Derywator* oraz semantyczny filtr, została wdrożona w ramach prac nad Słowosiecią do generowania potencjalnych instancji relacji motywowanych słowotwórczo, które później były oceniane ręcznie przez leksykografów przed włączeniem do Słowosieci.

4.6. Automatyczne rozszerzanie wordnetu w oparciu propagację aktywacji

W wyniku przedstawionego powyżej konsekwentnie realizowanego programu badawczego, opracowaliśmy system narzędzi do wydobywania z korpusu wiedzy, dotyczącej znaczeń reprezentowanych przez lematy oraz wiążących je relacji leksykalno-semantycznych. Z natury rzeczy rezultaty poszczególnych narzędzi mają różny charakter (na przykład symboliczny, numeryczny lub probabilistyczny), są częściowe i obarczone często wysoką niepewnością. Na potrzeby dalszych rozważań będziemy nazywać rezultat zwrócony przez pojedyncze narzędzie *źródłem wiedzy*. Pojedynczy lemat X może być opisany przez kilka źródeł wiedzy. Interesującym wyzwaniem badawczym okazało się automatyczne określenie opisu danego lematu X w strukturze wordnetu w oparciu o wydobyte źródła wiedzy.

Głównym celem opracowanej metody było zapewnienie automatycznego wsparcia przy budowie wordnetu, szerzej, leksykalnej sieci semantycznej. Ponieważ jakość wordnetu była zawsze absolutnym priorytetem dążyłem do opracowania rozwiązań półautomatycznych, czyli wspierających pracę leksykografów, których decyzje były zawsze ostateczne. Po wstępnej analizie zdecydowałem się, aby skoncentrować wysiłek nie na tyle na metodzie automatycznej konstrukcji wordnetu, co na metodzie automatycznego rozszerzania wordnetu o nowe lematy i znaczenia. Były po temu dwa istotne powody. Po pierwsze,

na szczycie hiperonimicznej hierarchii w wordnecie znajdują się znaczenia bardzo abstrakcyjne i ogólne, które byłoby bardzo trudno automatycznie opisać na podstawie danych korpusowych. Ponadto decyzje odnośnie struktury wordnetu na tych poziomach kształtują całość struktury wordnetowej, wyrażają pewną koncepcję leksykograficzną i nie da się ukryć, że są w pewnym stopniu subiektywne. Po drugie, górna część struktury wordnetu jest budowana ręcznie na dość wczesnym etapie konstrukcji jako część procesu koncepcyjnego. To, co zajmuje gros czasu i wysiłku, to rozbudowa sieci o dziesiątki tysięcy znaczeń leksykalnych. Przyjęta przeze mnie perspektywa jest również rzadko spotykana w literaturze. Większość zaproponowanych metod, w odróżnieniu od podejścia opracowanego przeze mnie, to metody automatycznej konstrukcji sieci semantycznych, które były oceniane przez porównanie z wordnetami, ale nie były stosowane w praktyce leksykograficznej jako narzędzie wspierające budowę sieci.

Większość czołowych metod automatycznej konstrukcji sieci semantycznej można podzielić na dwie główne grupy:

— oparte na wzorcach leksykalno-syntaktycznych,

— oparte na reprezentacji probabilistycznej powiązań znaczeniowych.

Metody z pierwszej grupy są stosowane najczęściej w obszarze uczenia się lub wydobywania ontologii z tekstów i wykazują wszystkie ograniczenia i wady metod opartych na wzorcach, o których wspomnieliśmy w sekcji 4.4.3. Mogą być stosowane tylko dla tekstów o charakterze informacyjnym. Podejścia z drugiej grupy wymagają dużych ilości danych do wiarygodnego oszacowania rozkładów prawdopodobieństw. W przypadku rozszerzania wordnetu o coraz to rzadsze lematy jest bardzo trudno zapewnić tak duże ilości danych tekstowych, pomimo iż *Korpus Słowosieci* osiągnął wielkość ponad 4 miliardów segmentów w wersji 10 (zbudowanej na potrzeby konstrukcji *Słowosieci 4.0*). Mając świadomość powyższych ograniczeń, opracowałem metodę półautomatycznego rozszerzania wordnetu o charakterze heurystycznym, która umożliwia wykorzystania bardzo heterogenicznych i częściowych źródeł wiedzy. Finalna postać algorytmu została nazwana *Paintball* (Piasecki, Ramocki, and Kaliński, 2013), natomiast wcześniejsze wersje rozwojowe, [60, 61] i (Piasecki, Szpakowicz, and Broda, 2009), funkcjonowały najczęściej jako *WordnetWeaver*, czyli system wspomagający pracę leksykografów (były zaimplementowane w nim).

Wejściem do algorytmu *Paintball* jest zbiór lematów, które nie są jeszcze opisane w wordnecie[35]. Jego zadaniem jest wskazanie dla każdego lematu X synsetów w strukturze wordnetu, do których powinny być przypisane poszczególne jednostki leksykalnego reprezentujące znaczenia lematu X jako elementy danego synsetu lub jako nowe synsety powiązane z istniejącymi jedną z relacji synsetów. Przyjąłem założenie, że każde źródło wiedzy wydobyte z korpusu może być przedstawione jako zbiór trójek: $\langle X, Y_i, w_i \rangle$, gdzie X to jeden z lematów wejściowych, Y_i to lemat już obecny w strukturze wordnetu, a w_i waga – wartość z przedziału $(0, 1]$ – przypisana do danej pary. Algorytm nie zakłada żadnej określonej interpretacji źródłowej wag. Zakładamy, że wagi określają w pewien sposób siłę powiązania semantycznego pomiędzy X i Y_i. Waga liczbowa może być również przypisana do całego źródła wiedzy charakteryzując globalnie jego jakość lub poziom ufności.

W skrócie, *Paintball* jest oparty na ogólnym schemacie propagacji pobudzenia w sieci, tu grafie relacji wordnetowych, zob. (Piasecki, Ramocki, and Kaliński, 2013). Dla wejściowego lematu X każda pojedyncza trójka, w której X występuje to pojedyncza 'kropla farby', która upuszczona (lub rzucona) na sieć rozpływa się wzdłuż łuków relacji. W każdym węźle grafu odkład się część pobudzenia. Proces rozchodzenia się pobudzenia kończy się, gdy jego propagowana część spada poniżej przyjętej wartości granicznej (parametr algorytmu). Dla danego X na sieć wrzucane są kolejne pobudzenia wejściowe i ich wartości są sumowane w wyniku propagacji w poszczególnych węzłach. Semantyczna interpretacja grafu wordnetowego znalazła swoje odzwierciedlenie we wprowadzonych pojęciach *impedancji* i *transmitancji*. Pierwsze z nich jest reprezentowane jako wagą łuku grafu zależną od relacji wordnetowej jaka jest

[35] Algorytm może być zastosowany do lematów już opisanych, częściowo lub całkowicie, w wordnecie. Wtedy jego działanie może być narzędziem diagnostycznym. Wskaże brakujące znaczenia lub opisy znaczeń różne od już wprowadzonych.

reprezentowana przez dany łuk, na przykład powiązania hiperonimiczne mają niską impedancję, dzięki czemu duża ilość pobudzenia przepływa przez nie w górę struktury hiperonimicznej. Drugie pojęcie – transmitancja – jest określona dla par relacji wordnetowych i wyraża do jakiego stopnia pobudzenie może przepływać na styku łuków reprezentujących różne relacje. Dzięki transmitancji można wpływać na kształt ścieżek, przez które jest propagowane pobudzenie, na przykład blokować przejścia pomiędzy hiperonimią i antonimią lub też osłabiać możliwości propagacji poprzez długie łańcuchy meronimiczne. Zastosowany w *Paintball* algorytm propagacji jest stosunkowo prosty, jednak badania eksperymentalne, w których porównaliśmy go z innymi algorytmami propagacji, w tym rekurencyjnymi, na przykład PageRank, pokazały wyższość mojego oryginalnego rozwiązania. Prawdopodobnie spowodowane jest to naturą zadania oraz semantyką struktury grafowej wordnetu.

Dla każdego analizowanego lematu X, *Paintball* działa w dwóch etapach. W pierwszym następuje propagacja pobudzenia i ustalane są stany węzłów grafu. W etapie drugim wyznaczane są spójne podgrafy, w których wszystkie węzły wykazują wynikowe pobudzenie powyżej określonego progu (parametr metody). Następnie dla każdego wyznaczonego podgrafu określany jest węzeł o maksymalnym pobudzeniu i na podstawie tych wartości wyznaczane są podgrafy najlepiej dopasowane do X.

Podgrafy wyznaczone przez *Paintball* dla lematu X są następnie prezentowane interaktywnie leksykografom w systemie *WordnetWeaver* (Piasecki, Szpakowicz, and Broda, 2009) w oparciu o opracowaną przeze mnie wizualizację i projekt interfejsu użytkownika. Każdy podgraf opisuje potencjalną jednostkę leksykalną dla X, czyli potencjalne znaczenie X ustalone przez *Paintball* w odniesieniu do źródeł wiedzy wydobytych z korpusu oraz aktualnej struktury wordnetu. W zamierzeniu każdy podgraf przedstawia obszar wordnetu, w ramach którego jednostka leksykalna dla X może zostać dołączona do jednego z synsetów lub też może zostać utworzony nowy synset zawierający tę jednostkę i połączony bezpośrednią relacją do jednego z synsetów podgrafu. Zdecydowałem się na przedstawienie sugerowanych jednostek leksykalnych za pomocą podgrafów, a nie pojedynczych synsetów, aby:

— pokazać leksykografowi szerszy kontekst opisujący sugerowane znaczenie X,
— odzwierciedlić niepewność jaką charakteryzują się źródła wiedzy – jednocześnie w podgrafach poszczególne węzły opisane są kolorami odzwierciedlającymi poziom pobudzenia jaki został ustalony, pozwala to na porównywanie również podgrafów pomiędzy sobą,
— ułatwić leksykografowi podjęcie decyzji edycyjnej w innym miejscu niż synset podgrafu o najwyższym pobudzeniu.

Wordnet jest podwójnym grafem o dwóch warstwach: jedną tworzy graf synsetów powiązanych relacjami synsetów, ale drugą tworzy graf jednostek leksykalnych powiązanych relacjami jednostek. W każdym wordnecie obydwie warstwy są powiązane poprzez przynależność jednostek leksykalnych do synsetów. W wielu wordnetach warstwa jednostek jest słabo powiązana i graf ten ma małą gęstość w sensie średniej liczby łuków grafu na jednostkę leksykalną. W przypadku *Słowosieci* graf relacji jednostek jest znacznie gęstszy[36]. Dlatego relacje jednostek leksykalnych są bardzo ważne w opisie znaczeń oraz są potencjalnie bardzo istotne dla algorytmów opartych na propagacji pobudzenia w grafie. W większości metod budowy lub rozszerzania wordnetów wszystkie operacje są prowadzone na grafie synsetów. Również we wcześniejszych wersjach mojego algorytmu przypisanie pobudzenia i jego propagacja odbywała się na grafie synsetów, a relacje jednostek leksykalnych były heurystycznie rzutowane na graf synsetów, [60, 61] i (Piasecki, Szpakowicz, and Broda, 2009). Utrudniało to jednak efektywne wykorzystanie wszystkich relacji, dostrojenie wag, a prostota i przejrzystość modelu gubiła się.

Proste rozwiązanie, polegające na przeniesieniu wszystkich relacji na poziom jednostek (na co model *Słowosieci* pozwala), spowodowało, że wygenerowany w ten sposób jednowarstwowy graf stał się bardzo duży, ścieżki bardzo wydłużyły się, a, co za tym idzie, wzrósł znacząco czas obliczeń, co z

[36] Warto również przypomnieć, że w przypadku *Słowosieci* nie ma różnicy w statusie ontologicznym relacji jednostek i relacji synsetów. Te drugie są skrótami notacyjnymi sygnalizującymi występowanie określonej relacji pomiędzy wszystkimi jednostkami z dwóch synsetów.

kolei skomplikowało to dostrajanie algorytmu. Ostatecznie przyjąłem rozwiązanie pośrednie, w którym wordnet jest rzutowany na graf jednostek leksykalnych, ale relacje synsetu są przypisywane jego głowie – czyli pierwszej jednostce leksykalnej synsetu – natomiast jednostki jednego synsetu są powiązane skierowanymi relacjami synonimii. Mechanizmy transmitancji i impedancji zostały wykorzystane, aby czasami długie niekiedy łańcuchy synonimii nie powodowały straty pobudzenia w ramach jednego synsetu oraz aby nie powstawały pętle synonimiczne w ramach jednego synsetu (Piasecki, Ramocki, and Kaliński, 2013). Pokazało to, że zaproponowany stosunkowo prosty model może być łatwo dostosowany do różnych realizacji przetwarzania.

W literaturze bardzo trudno znaleźć informację o algorytmie wspomagającym budowę wordnetu, który byłby tak kompleksowy, w pełni zaimplementowany, wdrożony oraz opierałby swoje działania na wykorzystaniu tak różnorodnych źródeł wiedzy, na przykład systemy wspomagające budowę ontologii opierają swoje działanie głównie na wydobywaniu wiedzy za pomocą wzorców. Algorytm *Paintball* (tj. jego kolejne wersje rozwojowe) został zaimplementowany i wdrożony w praktyce leksykograficznej w ramach systemu *WordnetWeaver* jako narzędzie wspomagające pracę leksykografów. *WordnetWeaver* to rozszerzenie systemu *WordnetLoom* do rozproszonej edycji wordnetu przez zespół leksykografów. Dzięki temu leksykografowie mogę płynnie przejść od prezentacji sugerowanych jednostek leksykalnych dla lematu X do jego pełnej edycji w strukturze wordnetu. *WordnetLoom* wraz ze skojarzonymi narzędziami wspiera w pełni proces budowy wordnetu oparty na korpusie [39] i jest stosowany nieprzerwanie od 2005 roku. *WordnetWeaver* to najbardziej zaawansowane narzędzie do semantycznej eksploracji danych korpusowych. Można było często zaobserwować sytuacje, w który znaczenia podpowiadane przez *WordnetWeaver* zwracały uwagę leksykografów na różne aspekty znaczeń edytowanych lematów, które mogły umknąć ich uwadze. Ponieważ *Paintball* opiera się na możliwości wydobycia wiedzy z korpusów, to jego użyteczność spada, gdy przechodzimy do rzadszych lematów. Można było zaobserwować, że gdy liczba wystąpień danego lematu jest niższa niż 100, to trudniej uzyskać jego dobry opis w *WordnetWeaver*. Dlatego wraz z rozwojem *Słowosieci* (opis w wersji 4.0 sięgnął rzeczowników występujących rzadziej niż 20 razy w dwumiliardowym korpusie) znaczenie *WordnetWeaver* malało na rzecz prostszych narzędzi jak na przykład systemu do wydobywania przykładów użycia (omówionych wcześniej).

Kolejne wersje algorytmu *Paintball* zostały również poddane ocenie empirycznej i porównane z czołowymi metodami z literatury. Ocena tego typu metod nie była jednoznaczna. Ostatecznym celem takich algorytmów jest wygenerowanie rozszerzonego wordnetu. Porównywanie całych sieci: rozszerzonej i wzorcowej jest dość skomplikowane. Ponadto w decyzjach edycyjnych leksykografów jest zawsze pewien aspekt subiektywności[37], co rzutuje na strukturę sieci. Dlatego zdecydowaliśmy się dokonać oceny na poziomie trafności poszczególnych sugestii algorytmu oraz zbioru sugestii dla poszczególnych lematów. Zaproponowany został schemat oceny oparty na usunięciu testowego lematu X i, następnie, próbie przywrócenie informacji o nim automatycznie [60, 61] i (Piasecki, Szpakowicz, and Broda, 2009). W pierwszym kroku dla lematu X zostają usunięte ze struktury wordnetu wszystkie jego jednostki leksykalne oraz następnie ewentualne powstałe puste synsety (który były singletonowe). Propozycje algorytmu są oceniane pod względem odległości dołączenia od miejsca pierwotnego jednostki leksykalnej X. Proces oceny jest przeprowadzany dla kolejnych lematów z osobna, aby poprzez usuwanie informacji z wordnetu w jak najmniejszym stopniu zaburzać jego strukturę, która jest wykorzystywana w działaniu algorytmu. Zaproponowana metoda oceny została rozwinięta do pełnej formy w pracy [4]. Od podobnych metod zaproponowanych w literaturze wyróżnia się nie tylko bardziej kompleksową i dogłębną informacją o skutkach automatycznego rozszerzania w odniesieniu do struktury wordnetu, ale również o ich znaczeniu dla leksykografa. Ocena jest przeprowadzana zarówno na poziomie wszystkich sugestii jak również zbiorów sugestii dla poszczególnych lematów. W tym drugim przypadku wyróżniono dwa tryby oceny sugestii, jako *położonych najbliżej* oryginalnych jednostek leksykalnych (im

[37] Struktura wordnetu jest abstrakcją narzuconą na ciągłość systemu znaczeń leksykalnych danego języka i na pełną, kontekstową interpretację możliwych użyć jego wyrażeń. Dlatego nie możemy oczekiwać pełnej zgodności i powtarzalność działań leksykografów.

proponowana jednostka leksykalna jest bliższa odpowiedniego miejsca, tym jest to wartościowsza sugestia dla leksykografa) oraz *najwyżej ocenionych* przez algorytm (one są proponowane jako najbardziej wyróżniające się). Ponieważ trudno zakładać, że większość sugestii będzie trafiać idealnie w punkt, algorytm oceny umożliwia również tryb pozytywnej oceny z dokładnością do długości i typu ścieżki w grafie wordnetowym łączącej sugerowane miejsce i oryginalne miejsce jednostki leksykalnej, na przykład rozpatrywane są pozytywnie tylko sugestie w odległości do 3 łuków według ścieżki składającej się tylko z łuków hiponimicznych – czyli rozpatrywane są pozytywnie tylko bardziej specyficzne sugestie niż oryginalna jednostka leksykalna.

Zaproponowana metoda oceny może być dostosowywana pod kątem oczekiwań leksykografów. W oparciu o nią porównaliśmy *Paintball* z czołowymi podejściami w literaturze zarówno dla języka polskiego [68], jak i dla języka angielskiego (Piasecki, Ramocki, and Kaliński, 2013). W pierwszym przypadku wykorzystaliśmy *Słowosieć* oraz duże korpusy języka polskiego, w drugim *Princeton WordNet* i korpus *Wikipedii*. W porównaniach empirycznych szczególną uwagę poświęciliśmy metodzie z prac [101] i [100], która nie tylko wykazywała lepsze wyniki niż metody konkurencyjne[38], ale również charakteryzowała się kompleksowym modelem rozszerzania wordnetu, jednak opartym na założeniu probabilistycznego charakteru wszystkich źródeł[39]. Pomimo konieczności zastosowania ograniczonego zestawu źródeł wiedzy – metoda [101] może działać jedynie na źródłach o charakterze probabilistycznym – uzyskaliśmy wyniki zbliżone do niej dla częstych lematów, istotnie lepsze dla rzadszych oraz wyniki znacznie lepsze jeżeli wziąć pod uwagę sposób umiejscawiania sugerowanych jednostek leksykalnych. Metoda [101] wykazuje tendencję do ogólnego klasyfikowania nowych lematów poprzez przypisywanie ich do bardziej ogólnych synsetów wordnetu, podczas gdy *Paintball* proponuje dodawanie nowych jednostek leksykalnych w bliskim otoczeniu ich oryginalnego położenia. Wykazaliśmy w ten sposób, że *Paintball* jest lepszą podstawą do konstrukcji narzędzia leksykograficznego wspomagającego rozbudowę wordnetu.

Ogólne uwagi o współautorstwie

Mam niewiele publikacji, których jestem jedynym autorem. W swojej pracy naukowej kierowałem się cały czas ideą uprawiania nauki w sposób bezpośredni i służebny oddziaływujący na środowisko w którym żyję: naukowe, społeczne i kulturowe. Budowa bardzo dużych zasobów językowych wymaga współdziałania zespołów ludzi, budowa narzędzi językowych wiąże się z koniecznością przeprowadzania eksperymentów w dużej skali, najczęściej poprzedzonych przygotowaniem testowo-treningowych zasobów językowych, i późniejszej implementacji narzędzi na poziomie gotowości technologicznej, umożliwiającym szerokie zastosowania. Przy ogromnej złożoności problemu automatycznej analizy języka naturalnego, dopiero w momencie zastosowania technologii językowej do rzeczywistych danych językowych (tekstów, dokumentów, zapisów wypowiedzi, wiadomości itd.) w praktycznym wymiarze można poznać rzeczywistą wartość stworzonego dzieła.

Począwszy od pierwszego projektu naukowego w dziedzinie inżynierii języka naturalnego, w który się zaangażowałem w roku 2005, konsekwentnie budowałem i utrzymywałem powiększający się zespół naukowy, który umożliwił nam zmierzenie się z wyzwaniami ograniczającymi rozwój tej dziedziny w Polsce. Ogromna większość problemów wymagała pracy zespołowej. Przyjęliśmy zasadę uwzględniania wśród autorów publikacji każdego, kto wniósł chociażby najmniejszą cząstkę do jej powstania, na przykład programistów konstruujących program do przeprowadzenia eksperymentów. Bez części nie ma całości.

W przypadku prac składających się na cykl przedstawiony jako osiągnięcie badawcze mój wkład w każdą publikację wynika z oświadczeń Współautorów. W przypadku pozostałych prac ogromna większość nie powstałaby bez mojej inicjatywy i zaangażowania, w bardzo wielu też odgrywałem bardzo istotną rolę jako współautor.

[38] Ponadto praca ta zdobyła nagrodę Best Paper Award na bardzo prestiżowej światowej konferencji ACL 2006.

[39] W implementacji tej metody jednak jej autorzy odeszli od jej modelu formalnego na rzecz heurystycznych uproszczeń.

5. Zrealizowane oryginalne osiągnięcie projektowe, konstrukcyjne, technologiczne lub artystyczne

W ramach rozprawy doktorskiej obronionej w roku 2003 koncentrowałem się na opracowaniu formalnego modelu znaczenia dla wyrażeń języka polskiego. Model ten był pomyślany jako fundament dla systemów rozumienia języka naturalnego. Niestety, okazało się, że w roku 2003 konstrukcja takiego systemu natrafiła na potężną barierę w postaci niedostatecznego poziomu rozwoju otwartej technologii językowej dla języka polskiego. Dlatego, począwszy od roku 2004 jako misję programu moich badań przyjąłem działanie na rzecz budowy publicznie dostępnej technologii językowej dla języka polskiego umożliwiającej szerokie spektrum naukowych i praktycznych zastosowań. Technologia taka powinna się charakteryzować: szerokim pokryciem opisu języka polskiego na różnych jego poziomach oraz dokładnością, kompletnością i niezawodnością narzędzi językowych umożliwiających szereg ich zastosowań. W efekcie zaangażowałem się na przestrzeni lat w budowę dziesiątków zasobów i narzędzi językowych, jak również systemów do analizy języka naturalnego. Stanowią one bardzo istotną część moich osiągnięć naukowych z dwóch powodów:

— ich zbudowanie wymagało często opracowania wielu unikatowych rozwiązań,
— stały się katalizatorem umożliwiającym szereg dalszych badań i przyspieszających rozwój całej dziedziny badań.

W swoich pracach badawczo-rozwojowych konsekwentnie starałem się rozwiązywać problemy analizy języka naturalnego od poziomu opisu wyrazów po płytką, szybką analizę semantyczną tekstu. Pod względem skali rozwiązań przeszedłem od poziomu pojedynczych programów do kompleksowej infrastruktury badawczej CLARIN-PL[40], która zgromadziła i powiązała szerokie spektrum rozwiązań dla języka polskiego w ramach jednej otwartej platformy.

5.1. *TaKIPI* – tager morfosyntaktyczny języka polskiego

TaKIPI[41] to pierwszy publicznie dostępny tager morfo-syntaktyczny języka polskiego. Dokonuje analizy morfologicznej tekstu, który został wcześniej poddany analizie morfologicznej przy pomocy programu *Morfeusz* [105]. Następnie dla każdego słowa wybiera analizę, która jest właściwa dla kontekstu jego wystąpienia. Opis wyjściowy słowa obejmuje: lemat, klasę gramatyczną oraz wartości kategorii gramatycznych. Opis jest zgodny z formatem *Korpusu IPI PAN* [92]. *TaKIPI* umożliwia również *lematyzację*[42] polskich tekstów, tj. inteligentne, kontekstowe sprowadzenie wszystkich wyrazów do morfologicznych form podstawowych. Tager został później rozszerzony o *Odgadywacz* morfologiczny [72], co umożliwiło również rozpoznawanie form nieznanych *Morfeuszowi* (na przykład form obcych, neologizmów, form z literówkami).

Algorytm działania *TaKIPI* został oparty na koncepcji warstwowego, iteracyjnego ujednoznaczniania kolejnych części złożonych tagów (na przykład warstwa klasy gramatycznej, liczby i rodzaju, przypadka, itp.) oraz oryginalnej koncepcji podzielenia problemu tagowania na tzw. klasy niejednoznaczności. Dla każdej klasy niejednoznaczności były budowane osobne klasyfikatory oparte na drzewach decyzyjnych, dla których konstrukcji zaproponowałem unikatowe rozwiązanie zwiększające ich siłę ekspresji, oparte na wprowadzeniu atrybutów odwołujących się do wyników działania tzw. operatorów leksykalno-morfo-syntaktycznych, które były uruchamiane na tekście, (Piasecki and Godlewski, 2006), [57]. Operatory zostały napisane przeze mnie w języku ograniczeń morfosyntaktycznych mojego autorstwa, o nazwie *JoSKIPI* (Piasecki, 2006).

Zarówno język ten, jak i jego implementacja stały się potem podstawą na lata do budowy programów do wykrywania relacji syntaktycznych na poziomie wyrazowym zastosowanych w wielu zadaniach,

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

[41] `http://nlp.pwr.wroc.pl/takipi/`

[42] Lematyzacja zwraca poprawne formy hasłowe, w przeciwieństwie do tzw. *stemmingu*, który opiera się na bezkontekstowym, heurystycznym obcinaniu pseudo-końcówek.

dotyczących praktycznego pozyskiwania wiedzy lingwistycznej z polskich tekstów na dużą skalę [73], (Piasecki, Szpakowicz, and Broda, 2009).

W momencie powstania, a także jeszcze przez kilka lat później *TaKIPI* był najlepszym tagerem dla języka polskiego uzyskując dokładność ujednoznaczniania przekraczającą 92%. Wręcz umożliwił wiele typów analizy języka polskiego. Można nawet stwierdzić, że tager *TaKIPI* otworzył możliwości wielu badań naukowych w oparciu o polskie dane, dzięki temu, że był łatwo dostępny poprzez sieć i dość łatwy w instalacji i użyciu. *TaKIPI* jest wykorzystywany do lematyzacji polskich tekstów, konwersji tekstu do formatu anotowanego lingwistycznie na poziomie morfosyntaktycznym oraz do przygotowywania danych do analizy statystycznej tekstów. Ze względu na kompleksową konstrukcję, łatwość instalacji i dość dużą szybkość działania, jeszcze przez długie lata był stosowany, a nawet bywa stosowany jeszcze dzisiaj, pomimo pojawienia się tagerów o wyższej dokładności działania.

Pierwsza wersja TaKIPI została zbudowana w ramach projektu badawczego realizowanego przez IPI PAN, dlatego jest on dostępny na otwartej licencji zarówno z witryny PWr, jak i IPI PAN, a także repozytorium CLARIN-PL: `https://clarin-pl.eu/dspace/handle/11321/31`. Użytkownicy pobierający TaKIPI z witryny PWr. są proszeni o dobrowolną rejestrację. Od roku 2009 zarejestrowało się 67 użytkowników: naukowców, pracowników firm komercyjnych oraz studentów. Zadeklarowali szeroką gamę zastosowań badawczych oraz komercyjnych. Cytowania (łącznie 103 cytowania) oraz wyniki wyszukiwania w sieci pokazały ponadto, że TaKIPI został wykorzystany w kilkudziesięciu projektach naukowych[43], w tym przynajmniej kilkunastu projektach zagranicznych (głównie z obszaru slawistyki i analizy języków słowiańskich) oraz przynajmniej jednym projekcie komercyjnym. Na potrzeby parametryzacji Wydziału Informatyki i Zarządzania PWr. w roku 2017 zebrano również 8 kart aplikacji TaKIPI potwierdzających oficjalnie jego zastosowanie w jednostkach naukowych z Polski oraz z Institute of the Czech National Corpus.

Miałem od samego początku główny udział w budowie *TaKIPI*: począwszy od koncepcji, poprzez opracowanie algorytmów łącznie z modelem lingwistycznym (obejmującym reguły tagowania), a także wkład w projektowanie programu oraz kierowałem pracami nad implementacją. Część prac projektowych i prace implementacyjne zostały wykonane przez mgr inż. Grzegorza Godlewskiego.

5.2. *Słowosieć* – wordnet języka polskiego

Słowosieć[44] była w pierwszych wersjach budowana jako wordnet (inaczej leksykalna sieć semantyczna) języka polskiego (Piasecki, Szpakowicz, and Broda, 2009), aby później zostać rozwinięta do postaci wielkiego, relacyjnego słownika semantycznego języka polskiego. W pewnym przybliżeniu można ją również opisać jako rodzaj elektronicznego, sformalizowanego tezaurusa. *Słowosieć* opisuje znaczenia leksykalne przy pomocy relacji leksykalno-semantycznych takich jak: synonimia, hiperonimia (ogólne-bardziej szczegółowe), meronimia (całość-część), antonimia, kauzacja, żeńskość, wartość cechy i bardzo wiele innych. Zawiera również elementy opisu słownikowego znaczeń, takie jak krótkie tekstowe ich definicje (nazywane glosami), przykłady użycia oraz podział na rejestry stylistyczne. Zapisana jest w bazie danych i formacie opartym na XML-u, co umożliwiaj jej liczne zastosowania w inżynierii języka naturalnego i lingwistyce. *Słowosieć* opisuje znaczenia przy pomocy 53 typów relacji leksykalno-semantycznych oraz łącznie 107 podtypów relacji. Pojęcie wordnetu i model *Słowosieci* zostały po krótce przedstawione w sekcji 4.2.

Po latach ciągłej budowy (2005-obecnie, ale z różnym natężeniem) *Słowosieć* stała się największym wordnetem na świecie, jednym z największych słowników języka polskiego zbudowanych w historii i jednym z najczęściej stosowanych podstawowych zasobów językowych dla języka polskiego. Słowosieć

[43] Praca [57] została zacytowana w 71 publikacjach i raportach, dane z Google Scholar, po ręcznym usunięciu autocytowań. Pozostałe prace (Piasecki and Godlewski, 2006) 9 cytowań, (Piasecki, 2006) 10 cytowań, [65] 13 cytowań, [87] 9 cytowań – razem 103 cytowania.

[44] `http://plwordnet.pwr.edu.pl`

jest podstawą do analizy semantycznej tekstów na poziomie leksykalnym. Tabela 3 przedstawia statystyki *Słowosieci* obrazujące jej wielkość. Liczba lematów to *de facto* liczba haseł, a liczba jednostek leksykalnych to liczba różnych znaczeń leksykalnych opisanych w ramach tego zasobu.

Tabela 1. Podstawowe statystyki dla *Słowosieci 4.0 emo*

Elements	Verbs	Nouns	Adv.	Adj.	All
Lematy	19 941	133 843	8 010	29 228	**191 022**
Jednostki leksykalne	40 799	176 935	14 040	54 021	**283 795**
Synsety	29 650	132 623	11 260	46 705	**220 238**

Wielkość *Słowosieci* ujawnia się poprzez jej porównanie z innymi największymi wordnetami świata pokazane w Tabeli 4. *Princeton WordNet 3.1* był przez lata największym zasobem leksykalno semantycznym. Natomiast *GermaNet* był największym wordnetem zbudowanym od podstaw dla języka innego niż język angielski. Trzeci z porównywanych zasobów – *enWordNet 1.0* – to *Princeton WordNet 3.1* rozszerzony przez nas ręcznie o ponad 10 000 angielskich jednostek leksykalnych po to, aby poprawić jakość ręcznego odniesienia *Słowosieci* do znaczeń języka angielskiego.

Tabela 2. Porównanie *Słowosieci 4.0 emo* z największymi wordnetami świata

Wordnet	Synsety	Lematy	Jednostki lek.
GermaNet	101 371	119 231	131 814
Princeton WordNet 3.1	117 659.	155 593.	206 978
enWordNet 1.0	125 500.	165 712.	218 611
Słowosieć 4.0 emo	222,137	191 447	288 074

Wielkość wordnetu jest jego bardzo istotną cechą, ponieważ warunkuje stopień pokrycia danych językowych – im większy wordnet, tym większa szansa, że słowa i znaczenia leksykalne występujące w przetwarzanych danych są opisane w danym wordnecie. Drugim istotnym czynnikiem jest dobór lematów opisanych w danym wordnecie i dokładność opisu ich znaczeń. We wszystkich tych aspektach *Słowosieć* wypada bardzo dobrze w porównaniu z innymi dużymi wordnetami. Ponieważ *Słowosieć* jest cały czas iteracyjnie rozbudowywana w oparciu o dane z bardzo dużych korpusów tekstów, to dzięki swojej wielkości osiągnęła bardzo wysoki stopień pokrycia słów w tekstach, na przykład jest on co najmniej dwukrotnie wyższy niż w *Princeton WordNet*, por. badanie na tekstach z Wikipedii [40]. Gęstość sieci relacji w *Słowosieci* jest wyższa niż w innych wordnetach, por. [40] i (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016), co oznacza, że jest większa jest ilość informacji przypadająca na jedną jednostkę leksykalną. Przeprowadzone analizy porównawcze wersji rozwojowych *Słowosieci* w stosunku do *Princeton WordNetu* w oparciu o grafowe pojęcie *małego świata* (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016) pokazały, że *Słowosieć* charakteryzuje się znacznie lepszymi własnościami w odniesieniu do wszystkich wskaźników, tj. średniej długości ścieżki (ang. average path lenght), współczynnika grupowania (ang. clustering co-efficient) oraz łączliwości (ang. connectivity).

Jak już było to wspomniane w sekcji 4.3, na potrzeby konstrukcji *Słowosieci* zaproponowaliśmy unikatowy w skali światowej model lingwistyczny oraz metodę budowy, opartą konsekwentnie na analizie dużych korpusów tekstów jako podstawowego źródła informacji. W tym celu zbudowanych zostało szereg narzędzi do analizy semantycznej tekstów, które wspierają lingwistów w eksploracji bardzo dużych zbiorów tekstowych (w przypadku Słowosieci 4.0 był to korpus przekraczający 4 miliardy segmentów) i podpowiadają relacje semantyczne, znaczenia i przykłady użycia. Jednak decyzje ostateczne odnośnie do opisu znaczeń w *Słowosieci* podejmuje zawsze lingwista.

Jednym z naszych priorytetów była zbudowanie *Słowosieci* w sposób wiernie odzwierciedlający system leksykalny języka polskiego. Dlatego została ona zbudowana całkowicie niezależnie od słynnego *Princeton WordNetu* dla języka angielskiego. Później została dużym nakładem pracy ręcznie zrzu-

towana na struktury *Princeton WordNet* według zaproponowanej metody lingwistycznej [97], której kluczowym elementem jest dopasowanie struktur relacji opisujących znaczenia leksykalne. W rezultacie powstał unikatowy w skali światowej dwujęzyczny zasób umożliwiający porównywanie leksykalnych systemów języka polskiego i angielskiego. Jednocześnie jest to największy publicznie dostępny słownik polsko-angielski z ponad 240 tysiącami powiązań pomiędzy synsetami (zbiorami synonimów) polskimi i angielskimi. Słowosieć jest udostępniana na otwartej licencji i jest dostępna również w postaci aplikacji mobilnej na platformie Android, por. `http://plwordnet.pwr.wroc.pl/wordnet/download`. Wywiera bardzo duży wpływ na rozwój technologii językowych i ich zastosowania (setki cytowań[45]). Prace nad rozbudową Słowosieci są nadal kontynuowane. Jest ona jednym z tych polskich osiągnięć w dziedzinie budowy zasobów językowych, szerzej infrastruktury naukowej, gdzie to Polska wyznacza standardy, które są punktem odniesienia dla świata.

W przeciwieństwie do wielu innych wordnetów w konstrukcji *Słowosieci* skoncentrowaliśmy się wyłącznie na opisie znaczeń leksykalnych języka polskiego i systemu ich relacji znaczeniowych. Natomiast na potrzeby rozlicznych zastosowań *Słowosieci* zbudowaliśmy wokół niej cały system zasobów leksykalnych i wiedzy [42]:

- *MWELexicon*[46] [29] – słownik ustalonych wyrażeń wielowyrazowych (około 60 000) (silnych frazeologizmów) opisujących ich strukturę leksykalno-składniową – wyrażenia tego słownika są jednocześnie lematami w *Słowosieci*,
- *Walenty* [94] – leksykon opisujący struktury argumentów wymaganych przez niektóre jednostki leksykalne, szczególnie czasowniki (ponad 15 000 opisanych lematów)– został połączony ze *Słowosiecią* na poziomie znaczeń,
- odniesienie do enWordNet 1.0 (rozszerzonego przez nas *Princeton WordNet 3.1*) – wordnetu języka angielskiego – poprzez dwujęzyczne powiązania pomiędzy synsetami (ponad 240 000 ręcznie utworzonych powiązań), buduje pomost pomiędzy polskim i angielskim systemem leksykalnym (największy, publicznie dostępny słownik polsko-angielski),
- warstwa anotacji emotywnej[47] dla jednostek leksykalnych (ponad 86 000 ręcznie opisanych jednostek[48]): polaryzacji wydźwięku emocjonalnego, ośmiu podstawowych emocji i fundamentalnych wartości [107],
- rzutowanie na zasoby wiedzy – powiązanie pomiędzy systemem znaczeń leksykalnych a zasobami opisującymi byty i pojęcia:
  - rzutowanie *NELexicon 2.0*[49] – bardzo duży słownik około 2,4 mln nazw własnych języka polskiego zbudowany przez nas – powiązany ze Słowosiecią na poziomie kategorii semantycznych nazw,
  - ręcznie zdefiniowane powiązania z polską Wikipedią (ponad 55 000), rozszerzane półautomatycznie przy pomocy opracowanych przez nas metod [67],
  - rzutowanie *Słowosieci* na pojęcia z SUMO Ontology [52] – ontologię ogólną – określone półautomatycznie, z niewielkim błędem dla większości synsetów [25].

Dzięki powiązaniom *Słowosieci* z wordnetem angielskim otwiera się szeroka gama dwujęzycznych i wielojęzycznych zastosowań.

Słowosieć jest podstawą do analizy semantycznej tekstów na poziomie leksykalnym. Została ona pobrana przez blisko 1 300 zarejestrowanych użytkowników (indywidualnych i instytucjonalnych), w tym ponad 200 komercyjnych (między innymi Agora, Orange, mBank, Google), deklarujących szereg zastosowań, np., *Słowosieć* znajduje się na liście zasobów wykorzystywanych przez usługę Go-

[45] Na przykład: (Piasecki, Szpakowicz, Maziarz, and Rudnicka, 2016): 6 cytowań, (Piasecki, Szpakowicz, and Broda, 2009): 91 cytowań, [41]: 4 cytowania, [40]: 4 cytowania, [43]: 37 cytowań, [80]: 6 cytowań, [14]:35, [13]: 14, ...

[46] https://clarin-pl.eu/dspace/handle/11321/508

[47] Wersje *Słowosieci* wyposażone w anotację emotywną oznaczane są poprzez sufiks *emo*

[48] Opracowaliśmy również metodę automatycznego rozszerzania anotacji polaryzacji wydźwięku na większość synsetów *Słowosieci* wysoką dokładnością [26].

[49] https://clarin-pl.eu/dspace/handle/11321/247

ogle Translate. W ramach raportu końcowego fazy konstrukcji infrastruktury badawczej CLARIN-PL (`http://clarin-pl.eu`) mgr Agnieszka Dziob opracowała zestawienie 142 zidentyfikowanych użytkowników niekomercyjnych *Słowosieci*: naukowców, doktorantów i studentów wykorzystujących ją w badaniach i dydaktyce. Do tej liczby należy doliczyć kilkudziesięciu naukowców i wielu studentów z jednostek naukowych tworzących CLARIN-PL wykorzystujących *Słowosieci*, którzy nie zostali ujęci w raporcie. Znalazła się w nim natomiast lista 51 zidentyfikowanych użytkowników komercyjnych z Polski oraz wielu ze świata. Na potrzeby parametryzacji Wydziału Informatyki i Zarządzania PWr. w roku 2017 zebrano również 29 kart aplikacji *Słowosieci* potwierdzających oficjalnie jej zastosowanie w jednostkach naukowych z Polski oraz m.in. z Bulgarian Academy of Sciences, Charles University in Prague, GieBener Zentrum Ostliches Europa (Gieben, Niemcy), Kamusi Project International (Szwajcaria), Lund University (Szwecja), Nanyang Techological University (Singapur), National University of Ireland Galaway, PONS GmbH (Niemcy), University of Copenhagen.

Na podstawie cytowań, wymienienia nazwy *Słowosieci* w internecie oraz deklaracji przy rejestracji użytkowników, wyłania się bardzo bogaty obraz jej zastosowań w nauce i przy konstrukcji systemów do analizy języka naturalnego (jedno i wielojęzycznych). Poniżej wymieniono wybrane, a obszerniejszą listę można znaleźć w publikacjach:: (Maziarz, Piasecki, Rudnicka, Szpakowicz, and Kędzia, 2016), [81], [40] oraz prezentacjach z warsztatów szkoleniowych CLARIN-PL[50]. Wybrane zastosowania z literatury to:

— anotacja semantyczna korpusów,
— konstrukcja słowników wielojęzycznych: Ling.pl (`http://ling.pl`), Open Multilingual Wordnet, WordTies, PanLex,
— korekta językowa,
— ujednoznacznianie znaczeń słów w tekstach,
— wydobywanie terminologii z korpusów i grupowanie terminów,
— wydobywanie relacji semantycznych i wiedzy z tekstów,
— konstrukcja systemów odpowiadających na pytania.

Wśród zadeklarowanych zastosowań możemy znaleźć:

— badania nad ontologiami i generowanie ontologii, leksykalizacja pojęć ontologicznych,
— badania nad polską leksyką,
— ujednoznacznianie wyjściowych analiz struktury zdania z parsera,
— indeksowanie dokumentów, opisywanie metadanymi lub tagowanie,
— klasyfikacja tekstów: fragmentów i dokumentów,
— wykorzystanie w ramach badań korpusowych, np. nad określonymi klasami słów,
— budowa systemów wspomagających naukę języka,
— analiza zapożyczeń i frazeologizmów,
— komunikacja z robotami w języku naturalnym.

Jednak najbardziej dumni jesteśmy z praktycznego wdrożenia *Słowosieci* w jednej z wrocławskich firm w ramach leczenia afazji.

Kieruję projektem budowy *Słowosieci* od samego początku w 2005, od sformułowania pierwszego wniosku projektowego i sformułowanie pierwszej koncepcji. Miałem znaczący udział w opracowaniu unikalnej koncepcji modelu oraz koncepcji półautomatycznej metody tworzenia. Od samego początku do dnia dzisiejszego sprawuję ogólne kierownictwo i nadzór nad realizacją projektu przez interdyscyplinarny zespół, powierzając szczegółowe prace i decyzje leksykograficzne koordynatorom zespołów lingwistów *Słowosieci*: polskiej części (w chronologicznej kolejności, dr hab. Magdalena Zawisławska, dr Marek Maziarz, mgr. Agnieszka Dziob), angielskiej części (dr Ewa Rudnicka) oraz anotacji emotywnej (dr hab. Monika Zaśko-Zielińska). Brałem udział w zdecydowanej większości prac badawczych

[50] na przykład `http://clarin-pl.eu/wp-content/uploads/2016/04/konferencja/Premiera30-zastosowania.ppt` lub `http://clarin-pl.eu/wp-content/uploads/2017/09/SÅČowosieÄĞIPI.pdf`

dotyczących *Słowosieci* i jestem współautorem większości prac naukowych, na przykład [16], [96], [42], [39], [44], [108], [45], [37], [46], [38], [82], [14], [13].

Na potrzeby pracy zespołu lingwistów zbudowaliśmy sieciowy system WordnetLoom [50, 70] umożliwiający równoległą pracę na centralnej bazie banych oraz zarządzanie zespołem lingwistów. Zespół liczył przeciętnie około 20 lingwistów (w niektórych okres do 50 osób) i kilku wspomagających informatyków. Łączny nakład pracy na zbudowanie Słowosieci 3.0 emo przekroczył 40 osobolat.

5.3. Infrastruktura badawcza CLARIN-PL oraz Centrum Technologii Językowych CLARIN-PL

CLARIN ERIC[51] – *Common Language Resources & Technology Infrastructure* (pol. Wspólne zasoby językowe i infrastruktura technologiczna) – to konsorcjum naukowe typu ERIC (European Research Infrastructure Consortium) utworzone pod koniec 2011 poprzez osiem państw (Austrię, Bułgarię, Czechy, Danię, Estonię, Holandię, Niemcy i Polskę) i jedną organizację międzypaństwową (The Dutch Language Union). Warto podkreślić, że Polska jest jednym z członków założycieli CLARIN ERIC. Obecnie CLARIN ERIC tworzy 22 członków i dwóch obserwatorów. W budowę konsorcjum, a następnie infrastruktury CLARIN ERIC byłem zaangażowany od samego początku, tj. od roku 2005 (dokładnie od drugiego spotkania roboczego), zarówno na etapie przygotowawczego projektu europejskiego (finansowanego z 7FP), jak również od 2011 w zasadniczej budowie infrastruktury. Byłem jedynym polskim członkiem wąskiej grupy roboczej pracującej nad wnioskiem o projekt 7FP w latach 2006-2007, później przedstawicielem MNiSW w radzie naukowej CLARIN (2008-2010), ekspertem oddelegowanym przez MNiSW do prac w komitecie sterującym przygotowującym strukturę konsorcjum, natomiast od roku 2012 nieprzerwanie pełnię rolę Polskiego Koordynatora Narodowego i członka rady zarządzającej pracami konsorcjum CLARIN ERIC (od 2018 jestem przewodniczącym tej rady). W bardzo dużym stopniu miałem i mam wpływ na kształtowanie się i rozwój CLARIN ERIC jako jednej z wielkich infrastruktury badawczych Unii Europejskiej. CLARIN ERIC to pod wieloma względami największa europejska infrastruktura badawcza z obszaru nauk humanistycznych i społecznych. Budowana jest jednak w znacznej mierze przez instytucje naukowe i naukowców z obszaru szeroko pojętej informatyki.

Od samego początku byłem liderem, a później zostałem koordynatorem konsorcjum naukowego i infrastruktury badawczej CLARIN-PL[52] – polskiej części CLARIN ERIC. CLARIN-PL jest również częścią Polskiej Mapy Drogowej Infrastruktury Badawczej od samego początku jej istnienia i jedną z zaledwie kilku infrastruktur z tej mapy, które osiągnęły zaawansowany poziom realizacji. Miałem znaczący wkład w budowę CLARIN-PL poprzez planowanie, koordynację, kierowanie pracami i osobisty udział w wielu pracach badawczo-rozwojowych, często opartych na interdyscyplinarnej współpracy naukowej. Budowa i utrzymanie CLARIN-PL przyczyniło się niezmiernie do bardzo znaczącego rozwoju jakościowego i ilościowego technologii językowej dla języka polskiego.

CLARIN ERIC wyrósł z idei lepszego udostępnienia narzędzi i zasobów językowych jako językowych jako narzędzi badawczych badaczom z obszaru nauk humanistycznych i społecznych jako narzędzi badawczych. Pod pojęciem zasobów językowych rozumiemy opisy języka naturalnego, które są sformalizowane na różnych poziomach, np. leksykalne bazy danych, modele językowe, gramatyki formalne, itp. Pod pojęciem narzędzi językowych rozumiemy programy komputerowe przetwarzające język naturalny, na przykład do analizy składniowej i semantycznej, rozpoznawania mowy, wydobywania wiedzy z tekstów, itp. Narzędzia językowe najczęściej wykorzystują różnorodne zasoby językowe. Zasoby i narzędzia językowe oraz budowane z nich systemy składają się na technologię językową. Ostatecznym celem CLARIN jest budowa rozproszonych, niezawodnych aplikacji badawczych, które umożliwią rozszerzony dostęp oraz zautomatyzowaną analizę dużych zbiorów dokumentów tekstowych, nagrań języka mówionego oraz zasobów multimedialnych reprezentujących komunikację przy pomocy języka

[51] http://clarin.eu
[52] http://clarin-pl.eu

naturalnego. Cel ten staramy się osiągnąć poprzez budowę paneuropejskiej infrastruktury naukowej, która łączy w interoperacyjną technologię językową sieć centrów technologicznych zlokalizowanych w różnych krajach członkowskich CLARIN ERIC.

CLARIN ERIC jest rozproszoną sieciową infrastrukturą badawczą, której poszczególne elementy są konstruowane i utrzymywane przez członków konsorcjum. Obejmuje ona obecnie 45 certyfikowanych centrów technologicznych i centrów wiedzy w 22 krajach członkowskich. Dwa podstawowe rodzajów centrów to centra technologii językowych (tzw. centra CLARIN typu B) oraz centra infrastruktury szerzenia wiedzy (centra CLARIN typu K). Dla każdego typu centrum określona jest lista wymogów jakie musi ono spełniać, od czego zależy powodzenie procesu certyfikacji, przeprowadzanego co dwa lata przez międzynarodowe komisje eksperckie działające w ramach CLARIN ERIC. Centra typu B, jako repozytoria danych naukowych, dodatkowo muszą aplikować i odnawiać regularnie międzynarodowy certyfikat *Data Seal of Approval*[53].

Centra typu B to kluczowe elementy infrastruktury, które są wyposażone w funkcje rozproszonej autoryzacji i autentykacji, pełnią funkcję repozytoriów naukowych, dostarczając użytkownikom podstawowych funkcji sieciowych oraz dostępnych w sieci narzędzi językowych i aplikacji badawczych, opartych na technologii językowej. Każdy członek CLARIN ERIC jest zobowiązany do utrzymywania i rozwoju przynajmniej jednego centrum typu B. W Polsce, na mocy decyzji konsorcjum CLARIN-PL, powstało jedno centrum CLARIN typu B zorganizowane i zbudowane na Politechnice Wrocławskiej pod nazwą *Centrum Technologii Językowych CLARIN-PL*[54] (dalej CTJ CLARIN-PL). CTJ CLARIN-PL udostępnia repozytorium danych językowych spełniające standardy DSA i CLARIN ERIC, między innymi odnośnie do standardu metadanych. Jest ono zintegrowane z całością CLARIN ERIC w zakresie wymiany metadanych i udostępniania usług do przeszukiwania korpusów oraz przetwarzania języka naturalnego. CTJ CLARIN-PL przechowuje kilkadziesiąt zasobów językowych oraz oferuje dziesiątki podstawowych narzędzi językowych[55] dla języka polskiego oraz kilku innych języków w postaci usług sieciowych (mikroserwisów) oraz prostych aplikacji webowych. Ponadto CTJ CLARIN-PL udostępnia kilkadziesiąt aplikacji badawczych w większości zaprojektowanych we współpracy z naukowcami dziedzin nauk humanistycznych i społecznych pod konkretne zadania badawcze, a następnie uogólnianych do innych zadań badawczych. Unikatową funkcjonalnością CTJ CLARIN-PL jest zbudowanie i udostępnienie użytkownikom-naukowcom *CLARIN Cloud* – prywatnej chmury danych przeznaczonej dla naukowców, którą mogą oni wykorzystywać do przechowywania danych badawczych. Całość centrum, wszystkie jego moduły i funkcjonalności są dostępne bezpłatnie na otwartych licencjach, a samo centrum działa w trybie ciągłym i obsługuje kilkaset tysięcy zapytań i żądań przetwarzania rocznie. Łączny roczny wolumen przetwarzanych danych tekstowych to dziesiątki gigabajtów.

Wiele elementów technologii językowej zostało zbudowanych specjalnie na potrzeby CTJ CLARIN-PL w ramach CLARIN-PL, w tym wiele rozwiązań powstało przy aktywnym udziale pracowników CTJ.

Z CTJ CLARIN-PL jest ściśle powiązane jest *PolLinguaTec*[56] – Centrum Wiedzy CLARIN Technologii Językowej dla Języka Polskiego – certyfikowane centrum CLARIN typu K, które zostało powołane, aby świadczyć użytkownikom stałą pomoc w zastosowaniach badawczych technologii językowej dla języka polskiego. Zgodnie z wymagamiami CLARIN ERIC *PolLinguaTec* jest zobowiązane do udzielenia odpowiedzi w ciągu 48 godzin. Ponadto działa ono proaktywnie organizując warsztaty szkoleniowe oraz utrzymując bezpośrednią współpracę z użytkownikami-naukowcami w ramach ich projektów badawczych. Pomimo iż *PolLinguaTec* to głównie zespół ludzi, to centrum to jest bardzo istotnym elementem działania CLARIN-PL jako infrastruktury badawczej w ramach dynamicznie rozwijającego się nurtu humanistyki cyfrowej i metod cyfrowych w naukach społecznych.

CTJ CLARIN-PL oraz *PolLinguaTec* są budowane i utrzymywane przez kilkudziesięcioosobowy ze-

[53] https://www.datasealofapproval.org/en-gb/
[54] http://clarin-pl.eu
[55] http://ws.clarin-pl.eu
[56] http://kcentre.clarin-pl.eu/

spół badawczo-rozwojowy, który zorganizowałem i którego pracą kieruję od samego początku. Jestem autorem innowacyjnej koncepcji infrastruktury badawczej zorientowanej na użytkowników [58] i procesu jej tworzenia. Miałem znaczący udział w pracach nad koncepcją obu centrów oraz uczestniczyłem w pracach projektowych [104, 90, 91]. W dziedzinie projektowania i wytwarzania CTJ kluczowy wkład mają prace mgr inż. Marcina Pola oraz dr inż. Tomasza Walkowiaka. Brałem udział w opracowywaniu szeregu narzędzi językowych i aplikacji badawczych jakie są udostępniane poprzez CTJ CLARIN-PL. Jestem w dużej mierze odpowiedzialny za wizję dalszego rozwoju i utrzymania CTJ i CLARIN-PL. Ponieważ CLARIN-PL wytworzył i utrzymuje znaczącą część technologii językowej dla języka polskiego, ***CLARIN-PL ma on istotny wpływ na rozwój nauki w Polsce*** .

5.4. *MeWeX* – webowy system do wydobywania kolokacji

MeWeX[57] to webowy system służący do wydobywania z korpusów kolokacji o określonych typach strukturalnych, na przykład rzeczownik + przymiotnik (*broń jądrowa*) czy rzeczownik + przymiotnik + rzeczownik (*broń masowego rażenia*). W analizie kolokacji wykorzystywanych jest wiele miar statystycznych sprawdzających siłę powiązania. *MeWeX* można wykorzystać również do rozbudowy słownika leksykalnych jednostek wielowyrazowych (tj. niekompozycyjnych semantycznie oraz terminów). Jest bardzo cennym narzędziem wykorzystywanym zarówno w lingwistyce korpusowej, jak też w wielu innych dziedzinach humanistyki jako narzędzie do analizy tekstów pod kątem występujących w nich frazeologizmach i terminach.

Podstawą do konstrukcji *MeWeX-a* były metody wydobywania kolokacji z polskich tekstów wykorzystujące filtry w postaci ograniczeń leksykalno-morfo-syntaktycznych opracowane przy moim istotnym udziale i przez zespół kierowany przeze mnie [2]. Algorytmy te zostały poszerzone o kilkadziesiąt miar asocjacyjnych [89] i zaimplementowane w postaci pierwszej wersji systemu *MeWeX*. Następnie aplikacja webowa została uproszona do obecnej postaci w wyników testów z udziałem użytkowników, na przykład podczas warsztatów szkoleniowych CLARIN-PL.

5.5. *SuperMatrix* – system do wydobywania modeli semantyki dystrybucyjnej

SuperMatrix[58] [7, 6] to uniwersalny system do wydobywania modeli semantyki dystrybucyjnej z dużych korpusów tekstów, który został już krótko omówiony w sekcji 4.4.2. Implementuje dziesiątki algorytmów semantyki dystrybucyjnej (między innymi wiele sposobów transformacji macierzy koincydencji oraz wyliczenia podobieństwa wektorów na ich podstawie), umożliwia efektywne rozproszone przetwarzanie na sieci komputerów lub węzłów obliczeniowych oraz wspiera zaproponowaną przeze mnie unikatową metodą wzbogacania opisu zlematyzowanego tekstu o binarne relacje morfosyntaktyczne przy pomocy ograniczeń leksykalno-morfo-syntaktycznych.

Sformułowałem ideę *SuperMatrixa* jestem także autorem i współautorem większości jego algorytmów z zakresu semantyki dystrybucyjnej.Głównym wykonawcą, szczególnie od strony projektowej i implementacyjnej był ówczesny doktorant, obecnie dr inż. Bartosz Broda. Kierowałem pracami nad budową i dalszą rozbudową systemu.

SuperMatrix był bardzo intensywnie wykorzystywany przez lata przy rozwoju *Słowosieci*, m.in. do wydobywania przykładów z tekstów [8],[5], analizy stylometrycznej, na przykład wewnątrz wcześniejsze wersji *WebSty* [86], oraz klasyfikacji semantycznej, na przykład [36, 71]. *SuperMatrix* został wdrożony w Kanadzie do przetwarzania tekstów angielskich oraz w Słowenii do języka słoweńskiego [19].

[57] http://ws.clarin-pl.eu/mewex.shtml
[58] https://clarin-pl.eu/dspace/handle/11321/271

Maciej Piasecki