Michał Woźniak

Jak znaleźć igłę w stogu siana?

Automatyczna ekstrakcja wielosegmentowych jednostek leksykalnych z tekstu polskiego

PRACE INSTYTUTU JĘZYKA POLSKIEGO PAN

150

POLSKA AKADEMIA NAUK • INSTYTUT JĘZYKA POLSKIEGO

Michał Woźniak

Jak znaleźć igłę w stogu siana?

Automatyczna ekstrakcja wielosegmentowych jednostek leksykalnych z tekstu polskiego

Kraków 2017

Monice, Nataszy

Recenzent:
prof. dr hab. ADAM PAWŁOWSKI

Redakcja:

Korekta: Iwona Gądek

Projekt, opracowanie graficzne, skład, łamanie, druk i oprawa:
Grafpol Agnieszka Blicharz-Krupińska
ul. Czarnieckiego 1, 53-650 Wrocław
tel. 507 096 545; mail: argrafpol@argrafpol.pl

ISBN 978-83-64007-43-9

Instytut Języka Polskiego PAN
al. Mickiewicza 31
31-120 Kraków
www.ijp.pan.pl

Wykaz skrótów używanych w książce

Słowniki

- ISJP – *Inny słownik języka polskiego* pod redakcją M. Bańko
- SJPDor – *Słownik języka polskiego* pod redakcją W. Doroszewskiego
- SJPDun – *Słownik współczesnego języka polskiego* pod redakcją B. Dunaja
- SJPSzym – *Słownik języka polskiego* pod redakcją M. Szymczaka

Miary asocjacji

- Dice – współczynnik Dice'a
- LLR – *Log-likelihood ratio* (logarytm wskaźnika wiarygodności)
- MI – *Mutual Information* (informacja wzajemna)
- PMI – *Pointwise Mutual Information* (punktowa informacja wzajemna)
- t-score – test *t* Studenta
- z-score – wskaźnik *z*
- $\chi 2$ – test chi-kwadrat Pearsona

Części mowy

- N – rzeczownik
- A – przymiotnik
- V – czasownik
- Adv – przysłówek
- P – zaimek
- Pr – przyimek

Pozostałe skróty

- MWE – *Multiword Expression* (wyrażenie wielowyrazowe)
- NLP – *Natural Language Processing* (przetwarzanie języka naturalnego)
- SVM – *Support Vector Machine* (maszyna wektora nośnego)

Spis treści

I
Zagadnienia wstępne

Jedną z najbardziej fascynujących kwestii związanych z językiem naturalnym jest złożoność i wieloznaczność jego jednostek leksykalnych. Już pojedyncze wyrazy (podstawowe jednostki leksykalne) charakteryzuje często wielość znaczeń, nierzadko również zachodzi problem z ich klasyfikacją, cóż więc powiedzieć dopiero o wyrażeniach składających się z dwóch lub więcej wyrazów, które są również prostymi (niezłożonymi) jednostkami semantycznymi?

Weźmy za przykład wilka morskiego i jego upodobania: *Wilk morski był za pan brat z bocianim gniazdem.* Analizując to zdanie, zauważyć można kilka charakterystycznych zjawisk:

1. Wyrazy w przytoczonym przykładzie nie reprezentują znaczeń im przypisywanych: *wilk* to tak naprawdę człowiek, *bocianie gniazdo* przeznaczone jest raczej dla marynarzy, nie ptaków, a fraza *być za pan brat* nie mówi nic o więzach rodzinnych;
2. O ile wyrażenie *wilk morski* ma jednoznaczną interpretację, o tyle *bocianie gniazdo* może być rozumiane na dwa sposoby: jako zwykłe połączenie (‘gniazdo bociana’) lub jako związek frazeologiczny (‘kosz lub platforma obserwacyjna umieszczana na dawnych statkach’);
3. Sekwencja *za pan brat* jest pozornie niegramatyczna: forma biernikowa *pan brat* jest archaiczna i – z punktu widzenia współczesnej składni – niepoprawna;
4. Nie jest jasne, jaki rodzaj gramatyczny przypisać *wilkowi morskiemu*. Ośrodkiem tej jednostki jest wyraz *wilk*, który w normalnej sytuacji ma rodzaj męskożywotny, natomiast semantycznie wyrażenie to odnosi się przede wszystkim do mężczyzn, co sugerowałoby rodzaj męskoosobowy. *Słownik poprawnej polszczyzny PWN* pod redakcją W. Doroszewskiego nakazuje odmianę według paradygmatu męskożywotnego (w szczególności więc w bierniku liczby mnogiej będą to *wilki morskie*), jednak dla użytkowników języka nie jest to oczywiste, o czym świadczą odnotowane przez Kosek (2008, 116) przykłady użycia formy biernikowej *wilków morskich* czy pytania skierowane do poradni językowych[1];
5. W zdaniu pojawia się dziewięć wyrazów, natomiast liczba jednostek leksykalnych zależy od interpretacji. W zależności bowiem od tego, czy frazę *być za pan brat* uznamy za całość, czy nie (*być*, *za pan brat*), jednostek mamy cztery lub pięć. Co więcej – jeżeli wspomniany *wilk morski* zamiast przebywać w *bocianim gnieździe*

[1] Zob. http://www.ujk.edu.pl/ifp/index.php/poradnia-jezykowa/14-poradnia-jezykowa/478-wilki-morskie, http://poradnia-jezykowa.uni.lodz.pl/faq/jaki-rodzaj-ma-wyrazenie-wilk-morski/.

prowadziłby obserwacje z *pomostu nawigacyjnego*, komplikowałby sytuację jeszcze bardziej: pojawiłaby się kolejna wątpliwość, czy wyrażenie to (*pomost nawigacyjny*) stanowi semantycznie niepodzielną całość (a więc leksykalną jednostkę), czy zwykłe (choć skonwencjonalizowane) połączenie wyrazów.

Powyższy przykład pokazuje rolę, jaką pełnią w języku wielosegmentowe jednostki leksykalne, jednocześnie sygnalizując problemy badawcze z nimi związane. Wielosegmentowe jednostki to z jednej strony interesujące zagadnienie językoznawcze, z drugiej – niemałe wyzwanie dla badacza. Rozpatrywanie tego zjawiska jest o tyle istotne, że wyrażenia wielosegmentowe mają duży (przy tym często niedoszacowany – por. Sag i in. 2002) udział w tworzeniu wypowiedzi. Wskazują na to liczni badacze; Jackendoff (1997) szacuje, że ilość jednostek tego typu jest porównywalna z ilością pojedynczych leksemów. Podobnie ocenia to Bogusławski (1989, 19), pisząc, że „idą one jawnie w grube miliony".

Wieloczłonowymi jednostkami leksykalnymi zajmują się różne (choć sobie pokrewne) działy badań: językoznawstwo ogólne, leksykografia i lingwistyka komputerowa. Każda z tych dziedzin nieco inaczej podchodzi do tematu, akcentując różne aspekty złożonych jednostek.

Z punktu widzenia językoznawstwa ogólnego jest to istotny problem teoretyczny, zarówno z powodu powszechności występowania, jak i złożoności. Przyczyny tej złożoności są dwie. Po pierwsze, to zjawisko językowe bardzo niejednorodne, o szerokich i płynnych granicach. Idiomy, związki frazeologiczne, wielowyrazowe terminy, czasowniki złożone (ang. *phrasal verbs*), złożenia rzeczownikowe – to tylko niektóre ze zjawisk językowych, mogących stanowić wielosegmentowe jednostki leksykalne. Po drugie, granice między wymienionymi typami wyrażeń złożonych, a przede wszystkim między złożonymi jednostkami a zwykłymi połączeniami, tworzonymi według reguł składni (typu *kolorowy zeszyt*) są bardzo często nieostre i wymykają się precyzyjnym ustaleniom. Z tych względów właściwie każda definicja wielowyrazowego leksemu boryka się z problemami, a obecność lub brak niektórych wyrażeń w zbiorze wyznaczonym przez zakres definicji może być dyskusyjna.

Dla leksykografii kwestia jednostek wielowyrazowych jest również bardzo istotna. Wiele typów słowników (np. słowniki przekładowe, jednojęzyczne słowniki ogólne) uwzględnia wyrażenia tego typu. Istnieją też bardziej wyspecjalizowane słowniki, zawierające wyłącznie jednostki złożone, takie jak słowniki frazeologiczne lub słowniki kolokacji (rozumianych jako typowe połączenia wyrazowe). Wyrażenia te również w leksykografii sprawiają jednak wiele problemów, zarówno natury teoretycznej, jak i praktycznej – np. czasami funkcjonują jako osobne hasła (np. w SJPDun), częściej zaś jako rozwinięcie hasła jednowyrazowego (np. SJPD). Istnieje wiele prac omawiających różne kwestie problematyczne i sporne, dotyczące jednostek wielosegmentowych z punktu widzenia leksykografii (por. m.in. Bogusławski 1987, Bańko 2001, Żmigrodzki 2009).

Zagadnienie złożonych jednostek leksykalnych – właściwości i praktyczne aspekty z nimi związane – bada także lingwistyka komputerowa. Zostanie to szerzej przedstawione w sekcji 4 i 5 tego rozdziału.

Prezentowane w niniejszej książce podejście do ekstrakcji jednostek wielosegmentowych opiera się w istotnym stopniu na wykorzystywaniu danych korpusowych do badań języka. Bywało ono w przeszłości kwestionowane (kwestie te, zwłaszcza wpływ N. Chom-

skiego, szerzej omawia Karlsson, 2008). Współcześnie językoznawcy uznają jednak wartość danych korpusowych (zob. Żmigrodzki 2009, 32; Bańko 2001, 39-43), a w przetwarzaniu języka naturalnego jest to podejście, które w zasadzie nie ma konkurencji.

Prace poświęcone leksemom wielowyrazowym i pojęciom pokrewnym zaczęły powstawać już w latach sześćdziesiątych ubiegłego wieku, a od lat dziewięćdziesiątych ich ilość i stopień zaawansowania zauważalnie wzrosły. Mają jednak dla nas istotny mankament: nie dotyczą języka polskiego. Lingwistyka komputerowa zaczęła rozwijać się w Polsce dopiero po roku 1989 (z nielicznymi wyjątkami, jak np. *Słownik frekwencyjny polszczyzny współczesnej* – zob. podpunkt 4.1), kiedy zniknęły ograniczenia dotyczące swobodnej wymiany myśli z krajami zachodnimi, a dystans technologiczny zaczął się zmniejszać. Prac z zakresu lingwistyki komputerowej, podejmujących tematykę wyrażeń wielosegmentowych funkcjonujących w języku polskim, jest więc, głównie z powodu tego zapóźnienia – niewiele. Mało jest również narzędzi i zasobów językowych (jak np. anotowane korpusy czy gotowe listy jednostek wieloczłonowych), które mogłyby być pomocne przy pracy z takimi jednostkami – w przeciwieństwie do sytuacji w krajach anglojęzycznych i krajach Europy Zachodniej, gdzie akcenty badawcze przesuwają się już z ekstrakcji i opisu w stronę badań nad wieloaspektowym wykorzystywaniem jednostek wielowyrazowych w praktycznych aplikacjach.

Wyraźnie zauważalnym kłopotem dotyczącym wielosegmentowych jednostek leksykalnych jest brak jednej, spójnej i powszechnie akceptowanej definicji. Wskazują na to zarówno prace z lingwistyki komputerowej (por. Seretan 2011b, Baldwin i Kim 2010, Ramisch 2012), jak i językoznawstwa ogólnego (por. Chlebda 2001, Kosek 2008). Ten brak konsekwencji jest do pewnego stopnia zrozumiały – rozmaitość materii językowej jest w tym przypadku naprawdę duża. Jednocześnie jednak brak standaryzacji powoduje kilka problemów. Po pierwsze, przyczynia się do pewnego rodzaju zamieszania, prowadząc do nieporozumień między językoznawcami (zob. Evert 2008, 3). Po drugie, utrudnia pracę leksykografów i lingwistów komputerowych. Bartsch (2004, 13-14), komentując sytuację dotyczącą wyrażeń wielowyrazowych w języku angielskim, zauważa, że przypomina ona błędne koło: ze względu na brak wystarczająco precyzyjnego określenia i definicji pojęcia, brakuje motywacji do podjęcia empirycznych studiów na ten temat, z drugiej zaś strony brak wystarczająco szerokiego materiału zaczerpniętego z odpowiednio obszernych próbek tekstu powoduje, że trudno nad taką definicją pracować. Podobna sytuacja dotyczy w dużym stopniu warunków polskich. Wiele prac skupia się zatem w pewnej części nad określeniem precyzyjnej definicji „od podstaw", co wymaga wielkiego nakładu pracy, ale efekty takich działań trudno ze sobą porównywać, ze względu na nieuniknione – mniejsze lub większe – różnice w przyjętych założeniach.

Niniejsza książka jest próbą rozwiązania wymienionych wyżej problemów. Jej cele są zatem następujące:

1. Przedstawienie spójnej definicji jednostek, która obejmowałaby możliwie szeroki zakres zjawisk językowych, a jednocześnie byłaby stosunkowo łatwa w zastosowaniu z punktu widzenia lingwistyki komputerowej. Ujmując rzecz ściślej, przedstawiona zostanie metodologia umożliwiająca skonstruowanie własnej definicji, dostosowanej do specyficznych potrzeb, jakie narzuca dany język i określone zadanie badawcze, pozostającej jednak w pewnych ogólnych ramach.

2. Stworzenie algorytmu automatycznej ekstrakcji wielosegmentowych jednostek leksykalnych z tekstu polskiego, wykorzystującego dane językowe i statystyczne (algorytm hybrydowy). Algorytm taki będzie zaprojektowany z myślą głównie o jednostkach dwuelementowych (tzw. bigramach) z dwóch powodów: po pierwsze bigramy stanowią lwią część wszystkich wielosegmentowych jednostek, po drugie wiele metod statystycznych, które pełnią istotną rolę przy ekstrakcji jednostek, można zastosować tylko do par. Niezależnie od tego, zastosowanie algorytmu do dłuższych wyrażeń również będzie możliwe (choć najczęściej mniej efektywne). Należy tu zauważyć, że problem rozróżniania wielosegmentowych jednostek i swobodnych połączeń jest – z punktu widzenia automatycznego klasyfikatora – bardzo trudny, zarówno ze względu na złożoność pojęcia, jak i fakt, że o przynależności wyrażenia do którejś z powyższych grup decydują w dużym stopniu czynniki semantyczne i zewnątrzjęzykowe, które – na obecnym etapie rozwoju lingwistyki komputerowej – są bardzo trudne do zbadania. Konsekwencją tego jest niezbyt wysoka skuteczność algorytmów ekstrakcji jednostek. Z tego względu miarą sukcesu osiągniętego przez proponowany w niniejszej książce algorytm powinna być raczej jego efektywność w porównaniu z już istniejącymi rozwiązaniami, a nie skuteczność bezwzględna.
3. Stworzenie narzędzi umożliwiających ocenę skuteczności zaproponowanego algorytmu, co wymaga opracowania (na podstawie powyższej definicji) zbioru porównawczego klasyfikującego wyrażenia w nim zawarte jako jednostki i połączenia. Jak wspomniano wyżej, algorytm zaprojektowany jest głównie z myślą o bigramach, dlatego testowany będzie na właśnie takich jednostkach.
4. Możliwie dokładne omówienie problematyki wielosegmentowych jednostek leksykalnych i – w ograniczonym zakresie – lingwistyki komputerowej. To pedagogiczne zadanie ma na celu umożliwienie zapoznania się z tematyką przez osoby nią zainteresowane i stworzenie punktu odniesienia dla dalszych badań.

1. Wielosegmentowe jednostki leksykalne – uwagi o terminologii

Chlebda (1991, 6) zauważa, że we frazeologii (nie tylko polskiej) panuje dosyć duży chaos terminologiczny. Zaproponowano wiele terminów, których znaczenie podlega mniej lub bardziej subtelnym zmianom u różnych badaczy. Co więcej, wiele z tych znaczeń nakłada się na siebie, jeszcze bardziej komplikując nazewnictwo.

Termin *wielosegmentowa jednostka leksykalna* występuje w literaturze niezbyt często – najczęściej jako wyrażenie tworzone ad hoc na potrzeby tekstu. Przydawka *wielosegmentowy* jest używana – w kontekście oznaczającym zestawienie wyrazów – w niektórych pracach z zakresu frazeologii i leksykografii (zob. Wróbel 1995, Żmigrodzki 2009, Kosek 2008), przeważnie w postaci *związek / jednostka / całostka / wyrażenie wielosegmentowe*. Dużo powszechniejszym terminem, o zbliżonym zakresie semantycznym (a w myśl niektórych definicji właściwie synonimicznym), jest *związek frazeologiczny* (*frazeologizm*). W tym samym (albo bardzo podobnym) znaczeniu używane są niekiedy *nieciągła / wieloczłonowa jednostka leksykalna*, *idiom*, *frazem* (por. Kosek 2008). Andrzej Bogusławski w swoich pracach (m.in. Bogusławski 1976, Bogusławski 1989) stosuje termin *jednostka języka* w zbliżonym rozumieniu, choć termin ten obejmuje jednostki zarówno jedno-, jak i wielosegmentowe.

Lingwistykę komputerową trapią podobne problemy (por. Grégoire 2009, 1-3; Ramisch 2012, 21-23). Chronologicznie najwcześniejszym (szeroko używanym) terminem, którym zaczęto określać jednostki wielosegmentowe, jest kolokacja (ang. *collocation*)[2]. Nie ma wśród badaczy zgody co do jednoznacznej definicji tego pojęcia. Niektórzy za kolokacje uznają wszystkie połączenia wykazujące tendencję do występowania obok siebie, inni akcentują lingwistyczne właściwości tego fenomenu językowego (leksykalne, składniowe, semantyczne, pragmatyczne). Dyskusję na temat statusu kolokacji w językoznawstwie ogólnym i lingwistyce komputerowej można znaleźć w pracach (Evert 2005, Bartsch 2004, Seretan 2011b). Innym ważnym terminem oznaczającym mniej więcej to samo pojęcie jest wyrażenie *wielowyrazowe/jednostka wielowyrazowa* (ang. *Multiword Expression/Unit*). Badacze najczęściej odwołują się do tego określenia, chcąc podkreślić jego znaczenie językowe, całościowość znaczenia i inne cechy lingwistyczne (por. Sag i in. 2002, Baldwin i Kim 2010) – termin ten coraz częściej zastępuje kolokację, której rozumienie ogranicza się obecnie raczej do wyrażeń o częstotliwości występowania większej niż wynikałoby to z przypadku.

W literaturze wymienione terminy są często stosowane zamiennie, co jest zrozumiałe ze względu na stopień ich wzajemnego powiązania. W tej książce przyjęto podobną zasadę: jednostka wielosegmentowa, wyrażenie wielowyrazowe, związek frazeologiczny, kolokacja będą najczęściej traktowane jak synonimy. Oczywiście, sytuacjom, w których tego typu podejście mogłoby być mylące, towarzyszą odpowiednie uściślenia.

Podobnie rzecz ma się z nazwami dziedziny badawczej, w krąg której wpisuje się niniejsza książka. Jako że jest to obszar interdyscyplinarny, łączący wiedzę o języku, informatykę, a w pewnym stopniu nawet psychologię, można spotkać różne nazwy dotyczące w zasadzie niemal tego samego. Dwie podstawowe, tj. lingwistyka komputerowa i przetwarzanie języka naturalnego (ang. *Natural Language Processing*) są używane zamiennie. W książce używany jest również powszechnie stosowany w literaturze angielski skrót: NLP.

Pozostaje wyjaśnić powód, dla którego w tytule książki pojawia się wielosegmentowa jednostka leksykalna, a nie związek frazeologiczny będący terminem znacznie popularniejszym. Związek frazeologiczny najczęściej rozumiany jest jako wyrażenie nieregularne pod jakimś (przeważnie znaczeniowym) względem (zob. Lewicki i Pajdzińska 2001). Taki sposób ujęcia tematu akcentuje zatem opozycję do sytuacji normalnej, regularnej. Tymczasem niniejsza książka skupia się na wyodrębnieniu wyrażeń wielosegmentowych (składających się z więcej niż jednego wyrazu), które zachowują się w tekście tak jak pojedyncze wyrazy, zatem – podobnie jak one – stanowią jednostki leksykalne. Ponadto, choć wiele spośród tego typu jednostek cechuje nieregularność, istnieją jednak takie, które nieregularności nie wykazują (lub jest ona dyskusyjna). Podobne podejście przyjmuje Kosek (2008), używając terminu *nieciągła jednostka leksykalna*.

2. Lingwistyka komputerowa

Niniejsza książka podejmuje temat wielosegmentowych jednostek leksykalnych z punktu widzenia lingwistyki komputerowej. Na gruncie polskiej nauki jest to obszar stosunkowo nowy i dopiero zyskujący na popularności, z tego względu poniżej przedstawiony zostanie w miarę zwięzły opis tej dziedziny i jej podstawowych koncepcji.

[2] Wcześniejsze jest określenie polylog z początku XX wieku, wywodzące się z prac Harolda Palmera (por. Seretan 2011b, 9).

Lingwistyka komputerowa i przetwarzanie języka naturalnego (NLP) to dwie, bardzo do siebie zbliżone, dziedziny badań. Z uwagi na fakt, że są ściśle powiązane z komputerami i technologiami informatycznymi, ich faktyczny rozwój można datować od lat pięćdziesiątych XX wieku. Dotyczy to oczywiście Stanów Zjednoczonych i krajów Europy Zachodniej – w Polsce, ze względu na jej polityczne odcięcie od świata zachodniego, i w konsekwencji brak dostępu do odpowiedniego sprzętu i badań, lingwistyka komputerowa zaczęła się rozwijać dopiero w latach dziewięćdziesiątych XX wieku.

2.1. Zadania lingwistyki komputerowej

W obrębie lingwistyki komputerowej można wyodrębnić wiele obszarów zainteresowania. Poniżej wymienione zostają niektóre z nich (w nawiasie podana jest terminologia angielska):

1) **Tłumaczenie maszynowe (Machine Translation).** Automatyczne przekładanie tekstu na inny język naturalny jest dużym wyzwaniem. Złożoność języka jako systemu i ogromna ilość różnic – od fundamentalnych po subtelne – między różnymi językami przysparzają bardzo wielu trudności. Jest to jedno z pierwszych zadań, jakie podjęto w ramach lingwistyki komputerowej, i do dziś pozostaje zarówno jednym z najistotniejszych, jak i najtrudniejszych problemów.
2) **Rozpoznawanie i przetwarzanie mowy (Speech Recognition/Processing).** Podobnie jak tłumaczenie maszynowe, także i ten problem był jednym z pierwszych, z jakimi próbowano się zmierzyć, i również uważany jest za niezwykle złożony. Obejmuje zadania takie jak rozpoznawanie ludzkiej mowy, jej transkrypcję na tekst pisany, syntezę mowy.
3) **Optyczne rozpoznawanie pisma (Optical Character Recognition – OCR).** Systemy OCR mają za zadanie przetworzyć tekst przedstawiony graficznie na tekst pisany. Jest to szczególnie ważne zadanie towarzyszące skanowaniu tekstów.
4) **Rozumienie języka (Natural Language Understanding).** Polega na przypisywaniu tekstowi bardziej abstrakcyjnej reprezentacji logiczno-semantycznej, możliwej do interpretacji przez system komputerowy. Jest to konieczny element każdego systemu porozumiewającego się z użytkownikiem za pomocą języka naturalnego.
5) **Synteza języka (Natural Language Generation)**. Dotyczy prób automatycznego tworzenia wypowiedzi w języku naturalnym. Najczęściej są to systemy modelujące język z dobrze określonej dziedziny (np. prognoza pogody, tworzenie żartów), często są wykorzystywane do przekazywania informacji zawartych w bazach danych w sposób bardziej przystępny dla odbiorcy.
6) **Analiza morfologiczna/lematyzacja (Morphological Analysis / Lemmatization).** Pełni bardzo ważne zadania przy każdym nieco bardziej zaawansowanym przetwarzaniu tekstu. Obejmuje rozpoznawanie form tekstowych wyrazów, wyróżnianie tematu, formantów fleksyjnych i słowotwórczych (ang. *stemming*). Lematyzacja polega na znajdowaniu formy podstawowej danego wyrazu.
7) **Analiza składniowa (Parsing).** Analizatory składniowe (parsery) mają za zadanie przedstawić gramatyczną strukturę frazy lub zdania. Parsowanie płytkie (ang. *shallow parsing*) obejmuje analizę niewielkich fragmentów tekstu i wyróżnienie podstawowych struktur składniowych (np. połączenie czasownika i jego dopełnienia

lub grupę nominalną), głębokie (ang. *deep parsing*) bierze pod uwagę całe zdanie, tworząc pełną analizę składniową.

8) **Znakowanie częściami mowy (Part-of-speech Tagging).** Zwane także od angielskiego pierwowzoru tagowaniem, polega na przypisaniu analizowanemu wyrazowi właściwej części mowy.
9) **Ujednoznacznianie / dezambiguacja (Word Sense Disambiguation).** Określanie, w przypadku homonimicznych lub polisemicznych form wyrazowych, w którym ze znaczeń występuje wyraz w danym miejscu w tekście. Przykładowo, ciąg znaków *mam* ma w języku polskim dwie możliwe interpretacje: może być to forma czasownika *mieć lub* rzeczownika *mama*. Zadaniem systemu ujednoznaczniającego jest wskazanie odpowiedniego wyrazu.
10) **Ekstrakcja informacji (Information Extraction).** Obejmuje wiele zadań dotyczących wyodrębniania z tekstu różnego typu danych – np. nazw własnych, relacji semantycznych, interesujących faktów, danych liczbowych itp.
11) **Wyszukiwanie informacji (Information Retrieval).** Dotyczy wyszukiwania zadanej informacji w dużej grupie dokumentów lub bazie danych. Przykładem systemów wyszukiwania informacji mogą być wyszukiwarki internetowe. Obszar ten często traktowany jest jako odrębna wobec lingwistyki komputerowej dziedzina.

3. Znaczenie wielosegmentowych jednostek leksykalnych w lingwistyce komputerowej

Zagadnienie nieciągłych jednostek leksykalnych ma duże znaczenie dla lingwistyki komputerowej. Odpowiednie rozpoznawanie i przetwarzanie takich struktur językowych jest ważne dla wielu obszarów z tej dziedziny. Poniższa lista przedstawia niektóre z dziedzin, dla których wiedza o wielosegmentowych wyrażeniach ma duże znaczenie:

1) **Tłumaczenie maszynowe**. Każdy język dysponuje zasobem wielowyrazowych jednostek, których dosłowne tłumaczenie na inny język da w rezultacie nieprawidłowe lub brzmiące niezręcznie wyniki. Dla przykładu, tłumaczenie wyrażenia *łódź podwodna* na **underwater boat*, choć dające się zrozumieć, na pewno nie jest odpowiednie. Idiomy, specjalistyczne terminy czy specyficzne dla każdego języka sformułowania muszą być przekładane we właściwy sposób. Systemy tłumaczące, które potrafią radzić sobie z jednostkami wielosegmentowymi, są precyzyjniejsze, mniej podatne na błędy, a wyniki tłumaczenia brzmią bardziej naturalnie (por. Orliac i Dillinger 2003, Baldwin i Kim 2010, Ramisch 2012, Bungum i in. 2013).
2) **Wspomagana komputerowo leksykografia**. Nieciągłe jednostki leksykalne są uwzględniane w wielu słownikach ogólnych, słownikach frazeologicznych, słownikach kolokacji itd. Automatyczne pozyskiwanie takich jednostek może w dużym stopniu ułatwić pracę leksykografów, a także dostarczyć dowodów na słuszność niektórych rozstrzygnięć (np. wskazać, na podstawie badań korpusowych, częściej używaną formę danego wyrażenia – zob. Żmigrodzki 2009). Może być także podstawą specjalistycznych słowników – np. COBUILD (Sinclair 1995).
3) **Składniowa analiza tekstu**. Gramatyczny parser (analizator składniowy) dzięki wyrażeniom wielosegmentowym może uzyskać poprawniejsze wyniki – zarówno przy określaniu części mowy, jak i przy pełnej analizie gramatycznej zdania (por. Nivre i Nilsson 2004, Zhang i in. 2006, Green i in. 2011).

4) **Semantyczna analiza tekstu**. Wiedza o wyrażeniach wielosegmentowych pozwala zmniejszyć podatność na błędy interpretacyjne dotyczące zarówno pojedynczych wyrazów (jednostki złożone są dużo rzadziej wieloznaczne niż zwykłe wyrazy – zob. Finlayson i Kulkarni 2011), jak i większych struktur (znaczenie całości wyrażenia wielosegmentowego często jest inna niż jego poszczególnych składników). Ponadto wiedza taka pomaga w wykrywaniu zmian tematu w obrębie tekstu (zob. Ferret 2002) i szukaniu synonimów (Turney 2001, Terra i Clarke 2003).
5) **Automatyczna synteza tekstu**. Podobnie, jak w tłumaczeniu maszynowym, system tworzący wypowiedzi w języku naturalnym musi umieć wykorzystywać złożone jednostki leksykalne, aby efekty jego pracy były bardziej odpowiednie i naturalnie brzmiące dla ludzkiego odbiorcy. Przykładowo, z dwóch właściwie synonimicznych wyrażeń **silna herbata*, *mocna herbata* system wyposażony w odpowiednią wiedzę wybierze to drugie (por. Smadja 1991, Pearce 2001, Lareau i in. 2011).
6) **Wyszukiwanie informacji**. Systemy przeszukujące duże zasoby tekstowe, takie jak wyszukiwarki internetowe, przy interpretacji zapytania powinny traktować – i najczęściej traktują – wyrażenia wielowyrazowe tworzące jedną semantyczną całość jako samodzielne jednostki (por. Evans i in. 1991, Mitra i in. 1997).
7) **Nauka języka obcego**. Według Seretan (2011b) wyrażenia wielowyrazowe (kolokacje) są – z punktu widzenia użytkownika języka – łatwe do odkodowania (odczytania, zrozumienia), natomiast trudne do zakodowania (tworzenia i stosowania w rozmowie). Z tego powodu stanowią jeden z zasadniczych problemów dla osób poznających nowy język. Dlatego systemy wspomagające naukę języka obcego muszą dysponować wiedzą o jednostkach tego typu (Pearce 2001).
8) **Dezambiguacja** (określanie, w którym z możliwych znaczeń dany homonim występuje w konkretnym przypadku). Homonimiczne wyrazy mają silną tendencję do występowania w tym samym znaczeniu w zależności od wyrazów występujących obok nich („jeden sens w jednej kolokacji" – zob. Yarowsky 1995). Można zatem powiązać określone kolokaty homonimu ze znaczeniem, w jakim występuje.

Powyższa lista nie wyczerpuje oczywiście wszystkich możliwości. Specyficznych zastosowań wyrażeń wielowyrazowych w lingwistyce komputerowej jest o wiele więcej: usprawnienie systemu optycznego rozpoznawania pisma (Church i Hanks 1990), automatyczne streszczanie tekstu (Seretan 2011a), identyfikacja granic zdania z wykorzystaniem wielosegmentowych skrótów (Kiss i Strunk 2002), analiza opinii (Moreno-Ortiz i in. 2013), analiza dyskursu (Pęzik 2009) i wiele innych.

4. Stan badań

4.1. Korpusy

Metody automatycznej ekstrakcji wyrażeń wielowyrazowych, podobnie jak wiele innych zadań w lingwistyce komputerowej, opierają się na analizie dużej ilości tekstu, najczęściej wykorzystując techniki statystyczne. Ich rozwój umożliwiły więc dwa czynniki: wzrost możliwości komputerów i pojawienie się komputerowych korpusów tekstów. Korpus to zbiór tekstów wykorzystywany przy badaniach lingwistycznych. Taki zbiór jest przeważnie duży lub bardzo duży (brytyjski korpus **Bank of English** w 2012 roku liczył

650 milionów wyrazów) i często zawiera dodatkowe informacje (metadane), dotyczące zarówno poziomu tekstu (np. typ – pisany, mówiony; styl – publicystyczny, naukowy itp.; struktura – podział na akapity i zdania), poziomu zdania (gramatyczna struktura zdań), jak i poziomu wyrazu (informacje morfosyntaktyczne). Korpusy jako narzędzia pracy językoznawcy powstawały od dawna, jednak zarówno ich rozmiar, jak i użyteczność znacząco zwiększyły się wraz z pojawieniem się komputerów. Dzięki połączeniu tych dwóch elementów możliwa stała się efektywna praca nad przetwarzaniem dużych ilości danych tekstowych, a co za tym idzie – zastosowaniem na szeroką skalę metod statystycznych. Dało to początek nurtowi badawczemu zwanemu lingwistyką korpusową[3]. Istnieje bardzo dużo korpusów komputerowych, zróżnicowanych pod względem języka, rozmiaru, składu i dodatkowych informacji.

Najliczniejsze są korpusy anglojęzyczne, wśród których warto wymienić te najczęściej stosowane do ekstrakcji jednostek wielosegmentowych:

- *Brown Corpus* (Francis i Kucera 1964). Pierwszy powszechnie używany korpus elektroniczny, zawierający 500 próbek tekstów z różnych gatunków, liczący około miliona słów, z których każdemu przypisana jest część mowy.
- *Bank of English*. Rozwijany przez Uniwersytet Birmingham i wydawnictwo Collins, liczył w 2012 roku 650 milionów słów. Stał się podstawą bardzo rozbudowanego i popularnego słownika *Collins Cobuild English Language Dictionary* (Sinclair 1995).
- *British National Corpus*. Stworzony w latach 1991-1994 korpus liczy 100 milionów słów i składa się z tekstów pisanych i mówionych reprezentujących różne gatunki.
- *Corpus of Contemporary American English*. Zawiera 450 milionów słów i jest największym korpusem zawierającym teksty w amerykańskiej odmianie języka angielskiego.

Wiele krajów dysponuje także korpusami we własnym języku. Często nazywane są one „narodowymi", zwłaszcza jeśli mają ambicję być korpusami reprezentatywnymi – tzn. próbującymi jak najwierniej oddać współczesny im stan danego języka.

Pierwszym polskim korpusem jest korpus utworzony na potrzeby *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz i in. 1990). Tworzenie korpusu i opartego na nim słownika frekwencyjnego rozpoczęto w 1967 roku, lecz ukończenie projektu się opóźniało – ostatecznie prace nad nim trwały ponad dwadzieścia lat. Omawiany korpus składa się z krótkich, liczących około 50 wyrazów, próbek tekstów; jego sumaryczna wielkość wynosi 500 tysięcy słów, przy czym podzielony jest on na pięć równych części, z których każda zawiera teksty reprezentujące różne style (teksty popularnonaukowe, krótkie notatki prasowe, publicystyka, proza artystyczna, dramat artystyczny). W latach 1974-1977 opublikowane zostały listy frekwencyjne oparte na poszczególnych częściach korpusu, całość wydano w roku 1990. Projekt *Słownika* jest – z punktu widzenia lingwistyki komputerowej – o tyle ciekawy, że już w chwili rozpoczęcia planowano, że większość prac nad słownikiem zostanie przeprowadzona za pomocą komputerów. Materiały źródłowe zostały zakodowane na taśmach papierowych. W latach dziewięćdziesiątych zaczęła powstawać uwspółcześniona wersja korpusu, dostępna w postaci elektronicznej, w której dokonano licznych korekt, a także

[3] Wiele informacji na temat lingwistyki korpusowej i korpusów, zwłaszcza z punktu widzenia języka polskiego, można znaleźć w pracy (Rudolf 2004).

dodano informacje gramatyczne (zob. Bień i Woliński 2003, Ogrodniczuk 2003). Mankamentem korpusu, na który zwracają uwagę sami autorzy (Kurcz i in. 1990, VIII), jest to, że teksty, które posłużyły do jego budowy, pochodzą z lat 1963-1967, jest to zatem korpus nieco przestarzały.

Istnieje kilka dużych, współczesnych korpusów języka polskiego, z których najważniejsze to:

- *Korpus PWN* (Łaziński 2000). Tworzony od 1995 roku, obecnie zawiera około 100 milionów słów, z czego podkorpus wielkości 40 milionów jest dostępny w Internecie.
- *Narodowy Korpus Języka Polskiego* (Przepiórkowski i in. 2012). Częściowo składa się z zasobów powstałych wcześniej (korpus IPI PAN, fragmenty korpusów PELCRA i PWN). Jest to największy polski korpus, o wielkości 1,8 miliarda słów, a ściślej: segmentów, gdyż do podstawowych jednostek w korpusie zaliczają się, oprócz wyrazów, niektóre łączniki (np. wyrażenie *biało-czerwony* to trzy segmenty), niektóre partykuły (np. *by*, *-li*) i tzw. formy aglutynacyjne czasownika być (*-śmy*, *-ś* itd.) (por. Przepiórkowski i in. 2012). Korpus składa się z tekstów pisanych i mówionych, reprezentujących różne style (m.in. publicystyczny, artystyczny, literatura faktu) i pochodzących z różnych źródeł (m.in. prasy, Internetu, książek). Teksty opatrzone są metadanymi, czyli informacjami o autorze, tytule, dacie pozyskania itp. Każdy wyraz jest oznakowany morfosyntaktycznie: przypisany mu jest tzw. tag, na który składa się forma bazowa wyrazu, część mowy – a konkretnie fleksem (por. Bień 1991), a także specyficzny dla każdego fleksemu zestaw kategorii gramatycznych (np. liczba, rodzaj, przypadek dla rzeczownika).

4.2. Metody ekstrakcji wyrażeń wielowyrazowych

Automatyczna ekstrakcja jednostek wielowyrazowych nie jest zadaniem trywialnym. Przede wszystkim nie istnieją własności ściśle językowe, których identyfikacja pozwalałaby na efektywne i precyzyjne wyodrębnianie ich z tekstu (Todirascu i Gledhill 2008). Ponadto wyrażenia takie cechuje duża rozmaitość, jeśli chodzi o cechy morfologiczne, składniowe, semantyczne i pragmatyczne (Graliński i in. 2010, Baldwin i Kim 2010). Mimo to (a może właśnie dzięki temu) wiele wysiłku badawczego włożono w rozwijanie technik ekstrakcji wyrażeń wielowyrazowych za pomocą automatycznych narzędzi. Zasadniczo można wyróżnić dwa rodzaje podejścia: statystyczne, wykorzystujące wyłącznie dane o częstotliwości występowania poszukiwanych wyrażeń i wyrazów składowych, oraz hybrydowe, łączące dane statystyczne z wiedzą lingwistyczną (Villada Moirón 2005).

4.2.1. Etapy ekstrakcji

Istnieje duża rozmaitość podejść do ekstrakcji jednostek wielosegmentowych, stosuje się też wiele technik lingwistycznych, statystycznych i informatycznych, niejednokrotnie przenikających i nakładających się na siebie – z tego powodu wyróżnianie ogólnych etapów pracy jest siłą rzeczy trochę sztuczne. Mimo to zdecydowana większość prac może być podzielona na trzy ogólne etapy: przygotowanie korpusu, wstępna selekcja „kandydatów na jednostki", sortowanie i filtrowanie otrzymanej listy (Evert 2005).

4.2.2. Przygotowanie korpusu

Przetwarzanie wstępne (ang. *pre-processing*) polega na przygotowaniu korpusu do ekstrakcji. W zależności od stopnia anotacji (dodatkowych informacji towarzyszących wyrazom lub zdaniom) korpusu i potrzeb danej metody może to być segmentacja tekstu na wyrazy i zdania, lematyzacja, znakowanie morfosyntaktyczne, analiza składniowa (por. np. de Caseli i in. 2009). Naturalnie, jeśli korpus, który służy badaniom, jest już odpowiednio przygotowany, ten etap jest pomijany.

4.2.3. Wstępna selekcja kolokacji

W drugim etapie z korpusu pozyskiwane są grupy wyrazów będące kandydatami na jednostki wielosegmentowe.

Wyodrębnianie wstępnej listy kandydatów może być przeprowadzane na różne sposoby. Popularną i najwcześniej stosowaną techniką (por. np. Berry-Rogghe 1973, 1974; Choueka 1988) jest ekstrakcja wszystkich n-gramów[4] z korpusu: jest to tzw. metoda okienkowa – nazwa pochodzi od stosowania „okna" o zdefiniowanej szerokości (ilości wyrazów), które „przesuwane" jest wzdłuż tekstu, a jego zawartość zapisywana. Na tym działaniu bazuje inna popularna metoda – wyszukiwanie wzorców – struktur złożonych z określonych części mowy, np. V-PP, V-N, A-N itp. Metoda ta jest szczególnie użyteczna z dwóch powodów: po pierwsze, w związku z dużą ilością syntaktycznych typów wyrażeń wielowyrazowych, przytłaczająca ilość prac skupia się na wyodrębnianiu raczej jednej lub kilku struktur niż na próbie wyszukiwania wszystkich możliwych wyrażeń (są oczywiście wyjątki – np. Zhang i in. 2006) – w takim wypadku wzorce syntaktyczne od razu zapewniają selekcję połączeń, na których skupia się dana praca; po drugie, ograniczenie listy n-gramów tylko do uprzednio zdefiniowanych struktur pozwala na znaczące zwiększanie efektywności ekstrakcji, zwłaszcza jeśli technika bazuje na zwykłym zliczaniu frekwencji wyrazów, bez stosowania bardziej zaawansowanych metod statystycznych (por. Justeson i Katz 1995b, Manning i Schütze 1999). Zadanie to często jest dosyć proste, jeśli wzorce są niezbyt skomplikowane, a pod uwagę brane są wyrazy sąsiadujące ze sobą. Często realizowane jest to za pomocą wyrażeń regularnych lub prostych automatów skończenie stanowych (Kupiec 1993, Daille 1994). W przypadku bardziej zaawansowanych struktur lub wyszukiwania par wyrazów niekoniecznie występujących obok siebie, zadanie się komplikuje – najczęściej stosowany jest wtedy płytki parsing (*chunking*), polegający na wyszukiwaniu podstawowych struktur syntaktycznych typu: frazy nominalne, połączenia czasownik-wyrażenie przyimkowe itp., bez analizy całego zdania (por. m.in. Bourigalt 1992, Paulo i in. 2002, Ritz 2006). Odmienne podejście do wzorców części mowy prezentuje praca (Dias 2003): zamiast wyszukiwać zdefiniowane wcześniej struktury syntaktyczne, najbardziej charakterystyczne wzorce wyszukiwane są za pomocą algorytmu LocalMax (Silva i in. 1999), który wykorzystując metody statystyczne, wyszukuje połączenia wyrazów ściślej ze sobą złączonych niż ich otoczenie.

[4] N-gram (lub ngram) to sekwencja następujących po sobie n elementów (rzadziej spotyka się definicje, w myśl których elementy nie muszą następować bezpośrednio po sobie). Elementami w zależności od zastosowania mogą być fonemy, litery, sylaby lub – jak w tym przypadku – wyrazy. N-gramy, dla których n wynosi 2 lub 3 (a więc sekwencje dwóch lub trzech kolejnych elementów), są zwyczajowo nazywane odpowiednio bigramami i trigramami. Interesujące uwagi na temat nazewnictwa n-gramów można znaleźć w (Manning i Schütze 1999, 193).

Ekstrakcja n-gramów za pomocą techniki okienkowej jest wygodna i daje dobre efekty, ma jednak wady: nie bierze pod uwagę relacji składniowych lub – w przypadku płytkiego parsowania – nie wykrywa głębszych relacji składniowych. Może też mylnie uznać sekwencję słów będących w niewielkiej odległości i niezwiązanych bezpośrednią relacją syntaktyczną za realizację szukanej struktury. Należy też zwrócić uwagę na fakt, że maksymalny rozmiar okna jest ograniczony zarówno przez względy lingwistyczne (im większa odległość jest brana pod uwagę, tym większa ilość językowego „szumu" i, w konsekwencji, możliwość błędnej interpretacji – por. Villada Moirón 2005, 51), jak i praktyczne: wraz ze zwiększeniem szerokości okna wzrasta koszt obliczeniowy algorytmów (Choueka i in. 1983 – cyt. za Seretan i in. 2003). Z tego powodu nie spotyka się prac, w których rozmiar okna jest większy niż 6 wyrazów. Konsekwencją tego jest pomijanie konstrukcji, których człony dzieli większa odległość[5] – nie jest to problem bardzo palący w przypadku języków o względnie sztywnym szyku wyrazów, jak język angielski, natomiast dla języków o swobodniejszej strukturze (jak język polski) może stać się przyczyną zauważalnie zmniejszonej skuteczności. Rozwiązaniem tych problemów staje się analiza zależnościowa (*dependency parsing*), dająca jako rezultat zbiór trójek (<*wyraz1, relacja składniowa, wyraz*2>) lub pełna analiza składniowa (*full/deep parsing*). Pozwalają one na wykrywanie zaawansowanych relacji i operacji składniowych (np. struktury predykatowo-argumentowej, pasywizacji itd.), a w przypadku głębokiego parsowania również na całościowe określenie struktury syntaktycznej zdania. Szczegółowe informacje na temat parsowania i jego zalet przy ekstrakcji jednostek wielosegmentowych można znaleźć w pracy (Seretan 2011b); praktyczne zastosowanie parsera zależności m.in. w pracach (Lin 1998, Lü i Zhou 2004, Heid i Weller 2010), zastosowanie parserów głębokich m.in. w (Pearce 2001, Green i in. 2011 – parser statystyczny, Goldman i in. 2001, Seretan i in. 2003, Seretan i in. 2004, Seretan i Wehrli 2009, Seretan 2011b – parser symboliczny).

4.2.4. Ocena, sortowanie i filtrowanie listy kandydatów

Otrzymana lista kandydatów na jednostki wielosegmentowe jest w kolejnym etapie oceniana za pomocą różnych metod, i na tej podstawie tworzony jest ranking wyrażeń o największym prawdopodobieństwie przynależności do zbioru kolokacji lub wyznaczane są dwa zbiory: w zależności od tego, czy ocena wyrażenia lokuje się powyżej ustalonego progu, kandydaci są przypisywani do zbioru kolokacji lub zbioru zwykłych połączeń. W poniższej sekcji opisane zostaną wykorzystywane w literaturze podejścia do oceny kandydatów.

4.2.4.1. Miary asocjacji oparte na frekwencji wyrażeń

Najczęstszym sposobem oceniania jest wykorzystywanie do tego miar asocjacji (ang. *association measures*) – metod opierających się na statystycznych danych badanych wyrażeń. Najprostszą miarą asocjacji jest częstość występowania (frekwencja) danego wyrażenia w tekście (korpusie), zastosowana np. przez Chouekę (1988). Najczęściej stosowane są jednak bardziej zaawansowane miary, biorące pod uwagę również frekwencję brzegową – czę-

[5] Odległość między członami relacji syntaktycznej w niektórych językach może być zaskakująco duża: Evert (2008, 13) podaje przykład zdania w języku niemieckim znajdującego się w korpusie, w którym podmiot od orzeczenia dzieli 15 wyrazów, Goldman i in. (2001 – cyt. za Seretan 2011b) piszą o odległości 30 wyrazów.

stości występowania poszczególnych członów wyrażenia – mierzące stosunek frekwencji oczekiwanej i frekwencji rzeczywistej. Miar asocjacji proponowanych do ekstrakcji wyrażeń wielowyrazowych jest dużo: Pecina i Schlesinger (2006) wymieniają ich ponad 80. Choć niektóre badania (np. Evert i Krenn 2005) sugerują, że skuteczność różnych miar nie jest absolutna (zależy m.in. od typów wyrażeń, rodzaju i wielkości korpusu), kilka miar jest zdecydowanie preferowanych przez badaczy:

- wskaźnik z (z-score) – wykorzystywany we wczesnych pracach (Berry-Rogghe 1973, Berry-Rogghe 1974, Smadja 1993),
- informacja wzajemna (*Mutual Information* – MI), czyli miara pochodząca z teorii informacji. Jej najczęściej stosowana wersja – *Pointwise Mutual Information* (PMI) została pierwszy raz użyta w pracy (Church i Hanks 1990) i szybko zyskała na popularności (m.in. Smadja 1992, Lin 1998, Khokhlova 2008),
- test t Studenta (t-score) – zaproponowany w pracy Church i in. (1991), stosowany m.in. w pracach (Breidt 1993, Krenn 2000, Yagunova i Pivovarova 2010),
- logarytm wskaźnika wiarygodności (*log-likelihood ratio* – LLR) – zaproponowany przez T. Dunninga (1993), używany m.in. w pracach (Lin 1999, Lü i Zhou 2004, Todirascu i Gledhill 2008, Heid i Weller 2010).

Michelbacher i in. (2007) zwraca uwagę na to, że stosowane miary są symetryczne, tzn. nie zależą od kolejności wyrazów, i przedstawia dwie miary asymetryczne: jedną opartą na prawdopodobieństwie warunkowym, drugą bazującą na teście chi-kwadrat Pearsona.

Miary asocjacji w swojej podstawowej wersji przeznaczone są przede wszystkim do oceny bigramów (par wyrazów) – z tego powodu przytłaczająca większość prac na ten temat dotyczy właśnie bigramów. Podstawowy sposób na ominięcie tego ograniczenia polega na podejściu iteracyjnym: w pierwszym etapie za pomocą wybranej miary oceniane są bigramy, następnie tworzone są pary bigram + kolejny wyraz (a więc trigramy – trójki), które są oceniane tą samą metodą itd. (por. Evert 2005).

4.2.4.2. Miary związane z odległością i długością wyrażeń

Miary opisane we wcześniejszej sekcji badają fakt współwystępowania pewnych wyrazów, co często jest odzwierciedleniem ich frazeologiczności. Inną dającą się mierzyć cechą jest odległość dzieląca wyrazy tworzące kolokację (tj. ilość wyrazów, które rozdzielają człony kolokacji). Geffroy i in. (1973 – cyt. za Oakes 1998) badają kolokacje w języku francuskim za pomocą miary C, Hardcastle (2005) przedstawia *Inverse Distance Measure* (miarę odwróconego dystansu), Washtell i Markert (2009) proponują miarę *Co-Dispersion* i jej warianty (m.in. wersję asymetryczną). Wszystkie wymienione funkcje oceniające siłę kolokacyjności na podstawie odległości biorą pod uwagę średni dystans między wyrazami składowymi i frekwencję występowania wyrażenia i/lub jego składowych, różniąc się właściwie tylko w szczegółach.

Mierzyć można również długość badanych wyrażeń. Kita i in. (1994) stosują tzw. kryterium kosztu (*cost criterion*), które porównuje za pomocą odpowiedniego wzoru częstość występowania sekwencji słów o długości *n* z częstością występowania sekwencji o długości *n+1*, *n+2* itd. Zestawienie wyników uzyskanych dzięki tej mierze z wynikami otrzymanymi za pomocą informacji wzajemnej pokazuje, że kryterium kosztu jest szczególnie użyteczne przy wyodrębnianiu popularnych sformułowań typu *thank you very much* (‘dziękuję bardzo’) czy *is that so?* (‘naprawdę?’) – a więc fraz istotnych w nauce języka obcego.

4.2.4.3. Miary oparte na restrykcjach leksykalnych i syntaktycznych

Baldwin i Kim (2010) pokazują, że nieodłączną cechą większości jednostek wieloczłonowych jest pewnego rodzaju idiomatyczność (rozumiana jako specyfika wyrażenia) na jednym (lub, częściej, kilku) z poziomów: semantycznym, leksykalnym, składniowym, pragmatycznym i statystycznym. Miary opisane w podpunkcie 5.2.4.2 skupiają się na badaniu idiomatyczności statystycznej, w tym miejscu zostaną opisane techniki badające ograniczenia leksykalne i syntaktyczne.

Pearce (2001) za pomocą pomiaru frekwencji, kwerendy słownikowej i wyników wyszukiwarki internetowej (AltaVista) mierzy tendencje określonych wyrazów do łączenia się w pary tylko z wybranymi wyrazami z grupy synonimów lub niemal-synonimów.

Fazly i Stevenson (2007) badają stopień niezmienności wyrażenia, mierząc jego leksykalną i syntaktyczną niezmienność. Stałość leksykalna badana jest za pomocą porównania PMI (*Pointwise Mutual Information* – zob. 5.2.4.2) danej frazy werbalno-nominalnej i uśrednionej PMI wyrażeń utworzonych za pomocą podstawienia za rzeczownik jego synonimów. Do pomiaru stałości składniowej używana jest dywergencja Kullbacka-Leiblera (porównująca skłonność danego wyrażenia do występowania w określonej konfiguracji syntaktycznej z zachowaniem zwykłych par czasownik-rzeczownik).

Van de Cruys i Villada Moirón (2007) również badają obiekty czasownik-rzeczownik i ich preferencje do łączenia się z konkretnymi wyrazami, prezentując dwie formuły mierzące te właściwości. Najpierw rzeczowniki dzielone są na grupy według podobieństwa semantycznego, następnie dla każdego czasownika liczone jest prawdopodobieństwo wystąpienia w połączeniu z danym rzeczownikiem. Pierwsza z miar pozwala na identyfikację rzeczowników, z którymi najchętniej łączy się badany czasownik, bez brania pod uwagę ich przynależności do grupy semantycznej. Druga miara bada preferencję czasownika do wyboru rzeczownika spośród wszystkich przynależących do określonej grupy. Im wyższa wartość miar, tym większa leksykalna stabilizacja danej frazy.

Kopotev i in. (2013) używają dywergencji Kullbacka-Leiblera do określania, która z kategorii gramatycznych (takich jak przypadek lub rodzaj) jest najbardziej charakterystyczna dla danej konstrukcji składniowej.

Zhang i in. (2006) prezentują miarę entropii permutacji (*permutation entropy* – PE), która opiera się na założeniu, że człony wielosegmentowej jednostki leksykalnej mają skłonność do występowania w ustalonym szyku, zatem jeśli dane korpusowe poświadczają występowanie wyrażenia tylko w wersji o jednoznacznej kolejności, może to świadczyć o kolokacyjności. Ramisch i in. (2008) proponują modyfikację tej miary – EPI (*Entropy of Permutation and Insertion* – entropia permutacji i wstawiania), biorącą pod uwagę wszystkie możliwe warianty składniowe wyrażenia.

4.2.4.4. Metody ekstrakcji wyrażeń idiomatycznych semantycznie (niekompozycyjnych)

Wyrażenia niekompozycyjne to takie, których znaczenie nie wynika z sumy znaczeń składników (np. *wieczne pióro*). Baldwin i Kim (2010) określają to jako idiomatyczność semantyczną i uważają za najważniejszą cechę jednostek wieloczłonowych. Ekstrakcja

wyrażeń tego typu[6] jest powiązana z miarami opartymi na restrykcjach językowych: miary te w dużym stopniu wskazują właśnie na niekompozycyjne połączenia.

Lin (1999) bada stopień niekompozycyjności wyrażeń za pomocą testu opartego na substytucji: mierzona jest informacja wzajemna danego połączenia i porównywana z informacją wzajemną wyrażeń powstałych przez podstawienie za jeden z członów innego wyrazu (np. *red tape* – 'biurokracja' – dosł. 'czerwona taśma' i *yellow tape* – 'żółta taśma'). Uzyskanie dla badanego zestawienia słów wyraźnie wyższej oceny niż dla połączeń powstałych dzięki substytucji sugeruje, że jest to jednostka niekompozycyjna (założenia te są jednak krytykowane – zob. Bannard i in. 2003).

Inne techniki stosowane przy identyfikacji niekompozycyjnych jednostek opierają się na porównaniu kontekstu, przy wykorzystaniu założenia, że kontekst, w jakim występuje wyrażenie niekompozycyjne jako całość, jest istotnie różne od kontekstu, w jakim występują jego człony. McCarthy i in. (2003) wykorzystują dane statystyczne i kilka heurystyk do pomiaru podobieństwa kontekstów. Stosowane są też bardziej zaawansowane metody: Schone i Jurafsky (2001), Baldwin i in. (2003), Katz i Giesbrecht (2006), Krčmář i in. (2013) stosują LSA (*Latent Semantic Analysis* – niejawna analiza semantyczna – zob. Landauer i Dumais 1997).

4.2.4.5. Metody oparte na translacji

Do ekstrakcji jednostek wielowyrazowych można wykorzystać również ich tłumaczenia. Jest to podejście spotykane stosunkowo rzadko, gdyż wymaga najczęściej dostępu do równoległego korpusu – tworzonego przez zbiory tekstów w jednym języku i ich tłumaczeń w innym. Takich korpusów jest niewiele i dla niewielu języków, co ogranicza uniwersalność metody. Zaletą podejścia translacyjnego jest jednak otrzymanie listy jednostek w dwóch językach (lub w kilku, jeśli metoda dotyczyła większej ilości języków) z jednoczesnym połączeniem ich w pary (trójki, n-tki...).

Kupiec (1993) w odpowiadających sobie zdaniach angielskich i francuskich wyszukuje frazy nominalne, a następnie próbuje je połączyć, przypisując każdemu z połączeń odpowiednie prawdopodobieństwo. Procedura ta powtarzana jest iteracyjnie przy każdym wystąpieniu danej pary.

Metodę porównywania przekładów można zastosować również do badania idiomatyczności semantycznej. Villada Moirón i Tiedemann (2006) używają równoległego korpusu wielojęzycznego do określenia niekompozycyjności wyrażeń wielowyrazowych. W tym celu mierzą entropię (nieprzewidywalność) tłumaczeń wyrażenia w języku holenderskim na inne języki, wychodząc z założenia, że im większa entropia – a zatem różnorodność tłumaczenia – tym większa szansa na to, że fraza jest semantycznie idiomatyczna.

Ciekawą własnością jednostek wielosegmentowych jest fakt, że tłumaczone na inne języki mogą mieć inną długość (por. Ramish 2012, 2). Własność tę wykorzystują de Caseli i in. (2009): algorytm wyszukuje w równoległym korpusie portugalsko-angielskim te jednostkowe struktury, w których liczba wyrazów się różni, po czym odfiltrowuje kandydatów o zbyt niskiej frekwencji lub reprezentujących wcześniej zdefiniowane struktury leksykalno-syntaktyczne.

[6] Ujmując rzecz precyzyjnie, część omawianych w tej sekcji prac nie skupia się bezpośrednio na ekstrakcji, bada natomiast stopień niekompozycyjności wyrażeń – czego prostą konsekwencją jest możliwość wykorzystania tych metod właśnie do ekstrakcji.

Lü i Zhou (2004) proponują metodę porównywania tłumaczeń angielsko-chińskich w oparciu o parser zależności, dwa niezależne (nierównoległe) korpusy jednojęzyczne i słownik dwujęzyczny. Metoda ta bazuje na spostrzeżeniu, że, mimo oczywistych różnic obu języków, relacje zależności wykazują wysoką zbieżność, zatem istnieje wysokie prawdopodobieństwo, że najbardziej prawdopodobne tłumaczenie rzeczywistej kolokacji będzie również poprawną właściwą kolokacją w drugim języku.

4.2.4.6. Ekstrakcja terminów

Terminy i wieloczłonowe jednostki leksykalne są do siebie zbliżone. Pojęcie terminu jest z jednej strony szersze – obejmuje zarówno pojedyncze wyrazy, jak i ich połączenia, z drugiej strony węższe – wieloczłonowe terminy stanowią podgrupę wyrażeń wielowyrazowych. Podgrupa ta jest charakterystyczna ze względu na kilka cech: podstawową cechą (wynikającą z definicji) jest przynależność do specjalistycznej dziedziny; Savary i in. (2012) wymieniają trzy dodatkowe: i) terminy są kategorią wysoce produktywną, ii) wiele z nich ma postać frazy nominalnej, iii) wiele jest zagnieżdżonych, tj. mogą składać się z krótszych terminów.

Wielosegmentowe jednostki terminologiczne można – z punktu widzenia ekstrakcji – traktować właściwie tak samo jak inne jednostki wielowyrazowe. Często ogólne systemy ekstrakcji są w stanie wyodrębniać również terminy, np. *mwetoolkit* (Ramish i in. 2010), a wiele prac poświęconych wyłącznie ekstrakcji terminów korzysta z identycznych technik, jak w przypadku jednostek ogólniejszych. System *LEXTER* (Bourigault 1992), korzystając z płytkiego parsowania, wyszukuje maksymalnie długie frazy nominalne, wśród których wyodrębnia za pomocą własnego systemu reguł podstawowe jednostki, które z dużym prawdopodobieństwem są terminami. Daille (1994) wyodrębnia francuskie terminy za pomocą kombinacji wzorców części mowy i miar asocjacji (m.in. MI, $\chi 2$, LLR). System Termight (Dagan i Church 1994) stworzony został jako pomoc w tłumaczeniach dokonywanych przez firmę telekomunikacyjną AT&T. Wyszukuje terminologiczne frazy nominalne za pomocą wyrażeń regularnych. Justeson i Katz (1995b), opierając ekstrakcję terminów wielowyrazowych o ich frekwencję występowania, osiągnęli godne uwagi wyniki dzięki zastosowaniu filtrów części mowy. Pantel i Lin (2001) używają miar asocjacji do ekstrakcji terminów w języku angielskim i chińskim.

Do ekstrakcji stosowane są także metody biorące pod uwagę cechy charakterystyczne terminów. Frantzi i in. (2000) przedstawiają dwie miary: C-Value i NC-Value, z których pierwsza bierze pod uwagę zagnieżdżanie terminów, druga opiera się na badaniu kontekstu (wyrazy występujące często w pobliżu terminów wybranych na podstawie C-Value podwyższają ocenę „terminologiczności" badanego połączenia wyrazów). Maynard i Ananiadou (1999) wykorzystują C-Value i NC-Value, dodając do oceny miarę semantycznego podobieństwa między znanymi już terminami a ocenianymi kandydatami na terminy. Podobna idea (założenie, że wyrazy występujące w pobliżu znanych już terminów mają większą szansę również być terminami) przyświeca projektowi TTC (Terminology Extraction, Translation Tools and Comparable Corpora – Blancafort i in. 2010, Heid i Gojun 2012), którego jednym z zadań jest ekstrakcja wyrażeń terminologicznych: system ekstrakcji na wstępie dysponuje niewielką ilością terminów z danej dziedziny, i na ich podstawie robot internetowy wyszukuje teksty, które dotyczą tej dziedziny. Następnie terminy są wyodrębniane za pomocą reguł lingwistycznych i statystycznych.

4.2.5. Dodatkowe metody podnoszące skuteczność

W niektórych pracach poświęconych jednostkom wielosegmentowym stosowane są dodatkowe techniki pozwalające na zwiększenie skuteczności ekstrakcji (najczęściej techniki te same w sobie nie są wystarczająco skuteczne, by stanowić podstawę wyodrębniania).

Najpopularniejsze jest stosowanie różnego typu heurystyk związanych z własnościami lingwistycznymi albo statystycznymi. W obrębie pierwszej grupy to m.in.: wykluczanie wyrażeń, w których skład wchodzą wyrazy często występujące, o niewielkiej wartości semantycznej (tzw. stopwords – np. wyrażeń ze spójnikiem *i*, takich jak *i kot*), bardzo rzadko stanowiące część kolokacji (por. np. Lin 1998, Seretan 2011b, Ramisch 2012); odrzucanie wyrażeń, w których rozkład odległości pomiędzy członami jest zbyt równomierny – a więc mało charakterystyczny (Smadja 1993); sprawdzanie kolejności słów w przypadku wyrażeń, które mają określony szyk (Baldwin i Villavicencio 2002); odrzucanie kandydatów ze zbyt luźnym semantycznie jednym z członów (Woźniak 2011). Gayen i Sarkar (2013) stosują szereg heurystyk językowych (m.in. przeciętną długość wyrazów, fakt wystąpienia wyrazów w odmienionej formie) jako cechy brane pod uwagę przez algorytm maszynowego uczenia się.

Wśród heurystyk statystycznych najpopularniejszą (i często niezbędną z uwagi na naturę badań statystycznych) praktyką jest odrzucanie wyrażeń, których frekwencja nie przekroczyła ustalonego progu (jego wysokość może wahać się w zależności od potrzeb metody i wielkości korpusu od 5 do 100 wystąpień danego połączenia wyrazów – por. Khoklhova 2008, Pecina 2008b). Spotykane są także modyfikacje miar w celu zneutralizowania jakiegoś niepożądanego czynnika lub silniejszego zaakcentowania jakiejś własności: przykładem może być wariant informacji wzajemnej, nazywany sześciennym stosunkiem asocjacji (cubic association ratio, często oznaczany jako MI3), który do pewnego stopnia likwiduje tendencję MI do faworyzowania związków wyrazowych rzadko występujących (por. Evert 2005).

Metodą dającą dobre efekty jest również zastosowanie maszynowego uczenia się (machine learning). Algorytm uczący się analizuje szereg cech charakteryzujących każde wyrażenie i na tej podstawie stara się samodzielnie ocenić, czy jest ono rzeczywistą kolokacją, czy zwykłym połączeniem wyrazów. Cechy mogą być wyłącznie statystyczne (np. Pecina 2008a) lub statystyczne i lingwistyczne zarazem (np. Gayen i Sarkar 2013).

Niekiedy jako pomoc w ekstrakcji wykorzystywany jest Internet. Jedną z metod tego typu jest użycie wyszukiwarki internetowej (zgodnie z ideą sieci jako korpusu – por. Kilgarriff i Grefenstette 2003). Pearce (2001) mierzy liczbę wyników wyszukiwarki AltaVista dla badanego wyrażenia i traktuje to jako jeden z wyznaczników jego kolokacyjności. Zhang i in. (2006) i Ramish i in. (2008) używają miar entropii permutacji (które nie wymagają wiedzy o wielkości korpusu) do oceniania wyrażeń za pomocą wyszukiwarek Yahoo i Google. Attia i in. (2010) badają różnice w postaci haseł wielowyrazowych w różnych wersjach językowych Wikipedii.

4.3. Ekstrakcja jednostek wielosegmentowych w języku polskim

Prac poświęconych automatycznej ekstrakcji polskich jednostek wielosegmentowych jest niewiele. Praca magisterska A. Buczyńskiego (Buczyński 2004) opisuje niektóre metody ekstrakcji kolokacji – jej pokłosiem jest program *Kolokacje* (dostępny w Internecie pod adresem http://www.mimuw.edu.pl/polszczyzna/kolokacje/index.htm,

wykorzystany w pracy Savary i in. 2012). Vetulani i in. (2008) przedstawia słownik kolokacji werbalno-nominalnych wraz z opisem ich częściowo automatycznej ekstrakcji. *Toposław* (Czerepowicka 2011) to narzędzie do opisu jednostek wieloczłonowych bazujące na formalizmie *Multiflex* (Savary 2009), ułatwiające półautomatyczne tworzenie słowników jednostek złożonych. Na tym narzędziu bazują rozwijane projekty *SEJF* (*Słownik Elektroniczny Jednostek Frazeologicznych*, Graliński i in. 2010) i *SEJFEK* (*Słownik Elektroniczny Jednostek Frazeologicznych z Ekonomii*, Savary i in. 2012).

W ramach projektu NKJP stworzony został moduł *Kolokator* (Pęzik 2012), który dokonuje automatycznej ekstrakcji kolokacji (rozumianych w sensie statystycznym, czyli wyrazów najczęściej ze sobą współwystępujących) za pomocą testu chi-kwadrat. P. Pęzik jest także autorem słownika kolokacji *HASK* udostępnianego w formie serwisu internetowego pod adresem *http://pelcra.pl/hask_pl*/. Umożliwia on dostęp do kolokacji wygenerowanych na podstawie korpusu NKJP (dla języka polskiego) i BNC (język angielski). Serwis udostępnia również przeglądarkę kolokacji i narzędzie *Kolozaurus* służące porównywaniu kolokacji i wykorzystujące reprezentację grafową. *HASK* opisany jest w pracach Pęzik (2013) i Pęzik (2014).

Istnieje także pewna liczba prac i projektów, które – choć nie skupiają się bezpośrednio na ekstrakcji – są zbliżone tematycznie do tego zagadnienia badawczego. Moszczyński (2007, 2010) proponuje taksonomię wyrażeń wielowyrazowych w języku polskim, biorącą pod uwagę względy praktyczne, Bański i Moszczyński (2008) omawiają sposoby reprezentacji idiomów w słowniku angielsko-polskim. Korzycki (2008) przedstawia mechanizm pozwalający na reprezentację w słowniku złożonych wyrazów potencjalnych (jak np. liczby czy połączenia przymiotników typu biało-czerwony) za pomocą transducerów skończenie stanowych. Lubaszewski (2009, 23-26) omawia postać i rolę jednostek wieloczłonowych w elektronicznym słowniku, a Rokitiański (2009, 69-78) – sposób ich reprezentacji w słowniku.

Ważnym (nie tylko z punktu widzenia jednostek wielowyrazowych) zasobem dla języka polskiego jest Słowosieć (Maziarz i in. 2012, Piasecki i in. 2009) – sieć semantyczna wzorowana na anglojęzycznym WordNecie (Miller 1995), zawierająca ponad 178 tysięcy jednostek leksykalnych, w tym również wielosegmentowych[7].

Rudolf (2004) opisuje metody automatycznej ekstrakcji kolokacji, ilustrując to przykładami z języka polskiego.

[7] *Polska Słowosieć* jest udostępniona w internecie pod adresem *http://plwordnet.pwr.wroc.pl/*, anglojęzyczna: *https://wordnet.princeton.edu/*. Warto zwrócić uwagę, że *Słowosieć*, zawierająca ponad 178 tysięcy jednostek, jest największą tego typu bazą na świecie.

II
Wielosegmentowa jednostka leksykalna

Problematyka wielosegmentowych jednostek leksykalnych jest obszerna i wieloaspektowa. W zależności od potrzeb i celów można rozpatrywać ją na różne sposoby, między innymi jako zagadnienie semantyczne, leksykalne, składniowe, pragmatyczne, kognitywistyczne czy psychologiczne. Bardzo trudne jest zarazem precyzyjne zdefiniowanie pojęcia, na tyle różnorodnego i płynnego, że wymyka się jednoznacznym przyporządkowaniom. W literaturze przedmiotu, zarówno językoznawczej, jak i tej z kręgu lingwistyki komputerowej, można znaleźć uderzająco wiele różnorodnych propozycji podejścia, definicji, klasyfikacji, kryteriów, punktów widzenia i terminów. Niektórzy zauważają, że w tej sytuacji trudno mówić o jednorodnym pojęciu, wiele zależy od przyjętych założeń i punktu wyjścia. Według M. Bańki „o sposobie rozumienia frazeologii decyduje założony przez badacza cel i przyjęta przez niego perspektywa poznawcza" (Bańko 2001: 149).

W takim metodologicznym gąszczu łatwo o niejasności i nieporozumienia. Tymczasem dla każdej pracy naukowej ogromnie ważne jest możliwie dokładne i precyzyjne opisanie przedmiotu badań. Celem tego rozdziału jest zatem omówienie problematyki jednostek wielowyrazowych z teoretycznego punktu widzenia, opisanie istniejących teorii, ujmujących zagadnienie na różne sposoby, dyskusja problemów oraz przedstawienie koncepcji jednostki wielosegmentowej przyjętej w tej książce.

Lingwistyka komputerowa w badaniach nad jednostkami wielowyrazowymi skupia się przede wszystkim na praktycznych aspektach tej kwestii, co skutkuje mniejszą wagą przykładaną do definicji. Wiele prac dosyć swobodnie podchodzi do określenia przedmiotu badań, poprzestając najczęściej na ogólnych, lakonicznych (czasem nawet jednozdaniowych) i w konsekwencji mało precyzyjnych objaśnieniach. Istnieją opracowania bardziej systematycznie analizujące ten temat (np. Baldwin i Kim 2010, McKeown i Radev 2000), jednak nawet one nie dostarczają narzędzi pozwalających na odpowiednie zgłębienie tematu. Dlatego, mimo że niniejsza książka wyrasta z nurtu lingwistyki komputerowej, musi sięgnąć po narzędzia i metody z dziedziny frazeologii – zarówno rodzimej, jak i obcej – które umożliwiają zdefiniowanie obszaru badań z zadowalającą precyzją.

1. Definicje i typologie

W dziedzinie frazeologii bezdyskusyjna jest jedynie jej złożoność. Na przestrzeni lat podejmowano liczne próby zdefiniowania i uporządkowania zagadnienia. Przedstawię po-

niżej część z nich, szczególnie istotną ze względów historycznych bądź merytorycznych. Należy jednak od razu zauważyć, że pojęcia definicji i klasyfikacji częściowo nakładają się tu na siebie. Jednostki wielowyrazowe charakteryzują się bowiem szeregiem cech semantycznych, gramatycznych i pragmatycznych, i cechy te (najczęściej pewien ich podzbiór, który zostaje w danej teorii uznany za najważniejszy) mogą składać się jednocześnie na ich definicję – lub być traktowane jako charakterystyczne, ale drugorzędne. W koncepcji S. Skorupki definicyjną rolę pełni na przykład kryterium semantyczne, podczas gdy kryterium formalne, na podstawie którego przeprowadzony został drugi podział związków frazeologicznych, jest właściwie drugorzędne (zob. Skorupka 1982). Warto także dodać, że jeśli jakąś klasyfikację jednostek wielowyrazowych można uznać za wyczerpującą (do czego dąży wszak większość badaczy prezentujących swoją koncepcję), tworzy ona – poprzez wskazanie wszystkich elementów należących do danej grupy – definicję; nawet jeśli definicja ta nie jest wyrażona *explicite*.

Omówione niżej podejścia do opisu frazeologii da się umownie podzielić na dwie grupy. Pierwsza z nich obejmuje teorie, które składają się na frazeologię zanurzoną w polskiej tradycji językoznawczej, druga – podejścia prezentowane przez lingwistykę komputerową i korpusową. Choć opisywana materia językowa w obu przypadkach jest niemal taka sama, grupy te oddziela wyraźnie widoczna granica: odwołują się do różnych założeń, metodologii, terminologii; akcentują nieco inne aspekty zjawiska. Przykładowo, w kręgu prac z lingwistyki korpusowej / komputerowej ważne i często szeroko dyskutowane jest pojęcie kolokacji, podczas gdy we frazeologii polskiej idea ta, nawet jeśli zauważana, traktowana jest raczej marginalnie (por. uwagę P. Żmigrodzkiego o konieczności stworzenia słownika kolokacji: Żmigrodzki 2009, 198). Nie bez znaczenia jest także podstawowy język, którego dotyczą analizy[8]: w tradycji polskiej są to przede wszystkim języki słowiańskie (głównie polski i rosyjski), w zachodniej – przede wszystkim angielski (rzadziej francuski lub niemiecki).

1.1. Frazeologia tradycyjna

1.1.1. Koncepcja Winogradowa

Prace W. Winogradowa były jednymi z pierwszych zawierających spójną koncepcję związku frazeologicznego i wywarły bardzo duży wpływ na badaczy w krajach słowiańskich, przyczyniając się do powstania nurtu umownie zwanego „winogradowskim" (zob. Chlebda 1991, 6). Koncepcja ta oparta jest na kryterium semantycznym: relacji pomiędzy znaczeniem globalnym całego wyrażenia a znaczeniem jego poszczególnych elementów (ujmując rzecz precyzyjniej – stopniu rozbieżności pomiędzy tymi znaczeniami). Ta cecha, zwana często asumarycznością znaczenia (por. Bogusławski 1989), niekompozycyjnością (por. Manning i Schütze 1999) lub idiomatycznością semantyczną (por. Baldwin i Kim 2010), jest kryterium dominującym w badaniach nad jednostkami wielowyrazowymi i według większości badaczy świadczy o frazeologiczności danego połączenia.

Winogradow, w zależności od stopnia korelacji, wyróżnia trzy grupy:

1) **sraščenija** – połączenia, w których znaczenie wyrażenia w żaden sposób nie wynika ze znaczeń jego składników i nie da się go wydedukować; połączenia takie są

[8] Większość zjawisk opisywanych przez obie tradycje jest uniwersalna i daje się zastosować do więcej niż jednego języka, jednak wybór języka (lub grupy języków) wyraźnie rzutuje na sposób opisu.

nieprzetłumaczalne i całkowicie ustabilizowane leksykalnie i gramatycznie (np. *koń by się uśmiał*),
2) **jedinstva** – znaczenie nie jest sumą znaczeń składników, ale jest przejrzyste, można się go domyślić, cechuje je wysoka (ale niekoniecznie całkowita) stabilność leksykalna i gramatyczna (np. *świecić przykładem*),
3) **sočetanija** – znaczenie wyrażenia jest sumą znaczeń składników, przy czym jeden z członów jest często użyty metaforycznie; występują ograniczenia łączliwości (np. *dojść do władzy*).

1.1.2. Koncepcja Skorupki

Pierwszą polską klasyfikację związków frazeologicznych przedstawił w latach sześćdziesiątych XX wieku S. Skorupka. Była ona oparta na kryteriach dwojakiego rodzaju. Pierwsze, nazwane przez autora semantycznym (w istocie oprócz asumaryczności bierze również pod uwagę stopień zespolenia związku), dzieli związki frazeologiczne na trzy grupy:
1) **związki stałe** – połączenia, których znaczenie nie jest sumą znaczeń składników i które zachowują się jak pojedyncze wyrazy, są sztywno zespolone (np. *czarna śmierć*),
2) **związki łączliwe** – wyrażenia, których stopień spoistości jest duży, ale istnieje możliwość modyfikacji związku w ograniczonym zakresie; przynajmniej jeden ze składników użyty jest w dosłownym znaczeniu (np. *różowy*, *dobry humor*),
3) **związki luźne** – połączenia tworzone doraźnie, o czytelnym i dosłownym znaczeniu (np. *jeść zupę*, *bułkę*). W zamierzeniu Skorupki ta grupa ma przede wszystkim stanowić informację uzupełniającą, ilustrującą sposób użycia wyrazu.

Drugie kryterium, zwane formalnym, dzieli związki ze względu na ich budowę. Tu również wyróżnione są trzy grupy:
1) **wyrażenia** – połączenia o charakterze grupy nominalnej (ośrodkiem jest rzeczownik lub przymiotnik – np. *czarna owca*, *wielce szanowny*), a także przyimkowe, przysłówkowe i spójnikowe (np. *pod pozorem*, *na bok*),
2) **zwroty** – połączenia, których ośrodkiem jest czasownik (np. *ruszać z kopyta*),
3) **frazy** – połączenia złożone z członów czasownikowych i rzeczownikowych (np. *głowa komuś pęka*) mające postać zdania i mogące funkcjonować jako samodzielne wypowiedzi.

Podział semantyczny proponowany przez Skorupkę nosi wyraźne podobieństwo do klasyfikacji Winogradowa, nie jest jednak z nim tożsamy; pojawiają się w nim na przykład bardzo liczne związki luźne, które według Winogradowa w ogóle nie wchodziłyby w zakres frazeologii (więcej na ten temat pisze Chlebda (1991, 11-12)).

Obie powyższe klasyfikacje, choć dosyć mocno rozpowszechnione i stanowiące często podstawę „szkolnego" rozumienia frazeologii, były poddawane krytyce. Wskazywano, że kryterium spójności semantycznej jest mało wymierne i przysparza problemów przy próbie precyzyjnego rozgraniczenia między poszczególnymi grupami, zwracano uwagę na fakt, że klasyfikacje te łączą wyrażenia o różnym statusie językowym. Wobec koncepcji Skorupki wysuwano także inne zarzuty – przede wszystkim krytykowano wyodrębnianie związków luźnych, które właściwie nie powinny mieć statusu frazeologizmów (por. Żmigrodzki 2009, 101); nie wszyscy jednak zgadzają się z tym zarzutem (zob. Chlebda 1991, 12).

1.1.3. Koncepcja Bogusławskiego

Koncepcja A. Bogusławskiego pod wieloma względami różni się od zaproponowanych powyżej. Bogusławski za prymarne zadanie językoznawstwa uważa wyróżnianie podstawowych jednostek języka. Nie jest to bynajmniej zadanie banalne, gdyż – według słów F. de Saussure'a – „jest rzeczą niezmiernie trudną wydobyć w języku dźwięków grę spotykających się w nim jednostek oraz powiedzieć, jakimi konkretnymi elementami operuje język" (Saussure de 1961, 114). Ponadto bardziej oczywiste jednostki – wyrazy – stanowią tylko część zasobu leksykalnego danego języka. W licznych pracach Bogusławski skupia się na stworzeniu i dopracowaniu takiej definicji jednostki, która najlepiej oddawałaby rzeczywistość językową i brała pod uwagę zarówno pojedyncze wyrazy, jak i całostki wielowyrazowe. Ten punkt widzenia powoduje, że zacierają się tradycyjne granice pomiędzy leksyką a frazeologią: „Nasz punkt widzenia każe naturalnie porzucić również przeciwstawienie »leksyki« i »frazeologii« jako rzeczy różnych w jakimkolwiek nietechnicznym sensie (…) Wyrazy graficzne czy choćby fonologiczne nie tworzą odrębnego świata w języku i nie podlegają jako takie żadnym operacjom składniowym lub »frazeologizacyjnym«, a tym bardziej operacjom realizującym jakieś »związki« w różnym stopniu. »Frazy« nie są żadnym osobliwym dodatkiem do leksyki o na poły składniowych własnościach. Cechy »jest jednostką języka« i »jest pojedynczym wyrazem« nie są skorelowane nawet statystycznie: większość jednostek języka to nie pojedyncze wyrazy" (Bogusławski 1976, 357).

Podstawowym pojęciem w teorii Bogusławskiego jest kryterium, które autor (1989, 18) nazywa kryterium „elementarnej składnikowości". Każda jednostka języka (oznaczona tutaj jako A) spełnia następujące warunki:

1) wchodzi w proporcjonalny pod względem formalnym i funkcjonalnym układ AC:AD = BC:BD,
2) co najmniej jedna z klas substytucyjnych, do których należą elementy (A, B) i (C, D), musi być otwarta – to znaczy, że istnieje możliwość zdefiniowania jej w sposób ogólny, a nie tylko poprzez wyliczenie elementów składowych,
3) jest elementem minimalnym, tj. nie daje się podzielić na mniejsze jednostki,
4) ma potencję przypisania referencji do rzeczywistego obiektu w konkretnym akcie mowy, co wyklucza przysłowia, których referencja jest *a priori* ustalona (dopuszczając wszakże tzw. „zwroty przysłowiowe" w rodzaju *góra urodziła mysz*).

Definicję tę, mocno syntetyczną i dosyć zawiłą, ilustruje Bogusławski przykładami. Proporcja *ta książka* : *ta sztywna okładka* = *inna książka* : *inna sztywna okładka*, gdzie A = *ta*, B = *inna*, C = *książka*, D = *sztywna okładka* jest prawdziwa, a obie klasy substytucyjne są otwarte: klasę, do której zaliczają się wyrazy *ta*, *inna*, można scharakteryzować ogólnie jako klasę wyrażeń wskazujących, a klasę zawierającą wyrażenia *książka*, *sztywna okładka* jako klasę rzeczowników konkretnych. Wszystkie, poza jednym, elementy proporcji spełniają też warunek minimalności (nie da się ich podzielić na mniejsze jednostki), zatem stanowią jednostki języka. Dla wyrażenia *sztywna okładka*, które można podzielić, procedurę należy powtórzyć. Tworzona jest zatem nowa proporcja *ładna okładka* : *ładny materiał* = *sztywna okładka* : *sztywny materiał*. Według Bogusławskiego proporcja ta nie jest spełniona, gdyż między *sztywna* i *ładna* nie ma w tym wypadku symetrii funkcjonalnej: wyraz *sztywna* posiada tu znaczenie wyspecjalizowane, mówiące o pewnym typie okładki.

Sformułowana przez Bogusławskiego definicja jednostki języka ma wiele zalet, jest ścisła i elegancka, dzieli wyrażenia na dwie dobrze rozgraniczone grupy: jednostki języka

i zwykłe połączenia wyrazów. Nie oznacza to jednak, że rozwiązuje wszystkie problemy z definiowaniem frazeologizmów. Po pierwsze, nie usuwa konieczności podejmowania arbitralnych decyzji w niektórych mniej oczywistych przypadkach („Przyznać trzeba, że konkretna analiza materiałowa nastręczać może w różnych miejscach trudności"– pisze sam Bogusławski (1987, 21)). Po drugie, niejednokrotnie wymaga dużej biegłości i intuicji językoznawczej: przy rozważaniu przypadków takich jak *brat przyrodni*, *przybrane dziecko* Bogusławski sięga po złożoną i subtelną argumentację (por. Bogusławski 1987, 27-32). Po trzecie, decyzje oparte na tym kryterium bywają kontrowersyjne, tak jak zaliczenie w poczet jednostek wyrażeń typu *pada deszcz / śnieg* czy uznanie ciągu *kary koń* jednocześnie za jednostkę i połączenie. Bańko (2001, 154-159) szerzej omawia te i inne problemy (zwłaszcza w kwestii otwartości klas substytucyjnych) związane z teorią Bogusławskiego.

1.1.4. Klasyfikacja Lewickiego

A.M. Lewicki wprowadził podział frazeologizmów (Lewicki 1986), opierając go częściowo na klasyfikacji formalnej Skorupki. Jednak w ujęciu Lewickiego podział opiera się nie na obecności w związku odpowiednich części mowy, lecz na funkcji, jaką pełni w zdaniu. Według tej typologii związki frazeologiczne dzielą się na:

- **frazy** – gotowe znaki językowe niewymagające uzupełnień, pełniące funkcję wypowiedzenia, np. *wyszło szydło z worka*, *i po krzyku*,
- **zwroty –** znaki fragmentaryczne, pełniące podstawowo rolę czasownika; aby samodzielnie funkcjonować, wymagają uzupełnienia o komponent nominalny (rzeczownik, przymiotnik, przysłówek), np. *<ktoś> zbija bąki*, *krew <kogoś> zalewa*,
- **wyrażenia rzeczownikowe –** jednostki frazeologiczne pełniące w zdaniu funkcje rzeczownika, np. *kocie łby*, *to i owo*,
- **wyrażenia określające –** wyrażenia będące odpowiednikami przymiotników i przysłówków, określające rzeczowniki, czasowniki, przymiotniki i przysłówki – np. *całą gębą*, *na amen*, *chcąc nie chcąc*,
- **wskaźniki frazeologiczne** – frazeologizmy pełniące funkcje pomocnicze: o charakterze przyimka (*w związku z <czymś>*, *w ramach <czegoś>* itp.), spójnika (*zarówno... jak i...*, *bądź... bądź...*), partykuły (*że też...*, *też mi...*).

Lewicki wyróżnia też dwie kategorie frazeologizmów ze względu na znaczenie: idiomy (o ustalonym składzie, asumaryczne znaczeniowo, np. *pies na baby*) i połączenia frazeologiczne (frazemy), w których jeden z członów dominuje semantycznie (np. *złodziej kieszonkowy*, *gniew ogarnia*). Rzeczowniki wielowyrazowe (obejmujące również terminy) typu *samochód ciężarowy*, *płatki owsiane* czy *władza państwowa* nazywane są zestawieniami i uznawane za byty z pogranicza frazeologii.

1.1.5. Frazematyka

Frazematyka to koncepcja autorstwa W. Chlebdy (m.in. Chlebda 1991, Chlebda 2001). Bierze ona za punkt wyjścia moment tworzenia wypowiedzi, opisuje frazeologię od strony nadawcy komunikatu. Jedynym kryterium służącym wyodrębnianiu podstawowych jednostek – frazemów – jest odtwarzalność danego wyrażenia (ściślej: odtwarzalność „w danej sytuacji", a więc biorąca pod uwagę kontekst społeczny (zob. Chlebda 1991, 115)). Odtwa-

rzalność ta wynika z faktu, że w leksykonie nadawcy wyrażenie to istnieje jako gotowa, odrębna całość, która w razie potrzeby jest przywoływana. Frazemy odróżniają się od zwykłych konstrukcji składniowych, powstających w wyniku łączenia mniejszych jednostek.

Idea oparcia koncepcji wyłącznie na jednym kryterium podyktowana jest krytyką wcześniejszych „analitycznych" prób definiowania frazeologii, które starały się osiągnąć precyzję opisu przez wyróżnianie wielu cech dystynktywnych frazeologizmów, co „jest słuszne i pożądane z teoretycznego, genologicznego i systemowego punktu widzenia, co jednak nie jest relewantne w płaszczyźnie wypowiedzi" i stoi „w opozycji do metodologicznej zasady »brzytwy Ockhama«" (Chlebda 1991, 9).

Zakres frazematyki wyznaczony przez kryterium odtwarzalności jest dość szeroki: frazemami są zarówno wyrażenia, które można zaliczyć do „konwencjonalnej" frazeologii, takie jak *spiec raka* czy *jak filip z konopi*, jak i sformułowania typu *dziękuję, nie mam więcej pytań*, *zawieje i zamiecie śnieżne*, *co słychać* itp. Frazemami są również przysłowia, powiedzenia i skrzydlate słowa.

1.2. Lingwistyka korpusowa i komputerowa

1.2.1. Koncepcja Mielczuka

Definicja i typologia frazeologizmów przedstawiona przez I. Mielczuka nie wywodzi się, ściśle rzecz biorąc, z kręgu lingwistyki korpusowej lub komputerowej, jednak stała się inspiracją i istotnym punktem odniesienia dla prac z tego nurtu, dlatego też omawiana jest w tym miejscu. To jedna z najbardziej spójnych i precyzyjnych koncepcji wyrażeń nieciągłych. Koncepcja ta wyrasta ze stworzonej przez Mielczuka teorii sens-tekst. Jest to model języka opisujący proces tworzenia wypowiedzi jako złożony z dwóch zasadniczych etapów. W pierwszym nadawca komunikatu, dysponując mentalnym obrazem danego elementu rzeczywistości, nazywanym przez autora ConceptR (od Conceptual Representation), przetwarza go w postać obiektu semantycznego – SemR (od Semantic Representation), za pomocą procesu nazywanego CMM (Concept-Meaning Model). W drugim etapie nadawca wypowiedzi analogicznie przechodzi od SemR do reprezentacji fonetycznej, zwanej PhonR (od Phonetic Representation), posługując się modelem MTM (Meaning-Text Model).

Dodatkowo Mielczuk w odniesieniu do procesu tworzenia wypowiedzi wprowadza dwa pojęcia pomocnicze: nieograniczoność i regularność. Nieograniczoność to możliwość tworzenia elementu wypowiedzi za pomocą dowolnie wybranych (spośród możliwych do zastosowania w danym przypadku) reguł języka. Przykładowo wyrażenie *no parking*[9] ('nie parkować') jest ograniczone, gdyż nie jest przyjęte wyrażanie tego samego konceptu w tej samej sytuacji w inny sposób (np. *you should not park here* – 'nie powinieneś tu parkować')[10]. Regularność to możliwość tworzenia wyrażenia w sposób zgodny z regułami danego języka. Wyrażenie *the chip on one's shoulder* ('ktoś zachowuje się agresywnie, nietaktownie') jest nieregularne, gdyż nie ma możliwości uzyskania przypisanego mu znaczenia za pomocą regularnego połączenia znaczeń jego składników. Nieograniczoność podobnie jak regularność mogą dotyczyć zarówno strony semantycznej, jak i formalnej znaku językowego.

[9] Jako że ogromna większość prac z lingwistyki komputerowej dotyczy języków innych niż polski, w trosce o precyzję pozostawiam przykłady z języka oryginalnego (choć w praktyce dla przeważającej większości przykładów można znaleźć polskie odpowiedniki, nierzadko będące dosłownymi tłumaczeniami).

[10] Co ciekawe, wyrażenie to w tłumaczeniu na język polski podlega mniejszym ograniczeniom – spotykane są formy *parkowanie wzbronione*, *parkowanie zabronione*, *nie parkować* itd.

Mielczuk na podstawie tych kryteriów dzieli wyrażenia języka na dwie główne grupy: frazy swobodne (*free phrases*) – z którymi mamy do czynienia w przypadku wyrażeń nieograniczonych i regularnych, oraz frazy związane (*set phrases*) lub frazemy, które otrzymujemy, gdy co najmniej jeden z powyższych warunków jest niespełniony. Frazy związane są podzielone na typy, w zależności od tego, która reguła jest naruszona:

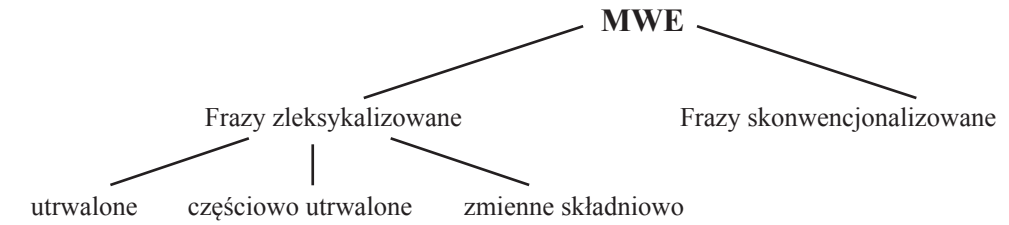
- **pragmatemy** – powstają, gdy warunek nieograniczoności w odniesieniu do SemR jest niespełniony, a więc SemR można otrzymać tylko na jeden sposób (lub spośród bardzo niewielkiego zbioru). PhonR może, lecz nie musi, podlegać restrykcjom. Spełniony jest natomiast warunek regularności, zarówno na poziomie znaczeniowym, jak i formalnym. Do pragmatemów należą m.in. utarte sformułowania stosowane w konkretnych sytuacjach, np. powitania, pożegnania, zwyczajowe zwroty używane w listach, znaki językowe typu *no talking please* (‘proszę nie rozmawiać’) itp.,
- **frazemy semantyczne –** warstwa znaczeniowa wyrażenia nie podlega ograniczeniom, jest natomiast nieregularna. Warstwa formalna frazemu semantycznego jest tworzona w sposób ograniczony (najczęściej dla danego SemR istnieje tylko jeden PhonR), przy czym może być regularna lub nieregularna. Ta grupa frazemów, w zależności od rodzaju nieregularności dzieli się na trzy podgrupy:
- **idiomy –** znaczenia całości wyrażenia nie da się wywieść ze znaczeń jego składników. Do tej grupy należą wyrażenia typu *[to] shoot the breeze* (‘gawędzić’) czy *[to] spill the beans* (‘zdradzić sekret’),
- **kolokacje** – jeden z członów wyrażenia ma regularne znaczenie, natomiast drugi w pewien sposób zależy od pierwszego. Przykładami kolokacji są np. *strong coffee* (‘mocna kawa’ – nie da się powiedzieć ***powerful coffee*** *‘silna kawa’), *[to] launch an attack* (‘rozpocząć atak’). Kolokacje, stanowiące według Mielczuka dominującą część frazemów, mogą być reprezentowane za pomocą konstruktów z pogranicza matematyki i językoznawstwa, zwanych przez Mielczuka funkcjami leksykalnymi. Są to funkcje przypisujące danej jednostce leksykalnej odpowiedni zbiór innych jednostek, zależny od rodzaju funkcji. Dla przykładu, funkcja **Magn** (oznaczająca intensyfikację) może mieć postać: **Magn** (*shave*) = {*close, clean*}, co jest tożsame ze stwierdzeniem, że *close shave* i *clean shave* to kolokacje,
- **quasi-idiomy –** wyrażenia, których znaczenie wywodzi się ze znaczeń składników, natomiast zawiera pewien dodatkowy element – np. *[to] start a family* (‘założyć rodzinę’) lub *bacon and eggs* (‘jajka na boczku’).

1.2.2. Koncepcja „stanfordzka”

W artykule Sag i in. (2002) przedstawiona jest koncepcja wyrażeń wielowyrazowych, którą można określić jako „kanoniczną” w kręgu NLP, nazwana tu „stanfordzką” ze względu na afiliację większości współautorów i znaczenie koncepcji dla projektu dotyczącego wyrażeń wielowyrazowych rozwijanego na Uniwersytecie Stanforda. Wyrażenia wielowyrazowe zdefiniowane są tu jako „swoiste interpretacje, które przekraczają granice wyrazów (lub spacji)” (Sag i in. 2002, 2)[11]. Zwraca uwagę duży stopień ogólności definicji, charakterystyczny dla lingwistyki komputerowej (szersze omówienie tej kwestii

[11] W oryginale: “idiosyncratic interpretations that cross word boundaries (or spaces)”.

znajduje się w podpunkcie 2.1). Oprócz popularnej wśród badaczy definicji przedstawiony jest też następujący podział wyrażeń wielowyrazowych:



Związki wielowyrazowe mogą być podzielone na dwie zasadnicze grupy: frazy skonwencjonalizowane (w oryg. *institutionalized phrases*) i zleksykalizowane. Podział ten wprowadził L. Bauer (1983), definiując wyrażenie jako skonwencjonalizowane, gdy podlega ono ograniczeniom semantycznym (tj. wyrażenie teoretycznie wieloznaczne na mocy konwencji językowej może być interpretowane tylko w jeden sposób, np. *maszyna do pisania*), a jako zleksykalizowane, gdy jego znaczenie lub forma są niemożliwe do uzyskania za pomocą konwencjonalnych reguł języka[12]. Frazy skonwencjonalizowane są regularne znaczeniowo, ale występują w danym kontekście z relatywnie wysoką frekwencją – idea ta zatem wykazuje dużą zbieżność z pojęciem kolokacji. Frazy zleksykalizowane natomiast charakteryzuje idiomatyczność różnego typu i można je podzielić na trzy klasy w zależności od stopnia, w jakim poddają się modyfikacjom morfologicznym i składniowym:

a) **wyrażenia utrwalone (fixed expressions)** – występujące tylko w jednej postaci, niemożliwe do modyfikacji morfosyntaktycznej, niedopuszczające przestawiania członów ani wstawiania dodatkowych członów (np. *in short* – ‘w skrócie’, *ad hoc*),
b) **wyrażenia częściowo utrwalone (semi-fixed expressions)** – mogą być modyfikowane w niewielkim stopniu, np. pod względem fleksji, natomiast ich szyk i skład leksykalny pozostają niezmienne. W obrębie tej grupy wyróżniane są typy, wśród których można wymienić:
   - **nierozkładalne idiomy (non-decomposable idioms)** – wyrażenia idiomatyczne, których części nie da się przyporządkować jednoznacznie do poszczególnych elementów znaczenia idiomu. Takim idiomem jest np. *kick the bucket* (‘umrzeć’, dosł. ‘kopnąć wiadro’, jego polskim odpowiednikiem może być frazeologizm *kopnąć w kalendarz*) w przeciwieństwie do wyrażenia *spill the beans* (‘zdradzić sekret’, dosł. ‘rozsypać fasolę’ – gdzie można wskazać odpowiedniość pomiędzy *rozsypać* i *zdradzić* oraz *fasolę* i *sekret*),
   - **złożenia rzeczownikowe (compound nominals)** – w tych połączeniach ośrodkiem wyrażenia jest rzeczownik, któremu towarzyszyć może przymiotnik (*personal computer* – ‘komputer osobisty’), wyrażenie przyimkowe (*part of speech* – ‘część mowy’) lub rzeczownik (*motion scanner* – ‘wykrywacz ruchu’). Ten ostatni przypadek, nazywany rzeczownikiem złożonym (noun compound) jest w języku angielskim szczególnie częsty; warto zauważyć, że w języku polskim większość konstrukcji tego typu realizuje się za pomocą połączenia rzeczownika i przymiotnika (por. *golf club* – ‘klub golfowy’, *police car* – ‘samochód policyjny’ itp.),

[12] Wydaje się, że konwencjonalizacja odpowiada pojęciu ograniczoności u Mielczuka, a leksykalizacja – nieregularności.

- **nazwy własne (proper names)**,

c) **wyrażenia zmienne składniowo (syntactically flexible expressions)** – związki wyrazowe, które podlegają szerszym modyfikacjom, np. dopuszczają rozdzielanie członów innym wyrazem, zmianę szyku, pasywizację itp. Wśród typów należących do tej kategorii autorzy wymieniają rozkładalne idiomy i wyrażenia werbalne (w tym konstrukcje charakterystyczne głównie dla języka angielskiego – phrasal verbs, light verbs); przykłady mogą stanowić tu także wyrażenia typu *sweep under the rug* (‘zamiatać pod dywan’) czy *look up* (‘sprawdzać, szukać informacji’).

1.2.3. Koncepcja Moon

R. Moon (1998) opisuje wyrażenia wielosegmentowe, posługując się narzędziami z zakresu lingwistyki korpusowej. Jednostki takie nazywa ustalonymi wyrażeniami i idiomami, tak zwanymi FEI (Fixed Expressions and Idioms). Za podstawowe cechy charakterystyczne uznaje konwencjonalizację (w praktyce mierzoną frekwencją wyrażenia w korpusie), stałość leksykogramatyczną i niekompozycyjność (asumaryczność znaczeniową). Jako kryteria drugorzędne wskazuje ortograficzną wielowyrazowość, spójność składniową (tworzenie niezależnych grup syntaktycznych) i specyfikę fonologiczną.

Moon dzieli FEI na grupy w zależności od typu nieregularności, jaki wykazują:

a) **nietypowe kolokacje (anomalous collocations)** – nieregularne pod względem leksykogramatycznym, dzielą się na cztery podgrupy:
   - **kolokacje niegramatyczne (ill-formed collocations)** – wyrażenia, które przekraczają reguły gramatyczne języka, np. *by and large* (‘ogólnie mówiąc’), *of course* (‘oczywiście’),
   - **kolokacje „żurawinowe” (cranberry collocations)** – zawierające morfemy lub wyrazy izolowane, występujące tylko w określonym połączeniu, niefunkcjonujące jako samodzielne wyrazy. Przykładami mogą być wyrażenia *to and fro* (‘tam i z powrotem’), *on behalf of* (‘w imieniu’). Nazwa pochodzi od wyrazu *cranberry* (‘żurawina’), jedynego wyrazu w języku angielskim, w którym występuje morfem *cran-* (zob. Hockett 1958, 127),
   - **kolokacje defektywne (defective collocations)** – zbliżone do poprzedniej grupy, obejmują wyrażenia, w których jeden z członów występuje w unikalnym znaczeniu, choć w innych kontekstach ma inne znaczenie (znaczenia): np. *beg the question* (‘przesądzać sprawę’), *in time* (‘punktualnie’),
   - **kolokacje frazeologiczne** – są to wyrażenia w zasadzie regularne, tworzące konstrukcje składniowe, których elementy mogą być dobierane tylko z pewnego mocno ograniczonego zbioru (np. *in action* – ‘w akcji’ / *into action* – ‘do działania’ / *out of action* – ‘nie działać’, *to a ... degree* / *to a ... extent* – ‘w … stopniu’),

b) **formuły (formulae)** – wyrażenia, które podlegają ograniczeniom natury pragmatycznej – są uzależnione od kontekstu sytuacyjnego, dzielą się na powiedzenia (skrzydlate słowa, cytaty itp.), przysłowia i porównania (skonwencjonalizowane, np. *as good as gold* – ‘wspaniały’, dosł. ‘dobry jak złoto’, *live like a king* – ‘żyć jak król’ itp.),

c) **metafory (metaphors)** – wyrażenia niekompozycyjne, nieregularne znaczeniowo:
   - **przejrzyste (transparent)** – metafory są łatwe do zdekodowania, wyraźnie motywowane, jak np. *behind someone’s back* – ‘za czyimiś plecami’, *breathe life into something* – ‘tchnąć w coś życie’ itp.,

- **częściowo przejrzyste (semi-transparent)** – wymagające pewnego stopnia wiedzy specjalistycznej, bez której są nieczytelne: np. *the pecking order* (‘hierarchia’, dosł. ‘porządek dziobania’, *throw in the towel* – (‘poddawać się, rezygnować’, dosł. ‘rzucić ręcznik’),
- **nieprzezroczyste (opaque)** – całkowicie niekompozycyjne, niemotywowane znaczeniowo, niemożliwe do poprawnego zinterpretowania bez znajomości historycznego źródła związku – np. *red herring* (‘fałszywy trop’, dosł. ‘czerwony śledź’).

Już na pierwszy rzut widać, że powyższe grupy nie są rozłączne – wiele wyrażeń może zaliczać się do więcej niż jednej klasy, czasem nie ma też pewności, do której klasy dane wyrażenie przyporządkować. Moon w swoich badaniach dopuszcza możliwość przypisania danego wyrażenia jednocześnie do dwóch klas.

Łatwo zauważyć też podobieństwo powyższej typologii do klasyfikacji Mielczuka (np. formuły odpowiadają pragmatemom, w obu typologiach występują kolokacje).

1.2.4. Inne definicje

Wielość prac dotyczących złożonych jednostek leksykalnych owocuje wielością definicji tego pojęcia. Poza przedstawionymi powyżej, są one w większości dosyć podobne do siebie nawzajem, różnią się przeważnie jedynie szczegółami lub sposobem ujęcia jakiegoś aspektu[13]. Wyróżnić można wśród nich dwa główne nurty definicyjne.

Pierwszy jest związany ze specyfiką wyrażeń wielowyrazowych pod względem statystycznym: podkreśla się tu częstsze niż wynikałoby to z przypadku współwystępowanie wyrazów składających się na jednostkę wieloczłonową (por. np. „Kolokacja to współwystępowanie dwóch lub więcej wyrazów w niewielkiej odległości od siebie w tekście” (Sinclair 1991, 170), „Powtarzająca się kombinacja słów, która występuje częściej niż wynikałoby to z przypadku” (Smadja 1993, 143)).

Drugi nurt podkreśla nieregularność (najczęściej semantyczną lub składniową) związku (por. „Pary wyrazów i frazy często używane w języku, w stosunku do których nie obowiązują ogólne reguły składniowe i semantyczne” (McKeown i Radev 2000, 507)). Evert (2005, 7) w swojej definicji zwraca uwagę na fakt, że taka nieregularność skutkuje koniecznością umieszczania tego typu jednostek w leksykonie. Bartsch (2004) w definicji *explicite* wymienia warunek, który milcząco zakładają niemal wszystkie inne definicje: między członami wyrażenia musi zachodzić bezpośrednia relacja składniowa. Niekiedy akcentowana jest kwestia jednostkowości wyrażeń wielowyrazowych: „sekwencja wyrazów, które zachowują się jak pojedyncza jednostka na którymś z poziomów analizy lingwistycznej” (Calzolari i in. 2002, 1934). Publikacja Krstev i in. (2010, 226) za element definicji uznaje, obok niekompozycyjności, konieczność posiadania przez wyrażenie unikalnej i stałej referencji. Todirascu i Gledhill (2008) proponują z kolei uznanie kolokacji nie za obiekty, lecz za relacje między wyrazami.

[13] Zestawienie ponad dwudziestu definicji można znaleźć w Seretan (2011b).

2. Definicja wielosegmentowej jednostki leksykalnej

2.1. Problemy z definicją i klasyfikacją

W pracach poświęconych leksykalnym jednostkom nieciągłym – zarówno z nurtu frazeologicznego, jak i lingwistyki komputerowej – uderzająco często powtarza się opinia, że zdefiniowanie tego pojęcia, określenie jego granic, a także klasyfikacja materiału w jego obrębie stanowią duży problem. Jest to fakt zauważany właściwie przez każdego, kto próbuje zmierzyć się z materią frazeologiczną (por. m.in. Kosek 2008, 22; Bogusławski i Wawrzyńczyk 1993, 11-13; Moon 1998, 19; Bartsch 2004, 13-14). W sposób niebezpośredni o tych trudnościach mogą świadczyć także liczne polemiki między badaczami. Bogusławski (1987) krytykuje znajdujące się w SJPDor hasło *robić coś na znak czego* jako błędne metodologicznie. Bańko (2001) z kolei za błąd uznaje decyzję Bogusławskiego o uznaniu związku *przyrodni brat* za zwykłe połączenie, a *przybrane dziecko* za jednostkę, zaznaczając, że w ISJP (który redagował) jest odwrotnie. Wyrażenie *sztywna okładka*, które Bogusławski podaje jako przykład jednostki języka, ilustrując własną definicję, uznaje Bańko (bazując na tej definicji) za zwykłe połączenie. W innym miejscu wskazuje też na niekonsekwentne traktowanie niektórych połączeń w SJPSzym, podważa zasadność uwzględnienia w tym słowniku i SJPDor połączeń typu *czerwone wino*, *głoska nosowa*, nazywając je „pozornymi frazeologizmami" (Bańko 2001, 160-164). Kosek (2008) zalicza do połączeń wyrażenia typu *kary koń*, które Bogusławski – z pewnymi zastrzeżeniami – uznawał za jednostki, natomiast *sztywną okładkę* klasyfikuje jako jednostkę.

Problem z rozróżnieniem między wielosegmentowymi jednostkami a zwykłymi połączeniami wyrazów sygnalizują również lingwiści komputerowi. Częstą praktyką przy automatycznej ekstrakcji wyrażeń wielowyrazowych jest uprzednia ręczna anotacja części materiału (oznaczane są wszystkie wielowyrazowe jednostki), co pozwala na ocenę wyników danego algorytmu. Anotacja taka jest najczęściej przeprowadzana równolegle przez dwóch lub więcej językoznawców. Co istotne, często zauważanym problemem jest tu niski poziom zgodności między anotatorami, czasami sięgający 30-50% (por. Pecina 2008b – dla języka czeskiego, Blaheta i Johnson 2001 – angielskiego).

Przyczyny tego stanu rzeczy są dwojakie. Po pierwsze, z powodu ogromnej różnorodności materiału frazeologicznego, założenia dotyczące zakresu i właściwości związków frazeologicznych niekoniecznie się u różnych badaczy pokrywają. Niektórzy skłonni są uznawać za właściwe frazeologizmy wyłącznie połączenia o wyraźnej nieregularności semantycznej lub gramatycznej, cechujące się przy tym obrazowością i ekspresywnością[14]. Inni z kolei (jak np. Skorupka) rozumieją frazeologię bardzo szeroko, traktując ją jako „naukę o połączeniach leksemów (i semów), a więc nie stawiają jej wyraźnych granic" (Kosek 2008, 17). Drugą, być może nawet ważniejszą, przyczyną rozbieżności jest duża niedookreśloność cech niektórych frazeologizmów, zmuszająca badaczy do subtelnych analiz i arbitralnych rozstrzygnięć: „Oto więc nasz problem główny: stałe baczenie, czy wielki dział wodny między jednostkami a złożeniami nie został aby przekroczony tak, iż jednostki umykają ku złożeniom, a złożenia infiltrują podstępnie obszar jednostek" (Bogusławski i Wawrzyńczyk 1993, 12-13). Ideałem byłaby jak najdalej posunięta ścisłość i jednoznaczność, możliwość dyskretnego rozróżnienia jednostka – swobodne połączenie:

[14] Bogusławski (1989) zauważa, że taki punkt widzenia jest bliski potocznemu rozumieniu frazeologii, ale w takim przypadku staje się ona dodatkiem do leksyki, pełniąc funkcję wyłącznie „ornamentacyjną".

„W badaniach naukowych (…) chodzi jednak o to, by poszukiwać narzędzi minimalizujących ów pas przejściowy i subiektywizm badacza, narzędzi pozwalających opisywać trudno uchwytny fenomen języka w sposób dyskretny, uporządkowany, spójny" (Kosek 2008, 54). Jednakże w obrębie każdej, zakreślonej konkretną definicją, grupy jednostek wielowyrazowych pojawia się wspomniany „pas przejściowy" – wyrażenia, które trudno sklasyfikować, a przypisanie ich do danej grupy sprowadza się ostatecznie do opartej nie wyłącznie na wiedzy, ale także na intuicji i arbitralnej opinii, decyzji badacza. Kosek jako przyczynę tego stanu rzeczy wskazuje właściwości analizy semantycznej, a także na zmiany samych wyrażeń, będące następstwem zmian języka w czasie. Przykładowo wyrażenie *z prądem* (w znaczeniu 'z dodatkiem alkoholu'), początkowo spotykane tylko w postaci *herbata z prądem*, *kawa z prądem*, wraz z upływem czasu rozszerza zakres łączliwości i obecnie spotykane są wyrażenia *sok z prądem*, *woda z prądem* itp. Kosek interpretuje to jako przejaw diachronicznej defrazeologizacji: pierwotnie *herbata / kawa z prądem* były samodzielnymi jednostkami, ale później, wraz z poszerzaniem się zakresu wyrażenia, utraciły ten status (por. Kosek 2008, 54 i nast.).

Dążenie do ścisłości i prostoty skłania część badaczy do jak najbardziej syntetycznego definiowania jednostek leksykalnych. Chlebda (1991, 9-10) wskazuje na tendencję do mnożenia rozmaitych kryteriów, jakie musi spełniać wyrażenie, by zostać uznane za frazeologizm, co w rezultacie powoduje silne i nieuzasadnione ograniczanie zbioru fenomenów językowych, które mogą być za frazeologizmy uważane[15]. Podobne stanowisko zajmują Bogusławski (1989) i Kosek. Postulują oni zastąpienie licznych kryteriów tylko jednym (albo przynajmniej wyróżnienie najważniejszego): u Chlebdy jest to odtwarzalność, u Bogusławskiego „elementarna składnikowość", u Kosek – w praktyce – otwartość / zamkniętość klas substytucyjnych. Kryteria frazeologiczności – te wymienione wyżej i inne – jako ważne z punktu widzenia w niniejszej książce zostaną dokładnie omówione w podpunkcie 2.2.3.

Zmierzanie do prostoty i syntetyczności definicyjnej ma z punktu widzenia naukowego istotne zalety: zwiększa precyzję i weryfikowalność. Dokładne trzymanie się tak ustalonych reguł posiada jednak również zasadnicze wady: pomijając nawet wspomnianą wyżej konieczność arbitralnych rozstrzygnięć, które – choć ograniczone – nie znikają całkowicie, powoduje, że zbiór jednostek staje się czasem zbyt szeroki (jak u Bogusławskiego, który w myśl własnych zasad uznaje za jednostki wyrażenia typu *pada deszcz*) lub dyskusyjnie zawężony (Kosek za zwykłe połączenia uznaje m.in. *pociąg pospieszny* czy wyrażenia typu *cichy wspólnik*, *ciche poparcie*).

Wyraźnie inne podejście do definiowania jednostek wielowyrazowych obecne jest w lingwistyce komputerowej: zdecydowana większość prac wykorzystuje lub tworzy definicje o dużym stopniu ogólności. Lingwistyka komputerowa – skupiając się przede wszystkim na praktycznych zastosowaniach – mniejszą wagę przywiązuje do teoretycznych rozważań i precyzyjnych rozstrzygnięć. Taka postawa nie dziwi: z jednej strony pojęcie jest

[15] Warto tu od razu zauważyć, że – czego zdaje się nie brać pod uwagę Chlebda – ograniczenie zakresu frazeologii następuje, jeśli od jednostki oczekujemy, że spełni ona każde z rozpatrywanych kryteriów. Jeśli za jednostkę uznawać będziemy wyrażenie, które spełnia jedno – lub kilka – kryteriów (czyli, stosując terminologię logiki matematycznej, bierzemy pod uwagę alternatywę a nie koniunkcję), zakres w rzeczywistości się powiększa.

intuicyjnie na tyle jasne[16], że możliwe jest zajmowanie się nim w praktyce, z drugiej strony próby dokładnego określenia jego istoty wymagają poświęcenia wiele uwagi i środków, co nie jest warunkiem *sine qua non* w przypadku zastosowań praktycznych. Przyjęcie takiego stanowiska powoduje jednak także negatywne konsekwencje: przede wszystkim wspomniane wyżej problemy ze zgodnością anotatorów prowadzą do zmniejszenia skuteczności metod oceniających efektywność algorytmu. Innym problemem jest trudność porównywania różnych technik automatycznego wyodrębniania jednostek nieciągłych ze względu na niedookreśloność zakresu materiału, do jakiego są stosowane.

2.2. Propozycja nowego podejścia do definicji wielosegmentowej jednostki leksykalnej

Niniejsza książka przedstawia własną definicję nieciągłej jednostki leksykalnej, której celem jest ściślejsze i bardziej może zgodne z rzeczywistością językową ujęcie zjawiska, pozwalające przy tym na zmniejszenie niektórych problemów towarzyszących innym definicjom. Definicja ta wykorzystana została w praktyce do stworzenia zbioru referencyjnego – zestawu jednostek wieloczłonowych występujących w korpusie IPI PAN. Zbiór ten niezbędny jest do dokładnej ewaluacji skuteczności algorytmu ekstrakcji jednostek wielosegmentowych przedstawionego w tej książce. Jego omówienie znajduje się w Rozdziale IV.

Istotą proponowanej koncepcji jest założenie, że frazeologiczność danego wyrażenia, a więc to, czy powinno być ono traktowane jako odrębna jednostka leksykalna, jest stopniowalna. Ów stopień da się określić, lokując tym samym badane wyrażenie w konkretnym miejscu skali, której bieguny to z jednej strony „swobodne połączenie składniowe", z drugiej – wyrażenie „ściśle frazeologiczne". Ocena danej sekwencji wyrazów jest wyrażona liczbowo: im większa liczba, tym wyższy jest stopień frazeologiczności. Wyrażenie oceniane jest za pomocą szeregu kryteriów – każdemu kryterium, w zależności od stopnia, w jakim wyrażenie je spełnia, oraz od ważności tego kryterium, przypisywana jest pewna punktacja. Suma punktów składa się na ostateczną ocenę wyrażenia.

Przedstawiony w punkcie 1. tego rozdziału wybór definicji i klasyfikacji związków frazeologicznych – mimo że daleki od kompletności – ilustruje ogrom różnorodności w przyjmowanych przez badaczy punktach widzenia, założeniach i interpretacjach. Patrząc na ilość podejmowanych prób opisu, można nabrać przekonania, że każda następna propozycja dotycząca definiowania zjawiska frazeologii może tylko powiększyć i tak już niemałe zamieszanie. Mimo tych zastrzeżeń niniejsza książka przedstawia nowe spojrzenie na ideę nieciągłej jednostki leksykalnej. Takie działanie podyktowane jest trzema względami, które – mam nadzieję – wystarczająco je usprawiedliwiają:

1. Żadna z istniejących definicji nie opisuje wystarczająco dokładnie omawianego fenomenu językowego w sposób, w jaki jest on rozumiany w tym miejscu. Koncepcje „tradycyjne" pomijają (lub traktują pobieżnie) niektóre istotne zjawiska, takie jak rzeczowniki wielowyrazowe czy unikalne statystyczne właściwości jednostek frazeologicznych; koncepcje lingwistyki komputerowej są zaś dostosowane głównie do języka angielskiego i najczęściej zbyt ogólne, by precyzyjnie oddawać istotę sprawy.

[16] Zauważa to już de Saussure, komentując problem wyróżniania jednostek wieloczłonowych: „Niewątpliwie osoba mówiąca nie odczuwa tych trudności; wszystko, co jest w jakimkolwiek stopniu związane ze znaczeniem, ukazuje się jej jako element konkretny, który zdoła odróżnić bezbłędnie w wypowiedzi" (Saussure de 1961, 114).

Przedstawiona propozycja wykorzystuje metody i pojęcia występujące w każdym z tych podejść, mając ambicje połączenia ich zalet.

2. Prezentowana koncepcja, nie tracąc z oczu wagi precyzyjnej definicji, daje się zarazem dobrze dopasować do celów i metod lingwistyki komputerowej, w nurt której wpisuje się ta książka.
3. Koncepcja przedstawia sposób na zmniejszenie istotnych problemów związanych z określaniem zakresu frazeologii i wyróżniania jej jednostek. Przede wszystkim rezygnuje z binarnej kategoryzacji „jednostka – luźne połączenie" na rzecz wyrażania stopnia frazeologiczności na szerokiej skali. Ogranicza w ten sposób problematyczność istnienia „szarej strefy", czyli połączeń, których jednostkowość jest niepewna i w przypadku których istnieją argumenty zarówno za, jak i przeciw umieszczeniu ich bądź w zbiorze jednostek, bądź w zbiorze swobodnych konstrukcji składniowych (zatem sytuacji, gdy każda decyzja jest w zasadzie wyborem „mniejszego zła"). Z przyjętego punktu widzenia każde połączenie wyrazów pozostających ze sobą w relacji semantyczno-składniowej charakteryzuje się pewnym stopniem „jednostkowości" (który w wielu przypadkach może być zerowy bądź znikomy). Nie jest zatem konieczny podział na kategorie jednostka i nie-jednostka. Nie znaczy to jednak, że podział jest wykluczony – nadal istnieje możliwość kategoryzacji, natomiast w łatwy sposób można modyfikować i dostosowywać granice między kategoriami (w zależności od praktycznych potrzeb). Koncepcja pozwala również na zmniejszenie stopnia subiektywności, nieodłącznie towarzyszącego decyzjom o włączeniu wyrażenia do zbioru jednostek (bądź wykluczeniu z niego). Dzięki unifikacji wielu kryteriów przedstawiona propozycja łączy zalety analitycznego i syntetycznego podejścia do definicji.
4. Koncepcja, ściśle rzecz biorąc, nie jest konkretną definicją nieciągłej jednostki, lecz metodologicznym „szkieletem", w ramach którego można zbudować własną wersję definicji. Stworzenie takiej definicji polega na ustaleniu, które kryteria brane będą pod uwagę, jaką wagę przykładać się będzie do każdego z nich i – opcjonalnie – na jakie kategorie będą dzielić się wyrażenia w zależności od stopnia frazeologiczności. W punkcie 2.2.4 przedstawiona zostanie taka właśnie konkretna definicja, przyjęta w niniejszej książce, jednak nic nie stoi na przeszkodzie, by poprzez zmianę poszczególnych jej aspektów, stworzyć własną, dostosowaną do specyficznych założeń i potrzeb.

W kolejnych podpunktach szerzej omówione zostaną poszczególne aspekty powyższej koncepcji.

2.2.1. Stopniowalność frazeologiczności

„Albo *przyrodni brat* to złożenie, albo tylko quasi-złożenie w rodzaju *starszy asystent*; tertium non datur" – to zdanie Bogusławskiego (1987, 27) dobrze ilustruje tradycyjny punkt widzenia na istotę wielosegmentowych jednostek. Według niego dane wyrażenie może przynależeć do jednej z dwóch kategorii: albo jest (w rozumieniu przyjętej przez danego badacza definicji) frazeologizmem, albo zwykłym połączeniem. W ten sposób w poczet jednostek mogą być zaliczane zarówno zestawienia wyrazów, co do których frazeologiczności nie ma wątpliwości, jak i przypadki wątpliwe, graniczne. Podobnie do grupy połączeń,

nie-jednostek, zaliczane są przypadki ewidentne i takie, co do których nie ma pewności. Przykładowo, Bogusławski do grona jednostek zalicza zarówno niewątpliwie frazeologiczną sekwencję *nie wylewać za kołnierz*, jak i wyrażenie *przybrany syn*, które dostaje ten status po skomplikowanych rozważaniach, a zatem którego „prawo" do bycia jednostką ma znacznie mniej solidne podstawy (Bogusławski 1987, 31-32). Natomiast w drugiej grupie (swobodnych połączeń) znajdują się obok siebie zarówno *żółty krawat*, jak i wspomniany wyżej *przyrodni brat*, o którym autor pisze, że jest to przypadek „autentycznie kontrowersyjny" (Bogusławski 1987, 27).

To przywiązanie do ostrego podziału binarnego, podyktowane troską o ścisłość[17] lub – znacznie częściej – automatycznym poparciem zastanego stanu rzeczy, jest charakterystyczne właściwie dla każdej koncepcji jednostek frazeologicznych. Nie istnieje – wedle mojej wiedzy – żadne opracowanie, które wykorzystywałoby do definicji fakt, że frazeologiczność różnych wyrażeń nie jest cechą dyskretną (binarną), lecz raczej ciągłą, stopniowalną, dającą się wyrazić na pewnej skali. Ciekawe jest przy tym, że wielu badaczy zauważa problem, odnotowując (i przeważnie na tym poprzestając) fakt stopniowalności frazeologizmów[18] albo pewnych ich cech (np. asumaryczności): por. m.in. McKeown i Radev (2000, 2), Blaheta i Johnson (2001), Pajdzińska (1991, 15-16). Wydaje się zatem (co potwierdza intuicja wielu badaczy), że postrzeganie frazeologiczności jako kontinuum, a nie binarnej opozycji, lepiej odpowiada istocie zjawiska i pozwala na bliższy rzeczywistości jego opis[19].

Najbliżej sformułowania takiego stanowiska był W. Chlebda (zob. Chlebda 1991, 16-18), który poddaje krytyce „postawę dychotomizującą wobec zjawisk językowych". Binarny podział narzuca się umysłowi, lecz na dłuższą metę prowadzi do braku precyzji. „Należy z całą mocą podkreślić, że dychotomizowanie obrazu świata, jakkolwiek nieodzowne we wstępnym poznawaniu rzeczywistości, znajduje się w stałym »konflikcie niedopasowania«" (Chlebda 1991, 16). Tę zasadę Chlebda stosuje w szczególności do relacji wyraz – frazeologizm, podkreślając, że „nawet najbardziej zestalonej jednostce frazeologicznej (…) przypisana jest *a priori* potencja defrazeologizacji, analityczności, rozkładalności na samodzielne wyrazy. I na odwrót: autonomiczność wyrazu jest względna wskutek przypisania mu ograniczeń intralingwalnych (np. walencyjnych) i ekstralingwalnych" (Chlebda 1991, 18).

Podejście do definicji nieciągłych jednostek leksykalnych w niniejszej książce jako fundament wykorzystuje właśnie koncepcję stopniowalności frazeologiczności. Każdemu wyrażeniu przysługuje ocena, umieszczająca je w odpowiednim miejscu kontinuum bytów językowych, w którym na jednym biegunie lokują się całkowicie regularne, tworzone *ad hoc* połączenia, a na drugim w pełni frazeologiczne wielowyrazowe jednostki.

Takie ujęcie zagadnienia przynosi kilka ważnych konsekwencji. Przede wszystkim podział na połączenia i jednostki złożone jest dużo mniej ostry: o konkretnym wyrażeniu można mówić jedynie, że jest „bardziej (lub mniej) frazeologiczne" od innego. Zmniejszona

[17] „Materia językowa jest skomplikowana i wątpliwości oczywiście nie da się uniknąć, chodzi jednak o próbę precyzyjnego jej opisania (uporządkowania) (…), o próbę niezakładającą a priori zjawisk pośrednich, tzn. w tym wypadku – swojego rodzaju »stopniowania« jednostkowości" (Kosek 2008, 46).

[18] Często przybiera to formę wskazywania na istnienie jednostek „pośrednich", leżących pomiędzy zbiorem swobodnych połączeń i związków frazeologicznych – takie podejście implikuje stopniowalność.

[19] Nie jest to oczywiście regułą, np. Kosek przyjmuje stanowisko odmienne, sprzeciwiając się idei stopniowalności: „[metoda Bogusławskiego – MW] stwarza szansę na uniknięcie (…) wyróżniania różnorakich bytów, grup pośrednich, w których ów przedmiot w końcu się rozpływa" (Kosek 2008, 46). Jednak i ona w wielu miejscach zauważa problemy z klasyfikacją binarną (por. Kosek 2008, 50-1; 54).

ostrość podziału może wydawać się problematyczna. Pojawiają się wątpliwości, czy nie wpłynie to na precyzję opisu, rodzi się też pytanie natury praktycznej: w jaki sposób zdecydować (np. przy tworzeniu słownika), co jest jednostką, a co nie jest. Dodatkowo: czy zaprzęganie matematycznego aparatu jest uzasadnione w badaniach językoznawczych? Jeśli tak, to w jaki sposób dokonywać oceny wyrażeń?

Pierwsze dwie wątpliwości łatwo rozwiać. Nadal można dzielić wyrażenia na grupy, ustalając przedziały, w jakich musi się mieścić ocena danego wyrażenia, by przypisać je do odpowiedniej grupy. Liczba tych grup nie jest z góry ustalona. Mogą być tylko dwie: powyżej pewnego progu wyrażenie uznawane jest za jednostkę, poniżej – za połączenie (odpowiada to tradycyjnemu, binarnemu podziałowi), mogą być trzy (jeśli ustalimy dwa progi): związki silnie frazeologiczne, związki z „szarej strefy" i zwykłe połączenia; może ich być dowolna inna ilość. Nawet zatem w przypadku pozostania przy tradycyjnym schemacie kategoryzacji otrzymujemy wygodne narzędzie pozwalające porównywać wyrażenia. Przyjęty tu punkt wiedzenia nie przeszkadza więc w klasyfikacji, a precyzja opisu raczej wzrośnie, niż zmaleje.

Stosowanie metod matematycznych (nawet tak nieskomplikowanych) nie należy do podstawowych narzędzi w językoznawstwie, jednak używa się ich coraz częściej. Pozwala to na wykorzystanie ich immanentnych, naukowo pożądanych cech – zwłaszcza precyzji i syntetyczności opisu. Ujęcie zjawisk językowych w matematyczne ramy ułatwia wydobycie mniej oczywistych cech bytów językowych, ich zależności i wzajemnych relacji.

Zwiększająca się popularność tych metod ma również związek z rozwojem komputerów, korpusów i – bardziej ogólnie – lingwistyki komputerowej. Istnieje także nurt analityczny lingwistyki matematycznej, wykorzystujący aparat matematyczny do analizy i opisu języków naturalnych[20], np. Lyons (1976) wykorzystuje w badaniach elementy teorii mnogości.

2.2.2. Kryteria wyróżniania jednostek wielosegmentowych

Każda definicja związku frazeologicznego operuje pewnym zestawem kryteriów, których obecność jest wymagana, by dane wyrażenie zaliczyć w poczet frazeologizmów. Takich kryteriów w literaturze przedmiotu można znaleźć wiele. Bogusławski (1989) wymienia ich dwanaście, Chlebda (1991) wspomina o „niemal dwudziestu", wymieniając jako najczęstsze jedenaście z nich, w innych pracach można znaleźć kolejne. Kryteria są również mocno zróżnicowane na rozmaitych poziomach. Po pierwsze, różni je poziom ogólności: wymóg, by związek „przekraczał granicę wyrazów" (Sag i in. 2002) czy „zachowywał się jak pojedyncza jednostka" (Calzolari i in. 2002), jest dużo mniej konkretny niż konieczność spełnienia proporcji analogicznej w rozumieniu Bogusławskiego. Po drugie, poszczególne kryteria dotyczą rozmaitych aspektów językowych: niektóre biorą pod uwagę semantykę związku, inne – składnię, jeszcze inne – kwestie pragmatyczne lub statystyczne.

Immanentną cechą jednostek złożonych, tak podstawową, że trudno uznać ją za kryterium, jest wielosegmentowość – tzn. fakt, że wyrażenia takie składają się z co najmniej dwóch członów nieciągłych, przy czym nieciągłość ta jest przez większość badaczy wyróżniana na podstawie zapisu ortograficznego (wyrazy rozdzielone spacją – zob. np. Kosek 2008, Baldwin i Kim 2010). Kryterium to bywa jednak uznawane za niewłaściwe (ortografia jest zasadniczo kwestią ustalaną arbitralnie) lub niewystarczające. Z tego względu niektórzy badacze proponują, by kryterium nieciągłości wywodzić raczej z relacji składniowych (zob.

[20] Zob. np. hasło *lingwistyka matematyczna* w *Encyklopedii językoznawstwa ogólnego.*

np. Sag i in. 2002). Wielosegmentowość może być w zasadzie utożsamiana z wielowyrazowością, jednak ze względu na brak pewności, czy części składowe jednostki złożonej można utożsamiać z wyrazami[21], wielosegmentowość wydaje się być precyzyjniejszym terminem.

2.2.2.1. Kryterium a cecha jednostki nieciągłej

Istnieje duże podobieństwo między pojęciem kryterium frazeologiczności a cechą frazeologizmu. Jednostki leksykalne, niezależnie od definicji, charakteryzują się specyficznymi własnościami. Najczęściej właśnie te własności stają się podstawą do odgraniczenia jednostek od zwykłych połączeń wyrazowych, zatem pełnią rolę kryteriów. Są jednak pewne różnice. W przypadku niektórych kryteriów dyskusyjne jest, czy mogą być one traktowane jako cechy: np. w koncepcji Mielczuka ograniczoność, o jakiej można mówić w przypadku części frazemów, jest właściwie cechą systemu językowego, a nie konkretnego wyrażenia. Podobne wątpliwości mogą pojawić się w przypadku konwencjonalizacji (stopnia utrwalenia danej jednostki w języku). Ponadto niejednokrotnie zdarza się, że autor koncepcji wymienia cechy jednostki frazeologicznej, nie przypisując im jednak roli definicyjnej, a raczej drugoplanową – cechy takie traktowane są bardziej jako pomocnicze, ułatwiające selekcję odpowiednich wyrażeń (co nie znaczy, że według innej koncepcji nie mogą stanowić one obligatoryjnych części definicji).

2.2.2.2. Dobór i łączenie kryteriów

Niekiedy, jak wspomniane zostało wyżej, definicja nieciągłej jednostki leksykalnej jest tworzona w oparciu o tylko jedno kryterium. Tak jest w przypadku niektórych prac z zakresu lingwistyki komputerowej: Cowie (1978) i Sinclair (1991) za jedyne kryterium uznają statystyczną istotność wyrażenia. Winogradow czy Choueka (1988) za warunek konieczny uważają nieregularność znaczenia. We frazematyce jest to odtwarzalność[22]. Formalnie jedno kryterium stosuje również Bogusławski, jednak w praktyce jego elementarna składnikowość łączy kilka kryteriów (m.in. asumaryczność znaczeniową, otwartość klasy substytucyjnej).

Rozwiązanie takie ma swoje zalety i wady. Przyjęcie tylko jednego kryterium powoduje, że definiowanie zbioru jednostek jest łatwiejsze (ogranicza się do oceny tylko jednej cechy). Powstały w ten sposób zbiór jest spójniejszy, a koncepcja jednostki wydaje się być bardziej elegancka (zgodnie z postulatem Chlebdy „nie mnoży bytów ponad potrzeby" – por. Chlebda 1991, 9). Z drugiej strony ewentualne błędy i subiektywność decyzji mają większe konsekwencje (decyzja podejmowana jest tylko raz i dotyczy całości wyrażenia, a nie któregoś z jego aspektów). Drugim mankamentem jest to, że powstały w ten sposób zbiór łączy wiele zjawisk językowych, których status jest różny (np. pragmatemy w rodzaju *dzień dobry* i niekompozycyjne, ekspresywne frazy typu *drzeć z kogoś pasy*). Nie są to wady dyskwalifikujące: jeśli proces wyróżniania jednostek będzie odpowiednio uważny, można zminimalizować

[21] Poglądy na ten temat są podzielone – kwestie sporne omawia m.in. Żmigrodzki (2009, 101-102). Kwestię wyrazu – w tym jednostek wielosegmentowych – z punktu widzenia lingwistyki komputerowej przedstawia Lubaszewski (2009).

[22] Czasem pojedyncze kryterium wspomagane jest przez dodatkowe warunki, które same w sobie nie stanowią kryterium: albo są zbyt ogólne i samodzielnie nie pozwalałyby na efektywne wyróżnianie jednostek (np. konieczność wchodzenia w relacje syntaktyczne), albo są ograniczeniami dotyczącymi kryterium (we frazematyce nie każdy ciąg odtwarzalny jest frazemem: musi być elementem podstawowym, nazywającym pojęcia, sądy itp. – zob. Chlebda 1991, 128).

ryzyko pomyłek, a problem niejednorodności da się rozwiązać, wprowadzając sekundarną, wewnętrzną klasyfikację. W praktyce największym problemem jest jednak znalezienie kryterium, które byłoby tu odpowiednie. Każde z wyżej wymienionych nastręcza problemów, mocno ograniczających skuteczność definicji. Wykorzystanie wyłącznie kryterium statystycznego skutkuje zbiorem zbyt ogólnym: obok rzeczywistych jednostek pojawiają się tam popularne połączenia w rodzaju *koniec marca* itp. Kryterium nieregularności semantycznej daje, odwrotnie, zbiór dosyć ograniczony, w którym nie ma miejsca np. na wyrażenia regularne znaczeniowo, ale nie leksykalnie, typu *po ciemku*. Z kolei dzięki odtwarzalności powstaje zbiór bardzo pojemny – co, w zależności od przyjętych założeń może stanowić zarówno wadę, jak i zaletę (zakres pojęcia nieciągłej jednostki przyjęty w tej książce i we frazematyce jest – jak się wydaje – w dużym stopniu zbliżony). Z odtwarzalnością są jednak związane istotne problemy metodologiczne (por. punkt 2.2.3.4), które stawiają pod znakiem zapytania akceptowalność tej definicji.

Inną strategią przy tworzeniu definicji jest łączenie poszczególnych kryteriów. Może ono przebiegać na różne sposoby. Przykładowo, Krstev i in. (2010) postuluje, by wyrażenie uznane za jednostkę spełniało wszystkie z wymaganych kryteriów (graficzna wielowyrazowość, niekompozycyjność, unikalna i stała referencja), u Mielczuka wystarczy natomiast, by wyrażenie spełniało warunek nieregularności albo ograniczoności. W przypadku wielu prac nie wiadomo właściwie, jaka kombinacja kryteriów jest konieczna, by wyrażenie spełniało warunki definicji – prace takie poprzestają tylko na wymienieniu kryteriów branych pod uwagę. Kontekst najczęściej sugeruje, że wyrażenie, które spełnia wystarczająco wiele spośród warunków, może być uznane za jednostkę, lecz nie jest sprecyzowane, co znaczy „wystarczająco wiele" – zob. np. Manning, Schütze (1999, 184)[23]. Mimo to łączenie kryteriów najczęściej pozwala na dokładniejsze określenie pojęcia jednostkowości i ułatwia wewnętrzną klasyfikację jednostek. Mankamentem takiego rozwiązania jest konieczność włożenia większego wysiłku w selekcję wyrażeń – każdą frazę trzeba ocenić, biorąc pod uwagę kilka aspektów.

Bogusławski (1989) pokazuje, jak wybór odpowiedniego zestawu kryteriów może określać zakres i sposób rozumienia frazeologii. Na tę zależność zwraca również uwagę Kosek (2008, 17), zauważając, że obszar zjawisk językowych uznawanych za frazeologiczne może się bardzo istotnie różnić: od rozumienia bardzo wąskiego, biorącego pod uwagę tylko wyrażenia asumaryczne znaczeniowo, charakteryzujące się przy tym metaforycznością i ekspresywnością, po bardzo liberalne, traktujące frazeologię raczej jako naukę o łączeniu wyrazów.

2.2.2.3. Wykorzystanie kryteriów w niniejszej książce

Koncepcja proponowana w tej książce opiera się na łączeniu różnorodnych kryteriów, jednak zasady ich współdziałania są określone konkretniej niż w większości przytoczonych prac. Przede wszystkim należy zauważyć, że ocena, czy dane wyrażenie spełnia jakieś kryterium, bardzo rzadko daje się sprowadzić do prostej decyzji: tak albo nie. W większości przypadków można mówić raczej o spełnianiu w większym lub mniejszym stopniu. Dla przykładu: *biały kruk* cechuje się zdecydowanie większą asumarycznością znaczenia niż *wieczór kawalerski*. Pierwsze wyrażenie jest metaforą, jego znaczenia w żaden sposób nie da się wyprowadzić z wyrazów składowych, w przypadku drugiego można mówić o pewnej

[23] Konfuzję pogłębia fakt, że często nie jest jasne, na ile twórcy danej koncepcji traktują kryteria jako część definicji, a na ile jako opis własności.

wartości dodanej: w końcu nie każdy wieczór, w którym biorą udział kawalerowie, to *wieczór kawalerski*. Jakkolwiek widać wyraźnie różnicę pod względem asumaryczności, nie da się powiedzieć, że jedno z nich spełnia kryterium, a drugie nie – oba je spełniają: *biały kruk* silnie, *wieczór kawalerski* w niewielkim stopniu. Podobnie jest w przypadku innych kryteriów.

Drugą rzeczą, na którą trzeba zwrócić uwagę, jest fakt, że niektóre kryteria są ważniejsze od innych. Wspomniana asumaryczność semantyczna nie bez powodu pełni pierwszoplanową rolę w zdecydowanej większości definicji: przesunięcie znaczenia to bardzo silny sygnał, że daną frazę należy traktować jako jednostkę. Z drugiej strony wysoka frekwencja danego zwrotu może sugerować jednostkowość, lecz nie jest to regułą.

Metoda klasyfikacji w oparciu o konkretne kryteria, proponowana przez niniejszą książkę, bazuje na tych spostrzeżeniach. Jej zasady to:

1) stopień, w jakim poszczególne kryterium jest spełniane przez badane wyrażenie, jest oceniany według trójstopniowej skali: **silny – zauważalny – zerowy**. Każdemu z tych stopni przypisywana jest wartość punktowa, odpowiednio 2, 1 i 0. Przykładowo *biały kruk* ze względu na asumaryczność znaczenia oceniany jest na 2 punkty, *wieczór kawalerski* na 1 a *piękny kwiat* na 0 punktów,
2) poszczególne kryteria mają przypisany sobie **współczynnik istotności**, przez który mnożona jest ocena z punktu 1. Asumaryczność znaczenia ma współczynnik istotności ustalony na 4, zatem *biały kruk* otrzymuje za to kryterium ocenę 8 (2 * 4), *wieczór kawalerski* 4 (1 * 4), a *piękny kwiat* 0 (0 * 4).

Suma ocen uzyskanych za poszczególne kryteria składa się na ostateczną ocenę wyrażenia.

Jak wskazano wyżej, opisana procedura nie jest tożsama z definicją. Aby uzyskać konkretną definicję, należy określić zbiór kryteriów branych pod uwagę, istotność każdego z nich, uściślić sposób oceny (tzn. warunki, jakie musi spełniać wyrażenie, by dane kryterium było uznawane za spełnione w sposób zauważalny lub silny), a także próg liczbowy (lub progi) dzielące wyrażenia na dwie lub więcej klas (np. klasa jednostek i klasa swobodnych połączeń). W zależności od przyjętych ustaleń mogą w ten sposób powstać definicje o skrajnie różnych zakresach. Przedstawioną propozycję należy zatem rozumieć jako koncepcję metodologiczną, a nie definicję samą w sobie (ta zostanie opisana w punkcie 2.2.4).

W sekcji poniżej omawiam większość kryteriów, jakie pojawiają się w literaturze, dyskutując przy tym ich zalety i wady, a także miejsce, jakie zajmują one w definicji nieciągłej jednostki leksykalnej przyjętej w tej książce. Niektóre z tych kryteriów są uniwersalne i mogą występować we wszystkich rodzajach jednostek, inne są wyznacznikami pewnych klas jednostek, ich występowanie świadczy najczęściej o konieczności przyporządkowania wyrażenia do określonej grupy. Na podstawie tych kryteriów opracowana została typologia jednostek wielosegmentowych, opisywana w sekcji 3.1.

2.2.3. Omówienie kryteriów jednostkowości

2.2.3.1. Nieregularność

Wielowyrazowe jednostki leksykalne ze swojej natury stoją w opozycji do regularnych połączeń[24]. Ta prymarna cecha powoduje, że nieregularność obecna na którymś z poziomów

[24] Warto tu zauważyć, że pojęcie „regularności" nie jest tożsame z „dowolnością". W języku właściwie każde połączenie jednostek (morfemów, wyrazów, zdań) podlega pewnym ograniczeniom (por. Kosek 2008, 48-9).

językowej charakterystyki jest bardzo silnym wyznacznikiem jednostkowości wyrażenia. Często nieregularność zawęża się do nieregularności semantycznej (asumaryczności znaczeniowej, niekompozycyjności), ale można wskazać również inne: nieregularność leksykalną i składniową. Baldwin i Kim (2010) wymieniają te trzy typy, uznając je za przejawy idiomatyczności, dodając dwa kolejne rodzaje: idiomatyczność statystyczną i pragmatyczną[25]. Każdy z tych typów może występować niezależnie od pozostałych i służyć jako osobne kryterium sygnalizujące frazeologiczność, choć w praktyce często wyrażenie łączy w sobie kilka typów nieregularności.

- **Nieregularność semantyczna**
  Wyrażenie nieregularne semantycznie to takie, którego znaczenie nie stanowi sumy znaczeń jego składników. Nieregularność semantyczna bywa nazywana asumarycznością znaczenia, (semantyczną) idiomatycznością, globalnością znaczeniową, niekompozycyjnością (ang. *non-compositionality*); Lewicki (1982, 6) określa ją jako „brak symetrii między planem treści a planem wyrażenia nie dający się opisać za pomocą reguł kategorialnych". W pracach anglojęzycznych czasem przez niekompozycyjność rozumie się dowolną nieregularność, dlatego w sytuacjach, które mogłyby prowadzić do konfuzji, stosuje się uściślenie: niekompozycyjność semantyczna (ang. *semantic non-compositionality*). Przez wielu badaczy kryterium to uznawane jest za najważniejsze kryterium frazeologiczności (zob. m.in. Pajdzińska 1988, 8; Manning i Schütze 1991, 184). Fakt, że ze znaczeń składników użytkownik języka nie jest w stanie uzyskać znaczenia całości wiąże się z dwoma istotnymi aspektami:
  a) wyrażenie takie musi być zapamiętywane i przywoływane w całości, a nie tworzone *ad hoc*, co implikuje konieczność traktowania go jako jednostki,
  b) wyrażenia tego typu mają szczególną wagę przy tłumaczeniu i nauce języka.

  Podstawą asumaryczności znaczeniowej jest najczęściej metafora (np. *czarna rozpacz*), metonimia (np. *głowa rodu*) lub hiperbola (np. *niewart splunięcia*), będące źródłem przesunięcia znaczeniowego. Przenośność wyrażenia nie jest jednak warunkiem *sine qua non*. Znaczenie wyrażenia typu *analiza techniczna* – 'w ekonomii: zbiór technik służących do przewidywania kursów giełdowych' jest asumaryczne, choć znaczenia obu członów są regularne. Zatem o ile metaforyczność automatycznie pociąga za sobą asumaryczność znaczenia, o tyle odwrotna relacja nie zawsze zachodzi. Z tego powodu metaforyczność jest w tej książce traktowana jako część kryterium nieregularności semantycznej.

  Należy zauważyć, że ustalenie, czy dane wyrażenie jest asumaryczne znaczeniowo, nie zawsze jest proste. O ile *biały kruk* jest w oczywisty sposób nieregularny, a *biały śnieg* jest w pełni regularny, o tyle w przypadku *białego sera* nasuwają się wątpliwości. Z jednej strony rzeczywiście mowa jest o serze, który jest biały, z drugiej – jest w tym wyrażeniu dodatkowa wartość semantyczna. Podstawowym kłopotem staje się fakt, że interpretacja znaczenia danego wyrazu jako metaforycznego bądź konwencjonalnego – nawet jeśli rzadko używanego – sama w sobie czasem bywa problematyczna. Teoretycznie każde, nawet jednostkowe, użycie wyrazu w ja-

[25] Jako że specyfika statystyczna i pragmatyczna nie jest konsekwencją nieregularności, a raczej – używając terminologii Mielczuka – ograniczoności wyrażenia, są one w niniejszej pracy potraktowane jako osobne kategorie.

kimś znaczeniu można uznać za regularne i umieścić je w leksykonie (a więc dla *białego kruka* byłyby to leksemy *biały* w znaczeniu ʻbardzo rzadkiʼ i *kruk* – ʻksiążkaʼ z adnotacją, że znaczenia te realizują się wyłącznie w połączeniu), jednak takie rozwiązanie było wielokrotnie odrzucane jako absurdalne i nieopłacalne (por. Kosek 2008, 51 i cytowane tam prace). Podejście takie – jako prowadzące do potencjalnej „nadprodukcji" (*overgeneration*) fraz – krytykuje też Sag i in. (2002). Należy zatem wyznaczyć granicę między metaforą a regularnym znaczeniem, lecz kwestia, w którym miejscu powinna ona przebiegać, pozostaje otwarta. Problem powiększa tendencja do przesuwania się tej granicy wraz z naturalnymi zmianami w obrębie języka (por. Kosek 2008, 54 i nast.). Ilustrować to może forma *salomonowy*, pierwotnie występująca wyłącznie w wyrażeniu *salomonowy wyrok*. Obecnie, jak pokazują badania (por. Mosiołek-Kłosińska 2002, cyt. za Kosek 2008), funkcjonują frazy łączące *salomonowy* m.in. z leksemami *decyzja*, *rozwiązanie*, *wybór*, *pomysł* czy *sposób*. W ten sposób pojawia się nowy leksem, którego metaforyczność (a na pewno motywacja) zaczyna się rozmywać. Z utratą frazeologiczności związana jest idea otwartości klas substytucyjnych występująca w koncepcji Bogusławskiego, która – jako osobne kryterium – omawiana jest dalej.

Bardzo kłopotliwe przy analizie nieregularności semantycznej są wyrazy niesamodzielne, takie jak przyimki czy spójniki, tworzące wyrażenia typu *na pozór*, *a mianowicie* itp. Znaczenie takich wyrazów jest często niezwykle trudne do uchwycenia. Zagadnienie to w szeroki sposób opisuje Przybylska (2002). Dla przykładu, *Słownik języka polskiego PWN* podaje dziesięć różnych sposobów użycia przyimka *na* w zależności od sytuacji, do których się odnosi i – w domyśle – różniących się znaczeniem. Nie odnotowuje przy tym konstrukcji typu *na znak*, *na zlecenie*, *na myśl*, w których znaczenie tego przyimka można by określić jako ʻwywołane przez Xʼ, gdzie X oznacza odpowiedni rzeczownik[26]. W związku z trudnościami w określeniu znaczenia wyrazów niesamodzielnych, tym większy kłopot stanowi próba analizy nieregularności semantycznej konstrukcji z ich udziałem.

Opisywane wątpliwości nie oznaczają jednak, że niemożliwe jest wskazanie asumaryczności znaczeniowej w niektórych przypadkach. Jedną z możliwości są na przykład konstrukcje typu *na barana*, *na wariata*. W tym przypadku znaczenie przyimka można określić jako ʻwskazujący na metodę przywodzącą na myśl X ze względu na semantyczne powiązanie między nią a Xʼ. Znaczenie *na* jest zatem ustalone, nie jest określony jednak aspekt znaczenia X, który w danym wypadku należy wziąć pod uwagę[27] – i z tego niedookreślenia wynika nieregularność. Interesujące jest przy tym, że mamy tu do czynienia ze zjawiskiem, które można by określić jako „strukturę frazeologiczną", umożliwiającą tworzenie serii wyrażeń nieregularnych według odpowiedniego schematu. Niektóre z tych połączeń – jak *na jeża* – są ustalone w języku (skonwencjonalizowane), inne mogą być tworzone *ad hoc*, np. *na żołnierza*,

[26] W świetle powyższego przykładu nie dziwi komentarz Przybylskiej na temat reprezentacji przyimków w słownikach ogólnych języka polskiego, „w których sposób opracowania haseł przyimkowych zdradza zagubienie i bezradność autorów połączone z brakiem konsekwencji widocznym w odmiennym uporządkowaniu leksykograficznym haseł poświęconych różnym przyimkom" (Przybylska 2002, 62).

[27] W przypadku frazy *na barana* tym aspektem będzie zwyczaj przenoszenia owiec na ramionach, w wyrażeniu *na wariata* nieprzewidywalne, lekkomyślne zachowanie osób niepoczytalnych.

na prezent – jednak w takim przypadku muszą być objaśniane, gdyż ich znaczenia nie można się domyślić.

Nieregularność semantyczna może być pełna – dotyczyć wszystkich członów wyrażenia (np. *zbijać bąki*, *czarna polewka*), lub częściowa – gdy nieregularność znaczenia dotyczy tylko części składników wyrażenia (np. *niedzielny kierowca*, *kłamać jak z nut*). Może się też różnić stopniem motywacji: w niektórych wyrażeniach motywacja jest wyłącznie historyczna (np. *smalić cholewki*), a więc z punktu widzenia synchronicznego nie istnieje, w innych jest tylko częściowo przejrzysta (np. *prawa ręka*), wreszcie zdarza się, że jest doskonale czytelna (*łysy jak kolano*).

Z pojęciem asumaryczności znaczeniowej wiąże się też przywołana wcześniej (1.2.2) koncepcja rozkładalności (ang. *decomposability*). Znaczenia niektórych wyrażeń, mimo ich metaforyczności, można rozłożyć na części składowe – przykładowo, w powiedzeniu *tonący brzytwy się chwyta* członowi *tonący* można przyporządkować znaczenie ‘człowiek w opałach’, a reszcie frazy znaczenie ‘ratuje się w każdy możliwy sposób’. Podobne działanie nie zawsze jest możliwe, np. w wyrażeniu *rozdzierać szaty*[28].

- **Nieregularność leksykalna**
  Nieregularność leksykalna to obecność w wyrażeniu form tekstowych, które nie występują nigdzie poza tą frazą, np. *pantałyku* we frazie *zbić z pantałyku* czy *kozery* w *nie bez kozery*. Takie formy są w pewnym stopniu wybrakowane, przypisanie im cech morfosyntaktycznych, kategorii gramatycznych, a czasem nawet wartości semantycznych stwarza problemy lub jest wręcz niemożliwe – z tego względu ich status jako wyrazu jest wysoce wątpliwy. Formie *pantałyku*, której pierwotne znaczenie nie jest znane, można by przypisać cechy gramatyczne poprzez analogię do leksemów, które wchodzą w syntagmatyczną relację z sekwencją *zbić z* (jak np. *trop*), jednak byłyby to formy wyłącznie potencjalne. Nie ma też sensu włączanie takich form do słownika (por. Sag i in. 2002).

  Kosek (2001) opisuje przypadki wyrażeń z przyimkiem *na*, powstałych z regularnych konstrukcji, ale poddanych procesowi leksykalizacji. Wyrażenia takie (np. *na nice*, *na oklep*), w których część rzeczownikowa bądź przysłówkowa występuje wyłącznie (lub niemal wyłącznie) w połączeniu z tym przyimkiem, można traktować jako nieregularne leksykalnie.
- **Nieregularność składniowa**
  Podobnie jak w przypadku nieregularności leksykalnej, idiomatyczność składniowa oznacza pojawienie się konstrukcji syntaktycznej wykraczającej poza ramy gramatyczne języka. Przykładami mogą być wyrażenia *na koń!* czy *wszem i wobec*.

Nieregularność dowolnego typu jest najczęściej bardzo silnym wykładnikiem jednostkowości. W praktyce nieregularność leksykalna i składniowa zdarzają się jednak dosyć rzadko, w przeciwieństwie do asumaryczności znaczeniowej.

W niniejszej książce nieregularność jest uznawana za najważniejsze kryterium, jej współczynnik istotności wynosi 4. Wyrażenie klasyfikowane jest jako silnie nieregularne (a więc otrzymuje 2 punkty, mnożone następnie przez współczynnik istotności), jeśli występuje nieregularność leksykalna lub składniowa. W przypadku asumaryczności znacze-

[28] Temat rozkładalności (choć nie nazywanej w ten sposób) podejmuje D. Buttler (1982, 55-56).

niowej, klasyfikowane jest jako silnie nieregularne, gdy przesunięcie znaczenia jest konsekwencją metafory. Jeśli asumaryczność ogranicza się do wartości dodanej (jak w *pociągu pospiesznym*) traktowane jest jako zauważalnie nieregularne (uzyskuje zatem 1 punkt, mnożony następnie przez 4).

2.2.3.2. Swoistość statystyczna (kolokacyjność)

Wielowyrazowe jednostki leksykalne mają tendencję do pojawiania się w tekście lub mowie częściej niż inne połączenia wyrazów. Ta statystyczna zależność była wielokrotnie odnotowywana i badana (zob. np. Choueka 1988, Smadja 1993, Pecina 2009). Jest konsekwencją utrwalenia danej jednostki w języku – z tego punktu widzenia może być miarą konwencjonalizacji (zob. podpunkt 2.2.3.7) i – ogólniej – rozpowszechnienia społecznego (zob. Chlebda 1991, 9). Warto zauważyć, że nie jest to cecha *stricte* lingwistyczna – oddaje bardziej zwyczaje językowe w określonej grupie społecznej.

Taka swoistość statystyczna – czyli kolokacyjność – jest na ogół dosyć mocnym wskaźnikiem frazeologiczności, jednak ma kilka wad. Po pierwsze, żeby ją w miarę obiektywnie badać, konieczny jest korpus tekstów – powinien być jak największy, zbalansowany i reprezentatywny – niedostatek na którymkolwiek polu obniża lub przekreśla wiarygodność oceny. Nawet jednak w przypadku korpusu spełniającego te warunki można mówić tylko o przybliżeniu: żaden korpus nie jest w stanie dokładnie odzwierciedlić rzeczywistości językowej. Druga kwestia wynika z poprzedniej: wiele jednostek frazeologicznych nie pojawi się wcale w korpusie lub występować będą w ilości statystycznie mało znaczącej. Jest to znany problem w lingwistyce komputerowej (zob. np. Manning i Schütze 1999, Evert 2005). Co więcej, badania pokazują, że wyrażenia silnie niekompozycyjne, a więc z dużo większą pewnością zaliczające się do grupy frazeologizmów, są rzadziej używane niż inne typowe zbitki wyrazowe (por. Baldwin i Kim 2010, 5). Kolejna niepożądana – z punktu widzenia wyodrębniania frazeologizmów – cecha kolokacyjności to wysoka frekwencja niektórych typowych konstrukcji, które niewątpliwie nie są jednostkami, jak np. *koniec marca*, *do lat*, *na siebie* itp. O ile połączenia, w których skład wchodzą wyrazy funkcyjne, nie stanowią problemu, ponieważ łatwo je odfiltrować, o tyle pozostałe sekwencje niełatwo odróżnić od rzeczywistych jednostek, zwłaszcza jeśli kryterium statystyczne jest jedynym lub głównym wskaźnikiem frazeologiczności.

Kolokacyjność jest niezwykle istotna z punktu widzenia lingwistyki komputerowej. Istnieje szereg technik wykorzystujących tę cechę do automatycznego wyodrębniania nieciągłych jednostek (zob. Rozdział I). Różne miary asocjacji (opisane szerzej w Rozdziale III) mogą być wykorzystane do oceny stopnia kolokacyjności. Wyrażone są one liczbowo, dlatego z powodzeniem można je dostosować do proponowanej w tej książce koncepcji, określając dwa progi: powyżej większego z nich wyrażenia uznawane są za silnie kolokacyjne, wyrażenia mieszczące się pomiędzy progami – za zauważalnie kolokacyjne, a te poniżej obu progów – za niekolokacyjne.

W niniejszej książce kolokacyjność ma status specyficzny. Choć jest ona uznawana za istotne kryterium, w niniejszej książce nie była wykorzystywana przy ocenie wyrażeń w zbiorze porównawczym (zob. Rozdział IV, podpunkt 3.3.1), służącym ocenie skuteczności prezentowanego w pracy algorytmu do wyodrębniania wielosegmentowych jednostek leksykalnych. Powodem jest to, że algorytm ten opiera się przede wszystkim właśnie na analizie kolokacyjności. Ewaluacja, która polega na porównaniu listy potencjalnych

jednostek wyodrębnionych przez algorytm z listą jednostek tworzącą zbiór referencyjny, zakłada, że listy te są niezależne. Gdyby więc konstrukcję tego zbioru opierać (choćby w części) na kolokacyjności, prowadziłoby to do błędu metodologicznego, a wyniki ewaluacji byłyby mniej wiarygodne.

2.2.3.3. Swoistość pragmatyczna

Użycie niektórych sformułowań jest powiązane z odpowiednim kontekstem sytuacyjnym. Na przykład *dzień dobry* jest wyrażeniem wykorzystywanym przy powitaniach, a *nie ma za co* jest typową odpowiedzią na czyjeś podziękowanie. Tego typu uwarunkowania sprawiają, że takie wyrażenia mają w języku specjalny status, a zatem nie są zwykłymi połączeniami, choć w wielu wypadkach są w pełni regularne pod względem znaczeniowym. Ciekawą rzeczą jest fakt, że to właśnie określony kontekst wymusza ich specyficzną interpretację, a gdy takiego kontekstu brak, stają się często zwykłymi połączeniami (por. *Niedziela to dzień dobry na spacery*).

Kryterium to nie jest stosowane powszechnie (w każdym razie *explicite*), ale przez wielu badaczy uznawane jest za ważne. W koncepcji Mielczuka pojawiają się pragmatemy, u Moon podobną rolę pełnią formuły, Bartsch jednostki wielowyrazowe definiuje m.in. jako „pragmatycznie ograniczone", we frazematyce za jednostki uznawane są wyrażenia odtwarzalne, a więc i te nacechowane pragmatycznie, kryterium Bogusławskiego odnosi się do „odpowiedniości funkcjonalnej", co zdaje się obejmować również aspekt pragmatyczny.

W niniejszej książce kryterium to jest brane pod uwagę, przy czym współczynnik istotności ustalono na 3. Na wysokość oceny ma wypływ fakt, że większość pragmatemów nie spełnia kryterium zamkniętości klasy[29] (ich znaczenie jest konwencjonalne, specyfikę związkowi nadaje kontekst, zatem należą do klas otwartych), co powoduje, że niejako „na wstępie" otrzymuje punkty ujemne. Problem ten koryguje właśnie wysoka punktacja współczynnika istotności kryterium, ale i tak w praktyce sama swoistość pragmatyczna nie wystarcza do uznania wyrażenia za jednostkę, musi ono spełniać także inne kryteria. W książce przyjęto, że stopień, w jakim to kryterium jest spełniane, zależy od ilości sytuacji, w których badane wyrażenie można użyć. Jeśli jest to tylko jedna bądź kilka dobrze określonych sytuacji (jak w przypadku *dzień dobry* lub *sto lat*), wyrażenie otrzymuje 2 punkty, w przypadku wyrażeń cechujących się mniejszą ograniczonością pod tym względem (np. *przy okazji*) ocena wynosi 1.

2.2.3.4. Odtwarzalność

Wiele sekwencji wyrazów, którymi posługują się ludzie, nie powstaje na bieżąco, w trakcie mówienia. Są one przywoływane z pamięci jako gotowe całości. Przekonać może o tym prosty przykład: w języku polskim mówi się *płatki kukurydziane*, inne formy, jak np. *płatki z kukurydzy* nie występują, mimo że są całkowicie poprawne zarówno pod względem gramatycznym, jak i semantycznym. Każdy człowiek dysponuje zatem leksykonem gotowych prefabrykatów językowych, które w odpowiednim momencie wykorzystuje. To zjawisko, nazywane odtwarzalnością, charakteryzuje wszystkie wielowyrazowe jednostki leksykalne.

[29] Jest to jedyne kryterium, za które wyrażenie może otrzymać ujemne punkty. Więcej na ten temat w podpunkcie 2.2.3.9.

Nic więc dziwnego, że w wielu koncepcjach frazeologii odtwarzalność pełni istotną rolę. W niektórych przypadkach (jak np. we frazematyce – zob. Chlebda 1991) odtwarzalność jest jedynym warunkiem, jaki powinno spełniać połączenie wyrazów, by zostało zaliczone w poczet frazeologizmów. W lingwistyce komputerowej koncepcja ta jest mniej popularna, choć czasem pojawia się pod inną nazwą (Seretan 2011b w definicji kolokacji podaje, że są to jednostki „prefabrykowane").

Jakkolwiek kryterium odtwarzalności jest intuicyjnie jasne i wygodne w użyciu, z jego stosowaniem wiąże się kilka problemów. Przede wszystkim nie jest jasne, na jakiej zasadzie uznawać dane wyrażenie za odtwarzalne. Człowiek, tworząc wypowiedź, najczęściej czyni to w sposób niemal automatyczny i – w wielu przypadkach – nie jest w stanie określić, z jakich jednostek faktycznie złożony jest komunikat. Analiza odtwarzalności wymagałaby specjalnie zaprojektowanej metodologii i mieściłaby się raczej w obrębie psycholingwistyki. Dodatkowo, wobec faktu, że każdy człowiek dysponuje nieco innym leksykonem, z punktu widzenia różnych ludzi te same wyrażenia mogą być odtwarzalne lub nie.

Chlebda (1991, 145 i nast.) wymienia kilka „wskaźników frazematyczności", które pomagają w ustaleniu, czy dane połączenie jest frazemem – a zatem czy jest odtwarzalne. Wskaźniki te (m.in. frekwencja, odpowiednia prozodia, wskaźniki metatekstowe sygnalizujące obecność jednostki, np. *tak zwane*) są jednak raczej heurystycznymi wskazówkami, niedającymi konkretnej odpowiedzi, jak badać samą odtwarzalność.

Warto zauważyć, że zastosowanie odtwarzalności jako jedynego albo głównego kryterium frazeologiczności powoduje, że zakres frazeologii staje się bardzo szeroki, obejmując frazy typu *tu cię boli*, *ktoś dzwoni z ogłoszenia*, *no tak, ale* itp., a także jednostki jednowyrazowe.

Wątpliwości związane z odtwarzalnością powodują, że część badaczy podchodzi do tego kryterium z ostrożnością. Według Kosek, odtwarzalność jest cechą właściwą jednostkom leksykalnym, lecz nie definicyjną – stanowi konsekwencję innych, ważniejszych własności. Przy tym: „Odtwarzalność jest niezwykle istotna, jeśli w analizie stawiamy sobie, tak jak to czyni frazematyka, pytania o twórczość i odtwórczość, o stopień schematyczności w posługiwaniu się językiem, a więc jeśli mówimy o pewnych aspektach *parole*. W spojrzeniu systemowym ta cecha nie wysuwa się na pierwszy plan" (Kosek 2008, 66). O problemach związanych z odtwarzalnością pisze również Bańko (2001, 151-152).

Ze względu na powyższe problemy, w prezentowanej tu koncepcji odtwarzalność nie jest brana pod uwagę (nie stanowi części definicji jednostki).

2.2.3.5. Stałość leksykalna

Stabilność składu leksykalnego związków frazeologicznych jest jednym z silniejszych wyznaczników frazeologiczności. Stałość leksykalna wielowyrazowej jednostki może mieć różne nasilenie. Istnieją wyrażenia, zwłaszcza cechujące się silną nieregularnością semantyczną, które wykazują się całkowitą niezmiennością, np. *udawać Greka* czy *niebieski ptak*. W innych wyrażeniach któryś ze składników może ulegać pewnym zmianom – najczęściej zastępowany jest synonimem bądź wyrazem bliskoznacznym, np. *stary pryk/dziad*, *budować/stawiać zamki na lodzie/piasku.* Czasem wymienność może przybierać takie rozmiary, że trudno określić, co właściwie jest podstawą leksykalną wyrażenia, por. *barania głowa / barani łeb / zakuty łeb / zakuta pała*. W wyrażeniach tego typu mamy właściwie do czynienia ze „szkieletem semantycznym" (nasuwa się skojarzenie z pojęciem SemR w teorii Mielczuka),

który jest wypełniany przez odpowiednie leksemy. W terminologii lingwistyki komputerowej stałość składniowa bywa określana jako niezastępowalność (non-substitutability).

Nie każde stabilne leksykalnie połączenie automatycznie zyskuje status jednostki językowej (por. wyrażenie *szczególny przypadek*, które mimo wyraźnego utrwalenia formy raczej nie zasługuje na miano jednostki języka). W tej książce przyjęto zasadę, że tylko całkowita blokada wymienności wyrazów składowych powoduje przypisanie wyrażeniu 2 punktów, natomiast jeśli liczba „zamienników" to 5 lub mniej, wyrażenie otrzymuje 1 punkt. Współczynnik istotności tego kryterium wynosi 2.

2.2.3.6. Stałość morfosyntaktyczna

Kryterium to oznacza stopień, w jakim dane wyrażenie opiera się próbom modyfikacji składniowych lub morfologicznych. Zwykłe połączenia mogą podlegać rozmaitym operacjom syntaktycznym, w przypadku wyrażeń związanych część lub wszystkie spośród takich operacji są niemożliwe. Wśród ograniczeń morfosyntaktycznych, które cechują te wyrażenia, wymienić można:

1. Niezmienność szyku – kolejność wyrazów w wyrażeniu jest ustalona, próba ich zamiany prowadzi do defrazeologizacji (por. *panna młoda* i *młoda panna*) lub powstawania form niepoprawnych, sztucznych lub nacechowanych stylistycznie (**rzeka wywiad*, *telewizji teatr*, *wino czerwone* itp.). Nie jest to cecha obligatoryjna jednostek, w niektórych inwersja jest całkowicie dopuszczalna (np. *większość absolutna – absolutna większość*) lub przynajmniej potencjalnie możliwa (np. *siostra zakonna – zakonna siostra*). Więcej na ten temat pisze Kosek (2008, 140-145);
2. Niemożność rozszerzenia – wiele złożonych jednostek opiera się próbom wstawienia dodatkowego składnika modyfikującego któryś z istniejących – np. **woda trochę utleniona*. W niektórych przypadkach może to prowadzić do defrazeologizacji (por. *bardzo cicha msza*, gdzie wyraz *bardzo* wymusza dosłowną interpretację przymiotnika *cicha*). Podobnie jak w przypadku szyku, niektóre wyrażenia dopuszczają takie modyfikacje (por. *woda na jego młyn*);
3. Niemożność stopniowania komponentu przymiotnikowego – w wyrażeniach, w których skład wchodzi przymiotnik, bardzo często zablokowana jest stopniowalność tego członu. Frazy typu **bielsza karta* czy **najlepszy duch* nie funkcjonują w języku polskim. Niemożność stopniowania oznacza, że przymiotnik ma właściwy sobie stopień (niekoniecznie równy – por. *lepsze czasy*, *najwyższy czas*). Sporadycznie zdarzają się przypadki, kiedy stopniowanie jest możliwe, jak w przypadku związków *wysokiej / wyższej / najwyższej próby* czy *grube / grubsze pieniądze*[30]. Ciekawe jest to, że stopniowanie niekoniecznie pełni tu swoją typową rolę: w zależności od kontekstu *grubsze pieniądze* mogą oznaczać więcej pieniędzy niż *grube pieniądze* (por. *zarabiałem grubsze pieniądze niż on*), ale najczęściej oba zwroty są synonimiczne;
4. Stałość liczby – istnieją nieciągłe jednostki leksykalne o ustalonej liczbie. Próba zmiany liczby prowadzi do defrazeologizacji lub zmiany znaczenia (dotyczy to zarówno *singulariów* – por. *kość słoniowa*, *obowiązek szkolny*, jak i *pluraliów tantum* – por. *ruchome piaski*, *kocie łby*) lub jest niemożliwa ze względu na specyficzne znaczenie jednostki (np. *dzieła zebrane*, *naczynia połączone*). Należy zauważyć, że nie-

[30] Nie jest jasne, czy możliwa jest forma *najgrubsze pieniądze*: w korpusie NKJP taka fraza nie występuje, wyszukiwarka Google podaje ponad 30 przykładów użyć.

możność modyfikacji liczby wynikająca z faktu, że komponent rzeczownikowy nie posiada formy pojedynczej – jak w przypadku frazy *mieć plecy* – nie stanowi kryterium frazeologiczności, gdyż niemożność ta wynika z natury wyrazu, a nie natury połączenia;

5. Defektywność paradygmatu odmiany – w przypadku niektórych wyrażeń dochodzi do zablokowania możliwości odmiany jednego członu, np. *dieta cud*. Niebrane są pod uwagę ograniczenia wynikające z realizacji określonej konstrukcji składniowej – takich jak: RZECZOWNIK + RZECZOWNIK W DOPEŁNIACZU (np. *zamach stanu*) – w takim przypadku restrykcje nie są związane z jednostkowością, a regularną strukturą wyrażenia.

Stałość morfosyntaktyczna częściowo (w zakresie szyku i niemożności rozszerzenia) pokrywa się z popularnym w lingwistyce komputerowej kryterium niemodyfikowalności (non-modifiability).

W prezentowanej w tej książce koncepcji stałości morfosyntaktycznej przypisany jest współczynnik istotności 2. 4 punkty (2 * 2) w tym przypadku otrzymują wyrażenia wykazujące co najmniej dwie spośród wyżej wymienionych własności, przy obecności tylko jednego z nich wyrażenie otrzymuje 2 (1 * 2) punkty.

2.2.3.7. Konwencjonalizacja

Koncepcja konwencjonalizacji jest właściwie niespotykana w polskiej literaturze dotyczącej frazeologii, odgrywa natomiast znaczącą rolę w anglosaskiej lingwistyce komputerowej. Termin ten wprowadził brytyjski lingwista L. Bauer (zob. Bauer 1983), definiując konwencjonalizację jako proces akceptacji danego połączenia przez użytkowników języka jako jednostki leksykalnej. Jednostki skonwencjonalizowane są regularne (również znaczeniowo), będąc tym samym w opozycji do jednostek zleksykalizowanych, których nie da się uzyskać, stosując reguły języka. Ich specyfika polega na tym, że potencjalna wieloznaczność jest zredukowana do jednego lub kilku bliskoznacznych znaczeń. Dla przykładu *pas bezpieczeństwa* mógłby teoretycznie oznaczać pas zabezpieczający wspinacza przed upadkiem z wysokości, jednak w praktyce oznacza on wyłącznie urządzenie do zabezpieczania osób znajdujących się w samochodzie lub samolocie. Dodatkowo, określony koncept jest wyrażany za pomocą skonwencjonalizowanego wyrażenia, a nie w inny, semantycznie ekwiwalentny sposób: *książka kucharska* nie może być zastąpiona wyrażeniem *książka przepisów*, *książka kuchenna* itp., mimo że byłoby to poprawne gramatycznie i semantycznie.

Konwencjonalizacja obejmuje szczególną grupę wyrażeń, która w polskiej tradycji frazeologicznej często jest pomijana, lub – w najlepszym wypadku – traktowana jako pojęcie z pogranicza frazeologii. Mowa o grupach nominalnych typu *szkoła podstawowa*, *samochód ciężarowy* itp., które w tej książce nazywane będą rzeczownikami wielowyrazowymi. Pojęcie to jest dokładniej omówione w podpunkcie 3.1.

Wyrażenie może być przyporządkowane do klasy rzeczowników wielowyrazowych, jeśli każdy jego składnik stanowi nośnik istotnej składowej znaczeniowej, a jego usunięcie powoduje wyraźne uogólnienie zwrotu. W myśl tego kryterium rzeczownikiem wielowyrazowym byłaby *czerwona krwinka* (przez usunięcie przymiotnika otrzymalibyśmy wyrażenie odnoszące się do innej – ogólniejszej – klasy), natomiast *czerwona wstążka* stanowiłaby zwykłe połączenie (tu przy braku przymiotnika i tak mamy do czynienia z tą samą klasą obiektu). Ustalenie, czy dana sekwencja jest rzeczownikiem wielowyrazowym, nie zawsze jest proste:

jak w przypadku większości innych typów jednostek, także tu występują przypadki graniczne. W rozstrzygnięciu przypadków wątpliwych może pomóc kilka reguł heurystycznych:

1) encyklopedyczność – wyrażenia, które mogłyby stanowić hasła w encyklopedii, można uznać za rzeczowniki wielowyrazowe,
2) istnienie odpowiednika jednowyrazowego – jeśli wyrażenie da się sparafrazować za pomocą jednego słowa, można je uznać za jednostkę. Przykładowo *samochód ciężarowy* potocznie określany jest jako *ciężarówka*. Tę metodę sugeruje Baldwin i Kim (2010, 8). Co prawda przykłady, jakie podają, dotyczą jedynie nieobecnych w języku polskim czasowników frazowych, ale nic nie stoi na przeszkodzie, by stosować tę regułę wobec innych typów jednostek złożonych,
3) blokada stopniowalności przymiotnika – ta własność, omawiana już przy okazji stałości morfosyntaktycznej, ma szczególne znaczenie przy rzeczownikach wielowyrazowych.

Warto podkreślić, że dwie ostatnie heurystyki można uznać za obiektywne (można określić, czy w danym przypadku własność występuje, czy nie), co nie zachodzi w przypadku encyklopedyczności. Jednocześnie encyklopedyczność dotyczyć może zdecydowanie większej ilości wyrażeń niż dwie pozostałe reguły.

Konwencjonalizacja z punktu widzenia niniejszej książki dotyczy wyłącznie jednostek o funkcji rzeczownikowej, a więc mogących nazywać klasy obiektów. Pomaga wyróżnić specjalną klasę jednostek – rzeczowniki wielowyrazowe – jest więc traktowana jako istotne kryterium (współczynnik istotności 3). Zasadniczą trudnością związaną z tym kryterium jest problem z określeniem stopnia, w jakim jest spełniane: granica między swobodnym połączeniem wyrazów tworzących grupę nominalną a rzeczownikiem wielowyrazowym jest nieostra. Przyjęto założenie, że spełnienie przynajmniej jednej z wyżej wymienionych heurystyk powoduje uzyskanie przez wyrażenie oceny 2, natomiast w przypadkach wątpliwych konwencjonalizacja oceniana jest na 1.

2.2.3.8. Nieprzekładalność

Charakterystyczną cechą wielu nieciągłych jednostek leksykalnych jest ich powiązanie z konkretnym językiem. To w obrębie danego systemu konkretną formę przybierają koncepcje, metafory, pojęcia, tworzą się typowe i chętnie wykorzystywane zbitki wyrazowe. Z tej przyczyny bardzo często jednostki wielowyrazowe sprawiają trudności w tłumaczeniu na inne języki. Rozpatrując tę kwestię od drugiej strony, można stwierdzić, że nieprzekładalność wyrażenia jest istotną wskazówką, że wyrażenie stanowi jednostkę: skoro tłumaczenia nie da się sprowadzić do prostego zastąpienia kolejnych członów ich odpowiednikami, może to oznaczać, że wyrażenie odnosi się do jakiegoś specyficznego bytu lub pojęcia, które jest „ubierane w słowa" przez poszczególne języki, z których każdy może to robić na własny sposób. Szczególnie znaczące są przypadki, gdy jednostka wielowyrazowa w innym języku ma postać jednowyrazową[31].

[31] Należy odnotować, że w niektórych językach (np. niemieckim) łączenie kilku wyrazów w jeden (*compositum*) jest stosunkowo częste i wynika z reguł systemowych. Nasuwa to wątpliwość, czy język taki jest odpowiedni do testowania kryterium. Wydaje się jednak, że jest wręcz odwrotnie: możliwość łączenia jest świadectwem, że dane połączenie można traktować (przynajmniej do pewnego stopnia) jako jednostkę. Oczywiście w takich przypadkach znaczenie dysproporcji między ilością wyrazów jest mniejsze.

Nieprzekładalność często związana jest z nieregularnością semantyczną, która sprawia, że próba dosłownego tłumaczenia przynosi wyrażenia niepoprawne, często brzmiące absurdalnie z punktu widzenia użytkownika języka docelowego. Dosłowne tłumaczenie idiomów w rodzaju *dziękuję z góry – *thank you from the mountain* jest często przedmiotem żartów.

Asumaryczność znaczenia nie jest jedyną przyczyną problemów z przekładem. Kolokacje – rozumiane jako typowe, ustalone połączenia wyrazów, zwroty skonwencjonalizowane, takie jak rzeczowniki wielowyrazowe, mają często inne formy w różnych językach. Tłumaczenie wyrażeń *łódź podwodna* jako **underwater boat* czy *kąt prosty* jako **straight angle* jest oczywistym błędem. Z drugiej strony nieprzekładalność dotyczy tylko części wyrażeń wielowyrazowych – np. *reakcja chemiczna* ma dosłowne odpowiedniki w niemal każdym języku europejskim. Nierzadkie są także przypadki, gdy dana jednostka jest przetłumaczalna na jeden język i jednocześnie nieprzekładalna na inny. Przykładem może być *wspólny mianownik* rozumiany jako termin matematyczny. W języku angielskim mamy dokładny odpowiednik: *common denominator*, natomiast w języku niemieckim jest to jednowyrazowy *Hauptnenner*, którego człon *Haupt* oznacza raczej ʻgłówny' niż ʻwspólny'.

Ekwiwalencja tłumaczenia jednostek wieloczłonowych wymaga odpowiedniości na dwóch poziomach: formalnym i semantycznym (por. Basaj 1982, Rejakowa 1982). W niniejszej książce przyjmuje się, że wyrażenie jest nieprzekładalne, jeżeli brak jest odpowiedniości na którymkolwiek z tych poziomów.

Z praktycznego punktu widzenia przy badaniu nieprzetłumaczalności konieczne jest ustalenie ograniczonego zbioru języków, w odniesieniu do których testowane jest to kryterium. W przypadku prezentowanej książki porównanie obcojęzycznych odpowiedników ograniczono do języka angielskiego, niemieckiego i rosyjskiego. Do tłumaczenia wykorzystano słowniki dwujęzyczne i internetową encyklopedię Wikipedia, której konstrukcja ułatwia porównywanie tych samych haseł w różnych językach. Współczynnik istotności kryterium ustalono na 2. Przyjęto regułę, że jeśli w co najmniej dwóch językach tłumaczenie dosłowne jest niemożliwe, kryterium jest spełniane silnie (2 punkty), w przypadku jednego języka – zauważalnie (1 punkt).

2.2.3.9. Zamkniętość / otwartość klas substytucyjnych

W wyrażeniu *cichy wspólnik* leksem *cichy* jest użyty w specyficznym znaczeniu: ʻtajny, nieujawniony'. Leksem ten w podanym znaczeniu łączy się także z innymi, np. *umowa*, *sprzymierzeniec*, *porozumienie* itp. Wyrazy te, pozostające w relacji paradygmatycznej, tworzą zbiór nazywany klasą substytucyjną. Warto zwrócić uwagę, że do tak określonej klasy substytucyjnej nie będą należeć wyrazy takie jak *zabawa* czy *dźwięk* – w wyrażeniach *cicha zabawa* i *cichy dźwięk* leksem *cichy* ma inne, konwencjonalne znaczenie. Podobnie nie zaliczymy do tej klasy wyrazów *msza*, *woda*: *cichy* w połączeniu z nimi ma również inne – choć dalekie od dosłownego – znaczenie[32].

Otwartość (czasem zwana niezamkniętością) klasy substytucyjnej i stojąca w opozycji do niej zamkniętość klasy to pojęcia wprowadzone przez Bogusławskiego (1976) jako istotna część testu na „elementarną składnikowość", pozwalającego według jego koncepcji na wyróżnianie jednostek języka. Kryterium to jest dość chętnie stosowane przez polskich

[32] Ściśle rzecz ujmując, leksemowi *cichy* w związkach frazeologicznych *cicha msza*, *cicha woda* trudno przypisać konkretne znaczenie, choć bierze on udział w tworzeniu wyrażenia o konkretnej wartości semantycznej – są to przykłady związków nierozkładalnych.

badaczy – obok Bogusławskiego używają go również Bańko (2001), Kosek (2008), Czerepowicka (2011). W myśl koncepcji Bogusławskiego wyrażenie wielowyrazowe może być uznane za jednostkę wtedy, gdy wszystkie jego elementy składowe należą do klas zamkniętych[33]. Precyzyjne zdefiniowanie tych pojęć nastręcza jednak pewnych trudności. Bogusławski (1976) określa klasę otwartą jako „scharakteryzowaną w sposób ogólny, nie zaś przez wyliczenie konkretnych obiektów", a nieco dalej podaje przykład klasy zamkniętej na przykładzie *za pomocą*, w którym przyimków łączących się z *pomocą* nie da się połączyć na podstawie cechy ogólnej, a zatem można je tylko wymienić. Jak zauważa Bańko (2001, 156) nie wynika z tego jasno, czy klasa otwarta m u s i być scharakteryzowana ogólnie (tj. jej elementów nie da się wyliczyć), czy też m o ż e być scharakteryzowana ogólnie (ponieważ istnieje ogólna cecha łącząca wszystkie elementy). W późniejszej pracy Bogusławski precyzuje: klasy otwarte to te, które „można przedłużać w sposób ograniczony co najwyżej przez przysługującą składnikom tych wyrażeń cechę ogólną, nie można zaś podać ich w postaci wyliczenia" (Bogusławski 1987, 20). Wynika z tego, że klasa otwarta to taka, której elementów nie da się wyliczyć, natomiast zamknięta to taka, której elementy wyliczyć można (ale niekoniecznie trzeba – może istnieć cecha ogólna). W sprzeczności z tym stoi jednak przykład (Bogusławski 1987, 29-31), w którym Bogusławski uznaje za otwartą klasę wyrazów łączących się z przymiotnikiem *przyrodni* (a więc – wydawałoby się – wyliczalnej, składającej się z trzech elementów: *siostra, brat, rodzeństwo*). Autor uzasadnia to faktem, że istnieje cecha ogólna charakteryzująca klasę: „właściwość ZMIANY pojęcia ʻmający CO NAJMNIEJ jedno z rodziców wspólne z___' na ʻmający TYLKO jedno z rodziców wspólne z___' oraz walencję wymagającą JEDNOWYRAZOWEJ frazy rzeczownikowej o właściwym ładunku semantycznym, tzn. z pierwszym z wymienionych zapisów". Wydaje się więc, że brak wystarczającej precyzji skutkuje tu niespójnością.

Nawet gdy pominie się nieścisłości, założenie, że klasa jest otwarta tylko w przypadku, gdy jej elementów nie da się wyliczyć, stwarza problemy. Bańko (2001, 156) pokazuje, że przyjmując tę definicję, trzeba by zaliczyć do jednostek języka wyrażenia typu *pięć po czwartej* (liczebniki określające minuty i godziny da się wyliczyć, zatem należą z definicji do klasy zamkniętej, co w myśl testu na „elementarną składnikowość" powoduje, że wyrażenie nie jest połączeniem jednostek, a zatem jest jednostką samo w sobie). Bańko zauważa też, że z formalnego punktu widzenia każda grupa elementów tworzących klasę substytucyjną może być uznana za możliwą do wyliczenia[34]. Te problemy z definicją powodują, że w pracach nad ISJP klasa otwarta definiowana jest jako taka, której elementy **można** scharakteryzować ogólnie. Tak samo klasę otwartą definiuje Kosek (2008, 48).

Z powyższą definicją klasy otwartej również wiążą się kwestie problematyczne. Przyjęcie jej powoduje, że zbiór jednostek staje się dosyć ekskluzywny: na tej podstawie Kosek uznaje za zwykłe połączenia wyrażenia typu *pociąg pospieszny*, *czerwone wino* itp., motywując to tym, że można wskazać na cechę ogólną, która definiuje ich klasy substytucyjne. W pierwszym wypadku klasę tworzą wszystkie „naziemne środki komunikacji publicznej

[33] Odróżnia to definicję Bogusławskiego od proponowanej przez Grochowskiego (1982), według której wystarczy, aby tylko jeden z elementów należał do klasy zamkniętej.

[34] Bańko pisze: „w ścisłym sensie nie ma takiego zbioru wyrażeń skończonej długości zbudowanych ze skończonej liczby symboli, którego elementów nie dałoby się policzyć i wyrazić skończoną liczbą", co jest prawdą, ale – chyba w sposób nie do końca uprawniony – wyklucza włączanie w obręb klas substytucyjnych wyrazów potencjalnych takich jak liczebniki złożone (typu *sto dwadzieścia cztery*) czy przymiotniki złożone (np. *biało-czerwony*), które nie są ograniczone (przynajmniej teoretycznie) pod względem długości.

(masowej)". W drugim – wyrazy, które zawierają w sobie element semantyczny *wino* – np. *Chianti*, *jabol*, *sikacz*[35]. Takie rozstrzygnięcia są oczywiście uprawnione z punktu widzenia przyjętych założeń, powodują jednak wykluczenie ze zbioru jednostek wyrażeń w pewnym stopniu nieregularnych znaczeniowo: mimo że *czerwone wino* faktycznie jest czerwone, a *pociąg pospieszny* najczęściej rzeczywiście przemieszcza się szybko, w obu wyrażeniach istnieje komponent semantyczny (można się o tym przekonać, analizując poprawne z semantycznego punktu widzenia wyrażenia *to czerwone wino jest właściwie różowe*, *ten pociąg pospieszny strasznie się dziś wlecze*).

W obu przypadkach można też mieć wątpliwości, czy powyższa argumentacja jest poprawna. Klasa substytucyjna łącząca się z *czerwone*, mimo że da się ją scharakteryzować ogólnie, obejmuje wyłącznie *wino* i jego hiponimy. Z tego względu jest to sytuacja inna niż na przykład w przypadku klasy łączącej się z leksemem *sąd*, obejmującej wyrazy określające typ sądu (*administracyjny*, *okręgowy*, *konstytucyjny* itp.). W przypadku *wina* mamy do czynienia nie z różnymi typami obiektów, lecz z jednym obiektem, który może być reprezentowany przez określenie ogólniejsze (*wino*) lub bardziej szczegółowe (*Chianti*, *jabol* itp.), zawierające dodatkową informację semantyczną, ale niezmieniającą jego istoty.

Problem *pociągu pospiesznego* jest inny: Kosek (2008, 53) formułuje następującą cechę ogólną, charakteryzującą klasę substytucyjną: „naziemne środki komunikacji publicznej (masowej)". Definicja ta jest zastanawiająco szczegółowa. Ponieważ z leksemem *pospieszny* łączy się bardzo ograniczona grupa wyrazów: *autobus*, *pociąg*, *tramwaj*, ewentualnie *trolejbus*, definicja jest tak skonstruowana, żeby wykluczać niepożądane środki transportu: „naziemne" wykluczają samoloty i urządzenia wodne, „komunikacji publicznej" eliminują rowery, samochody prywatne itp., „masowej" usuwa taksówki, riksze itp. Mimo tej szczegółowości i tak jest niedokładna: obejmuje również metro, które nie łączy się z wyrazem *pospieszny*, zatem nie powinno wchodzić w skład klasy substytucyjnej. Metro, co prawda, nie jest typowo naziemnym środkiem transportu, w większości przypadków porusza się pod ziemią, dlatego można próbować rozszerzyć definicję do postaci: „wyłącznie naziemne", ale wtedy problem stanowiłyby pociągi, które od czasu do czasu wjeżdżają w tunele pod górami. W rezultacie trzeba by przystać na określenie „poruszające się w większości po powierzchni ziemi środki komunikacji publicznej (masowej)". Rozumując w ten sposób, można znaleźć cechę ogólną dla niemal każdego zbioru wyrazów, co radykalnie ograniczyłoby wielkość zbioru frazeologizmów.

Podobnego problemu unika Bańko, wprowadzając pojęcie praktycznej niewyliczalności. Klasa jest według niego praktycznie niewyliczalna, jeśli jej wielkość przekracza pewien arbitralnie ustalony próg (np. 100), co jest równoznaczne z uznaniem jej za otwartą.

Punkt widzenia przyjęty w tej książce jest częściowo zbieżny z koncepcją Bańki. Liczba elementów klas substytucyjnych różnych wyrażeń waha się od jednego do „niewyliczalnie" wielu, rozmiar klas można zatem określić – precyzyjnie bądź w rozsądnym przybliżeniu – liczbowo. Wydaje się przy tym (co potwierdzają badania – zob. Woźniak 2011), że najsilniej frazeologiczne są wyrażenia, których klasy są jedno- lub kilkuelementowe, zaś stopień frazeologiczności wyraźnie spada, jeśli rozmiar (sumaryczny) klas przekracza 30. Jeśli ta liczba jest dużo wyższa, to bardzo silny sygnał, że mamy do czynienia ze swobodnym połączeniem wyrazów. W książce przyjęto, że wyrażenie spełnia silnie kryterium

[35] Z tego samego powodu wyrażenie to za „pozorny frazeologizm" uznaje Bańko (2001, 161).

zamkniętości (otrzymuje więc 2 punkty), jeśli suma elementów klas substytucyjnych jest równa lub mniejsza 5, a zauważalnie – jeśli suma jest równa lub mniejsza 30 (1 punkt). Powyżej tej liczby klasa uznawana jest za otwartą (zatem jej zamkniętość jest zerowa) i w tym przypadku – co jest odstępstwem od ogólnej zasady – wyrażenie zamiast 0 otrzymuje ocenę -1 (ponieważ otwartość klas istotnie zmniejsza szansę na uznanie wyrażenia za jednostkę), mnożoną przez współczynnik istotności wynoszący 3 (a więc całościowa punktacja, jaką wyrażenie może otrzymać za to kryterium, może wynosić -3, 3 lub 6).

2.2.3.10. Ekspresywność i obrazowość

Ekspresywność i obrazowość jako kryteria cechujące związki frazeologiczne przywoływane są często, jednak podkreśla się trudność w ich precyzyjnym zdefiniowaniu[36]. Wydaje się, że ogólną i łączącą je cechą jest ich opozycja względem neutralności stylistycznej. W przypadku ekspresywności wynikałoby to z obecności w wyrażeniu czynników wartościujących, intensyfikujących, w przypadku obrazowości – działających na wyobraźnię, zmysły, pozwalających na mentalną wizualizację wyrażenia (por. Skubalanka 1972, 125-126). Przykładem frazeologizmu cechującego się ekspresywnością i obrazowością może być wyrażenie *maczać w czymś palce* – wartościujące i wywołujące u słuchacza plastyczny obraz. Zbliżone do tych kryteriów jest pojęcie intensywności wyrażenia (zob. Straś 2008).

Cechy te, blisko związane z metaforycznością (por. Kosek 2008, 16), są na tyle wyraziste, że często traktowane zostają – zarówno przez językoznawców, jak i przez zwykłych użytkowników języka – jako podstawowe wyznaczniki frazeologiczności (por. Pajdzińska 1988). Jednak, jak zauważa Bogusławski (1989, 18): „zgodnie z tym stereotypem, frazeologia to szczególna, a zarazem drugorzędna, w znacznej mierze »ornamentacyjna«, ostatecznie zaś niekonieczna część leksyki". Ponadto opieranie definicji związku na tym kryterium powoduje wyodrębnienie grupy wyrażeń o niewątpliwie silnym stopniu frazeologiczności, ale stosunkowo niewielkiej.

W przedstawianej w tej książce koncepcji ekspresywność i obrazowość, jako powiązane ze sobą, potraktowane zostały zbiorczo (przy czym ekspresywność jest traktowana jako wyznacznik podstawowy, a obrazowość – pomocniczy). Przypisano im współczynnik istotności 2. Ocena siły, z jaką wyrażenie spełnia to kryterium, nastręcza problemów: dosyć trudno o konkretne, łatwo mierzalne wykładniki kryterium. W przypadku ekspresywności oceniana jest intensywność ładunku emocjonalnego, występowanie sformułowań kolokwialnych lub wulgarnych, w przypadku obrazowości stopień oryginalności obrazu i konkretność obiektów przywoływanych w obrazie (im większa konkretność, tym silniejsza obrazowość, np. *zielone idee* są obrazowe w mniejszym stopniu niż *stary piernik*).

2.2.4. Definicja wielosegmentowej jednostki leksykalnej

Przedstawione powyżej kryteria, ich sposoby oceny i istotność są podstawą przyjętej w tej książce konkretnej definicji, która brzmi następująco:

Za wielosegmentową jednostkę leksykalną uznawany jest przynajmniej dwuelementowy zbiór leksemów, który na podstawie punktacji uzyskanej przez zsumowanie ocen poszcze-

[36] Por. Bogusławski (1989) o obrazowości: „nie jest łatwo określić, na czym ona polega, ale wstępnie nie sposób jest odrzucić taką charakterystykę", a także Pajdzińska (1988, 9): „Mimo dużej popularności wymienione terminy [metafora, obraz, obrazowość – MW] są niedookreślone".

gólnych kryteriów jednostkowości uzyskał minimum 8 punktów. Brane pod uwagę kryteria i ich współczynnik istotności są następujące:

1) nieregularność: 4,
2) konwencjonalizacja: 3,
3) pragmatyczność: 3,
4) stałość leksykalna: 2,
5) stałość składniowa: 2,
6) nieprzetłumaczalność: 2,
7) ekspresywność / obrazowość: 2,
8) zamkniętość klas substytucyjnych: 3.

Każde kryterium może zostać ocenione jako spełniane przez wyrażenie w stopniu silnym, zauważalnym lub niewielkim / zerowym. Wyrażenie na podstawie danego kryterium może otrzymać następującą punktację:

- stopień silny: 2 punkty,
- stopień zauważalny: 1 punkt,
- stopień niewielki / zerowy: 0 punktów (wyjątek stanowi kryterium zamkniętości, w którego przypadku ostatni stopień powoduje przyznanie oceny -1).

Ocena, jaką wyrażenie otrzymuje za dane kryterium, jest mnożona następnie przez jego współczynnik istotności. Suma tych wyników (ocena x współczynnik istotności) daje ocenę całościową „jednostkowości" wyrażenia.

2.2.5. Przykłady klasyfikacji

Proces oceny wyrażenia stosowany w niniejszej książce jest dosyć złożony, dlatego w tym miejscu przedstawiam kilka przykładów ilustrujących przebieg całej procedury:

- ***czerwone wino***
  - **nieregularność**
    czerwone wino jest faktycznie czerwone, jednak „czerwień" w tym przypadku dotyczy nie tylko barwy wina, ale przede wszystkim określa sposób jego pozyskania, jego typ – zatem występuje tu dodatkowa wartość semantyczna. Z tego powodu nieregularność oceniona zostaje na 1 (pozostałe typy nieregularności – leksykalna i składniowa – nie występują)**,**
  - **pragmatyczność**
    nie występuje – ocena 0**,**
  - **konwencjonalizacja**
    wyrażenie *czerwone wino* reprezentuje klasę obiektów, które można zdefiniować jako 'napój alkoholowy uzyskany z czerwonych (ciemnych) winogron', zatem podstawowa przesłanka pozwalająca uznać frazę za skonwencjonalizowaną jest spełniona. Z drugiej strony w prosty sposób można uogólnić wyrażenie do postaci *wino*, kosztem pominięcia pewnej cechy – zatem ustalenie istotności tej cechy wydaje się kluczowe w celu określenia stopnia konwencjonalizacji. Jeśli pominąć wyraz *czerwony*, wyrażenie dalej pozostanie winem – ale jednocześnie przymiotnik ten ma wyraźnie inną jakość niż przymiotniki typu *stare*, *smaczne*, *mocne* itp. Ten pierwszy określa typ wina, kolejne – jego cechy. Co do „encyklopedyczności" czerwonego

wina nie ma jednoznacznej odpowiedzi, „tradycyjne" encyklopedie (PWN, Britannica) omawiają je pod hasłem *wino*, źródła „elektroniczne" (Wikipedia, WordNet) poświęcają mu osobne hasło[37]. Biorąc pod uwagę wszystkie powyższe argumenty, z których część przemawia za uznaniem konwencjonalizacji *czerwonego wina*, a pozostałe – przeciwko, wydaje się, że właściwe jest ustalenie stopnia tego kryterium na 1,

– **stałość leksykalna**
wyrazu *czerwone* nie można zastąpić innym bez zmiany jego znaczenia, natomiast *wino* można w określonych sytuacjach zastąpić hiponimami (np. *burgund*) lub hiperonimami (*trunek*) – z tego powodu stałość leksykalna wyrażenia oceniona zostaje na 1,
– **stałość morfosyntaktyczna**
wyrażenie charakteryzuje się umiarkowaną stałością leksykalną: szyk można zmieniać, składniki wyrażenia można rozdzielać innymi elementami, nie jest natomiast możliwe stopniowanie przymiotnika – wyrażenie za to kryterium otrzymuje ocenę 1**,**
– **nieprzekładalność**
tłumaczenie na język angielski i rosyjski jest dosłowne, natomiast w języku niemieckim odpowiednikiem *czerwonego wina* jest jednowyrazowe *Rotwein*, zatem wyrażenie w związku z tym kryterium otrzymuje 1 punkt,
– **zamkniętość klas**
z przymiotnikiem *czerwone* w tym znaczeniu może łączyć się wyłącznie *wino* (nie wlicza się tu hiponimów), zatem otwartość prawostronna jest zerowa, natomiast do klasy substytucyjnej, jaką otwiera wino, należą przymiotniki określające typu *białe*, *różowe*, *wytrawne* – ich liczba oscyluje w granicach 10. Zatem stopień kryterium określany jest jako zauważalny, co daje 1 punkt,
– **ekspresywność i obrazowość**
wyrażenie nie jest nacechowane ekspresywnie, natomiast występuje pewna obrazowość, jednak ze względu na niewielki stopień i mniejszą istotność obrazowości całościowo kryterium ocenione zostaje na 0.

[37] W opracowywaniu materiału korzystałem z następujących źródeł: PWN (http://encyklopedia.pwn.pl), Britannica (www.britannica.com), Wikipedia (http://pl.wikipedia.org, http://en.wikipedia.org), Słowosieć (http://plwordnet.pwr.wroc.pl) i WordNet (http://wordnet.princeton.edu). Te dwie ostatnie pozycje (polska i angielska) nie są wprawdzie encyklopediami, ale stanowią bardzo duże bazy wielosegmentowych jednostek leksykalnych dostępnych, przy czym filozofia umieszczania w nich haseł jest podobna do stosowanej w encyklopediach. Część ze źródeł jest anglojęzyczna, co najczęściej nie stanowi przeszkody, gdyż badana cecha – jako ontologiczna – jest w dużym stopniu niezależna od języka. Umowny podział na źródła „tradycyjne" i „elektroniczne" opiera się na podstawowej formie dystrybucji: encyklopedie PWN i Britannica są wydawane przede wszystkim jako materiały książkowe, ich baza internetowa jest wobec formy książkowej wtórna; Wikipedia, Słowosieć i WordNet istnieją przede wszystkim jako bazy elektroniczne.

Ocena *czerwonego wina* kształtuje się następująco:

Kryterium	Ocena	Wsp. istotności	Suma
nieregularność	1	4	4
pragmatyczność	0	3	0
konwencjonalizacja	1	3	3
stałość leksykalna	1	2	2
stałość morfosyntaktyczna	1	2	2
nieprzekładalność	1	2	2
zamkniętość klas	1	3	3
ekspresywność / obrazowość	0	2	0
Razem			**16**

Tabela II.1: Ocena wyrażenia ***czerwone wino***

Sumaryczna ocena *czerwonego wina* wynosi 16 punktów, zatem w myśl definicji (minimum 8 punktów) jest to pełnoprawna jednostka leksykalna.

- ***ręce do góry!***
  - **nieregularność**
    nie występuje w wyrażeniu – 0 punktów**,**
  - **pragmatyczność**
    wyrażenie jest przykładem typowego pragmatemu – używanego w określonych sytuacjach, którego pełne znaczenie musi brać pod uwagę kontekst sytuacyjny; jedynym ekwiwalentem jest *poddaj się!* (*rzuć broń!* używane w podobnej sytuacji posiada nieco inne konotacje) – zatem ocena stopnia kryterium wynosi 2,
  - **konwencjonalizacja**
    wyrażenie jest wykrzyknikiem, nie nazywa żadnej klasy rzeczywistych obiektów – ocena 0,
  - **stałość leksykalna**
    wyrażenie może występować wyłącznie w jednej formie – ocena 2,
  - **stałość morfosyntaktyczna**
    jako że nie jest możliwa żadna modyfikacja wyrażenia pod tym kątem, kryterium jest spełniane w silnym stopniu: ocena 2,
  - **nieprzekładalność**
    tłumaczenie zwrotu na język angielski i rosyjski może być dosłowne, w przypadku języka niemieckiego wyrażenie przyimkowe *do góry* zastępowane jest przysłówkiem *hoch* (*wysoko*) – zatem ocena 1,
  - **zamkniętość klas**
    w przypadku pragmatemów zamkniętość klas jest najczęściej niska i tak jest także tutaj: oba człony (wyrażenie przyimkowe *do góry* jest tu traktowane ze względów semantycznych jako jeden element) mogą łączyć się z wieloma wyrazami, zatem ocena stopnia kryterium wynosi 0 (co powoduje przyznanie ujemnych punktów),

– **ekspresywność i obrazowość**
ekspresywność wyrażenia jest wysoka, czego formalnym wykładnikiem jest wykrzyknik na końcu wyrażenia, podobnie jak obrazowość – wyrażenie dokładnie obrazuje czynność, którą jego nadawca pragnie wywołać u odbiorcy. Ocena stopnia kryterium wynosi więc 2 punkty.

Ocena wyrażenia *ręce do góry!*:

Kryterium	Ocena	Wsp. istotności	Suma
nieregularność	0	4	0
pragmatyczność	2	3	6
konwencjonalizacja	0	3	0
stałość leksykalna	2	2	4
stałość morfosyntaktyczna	2	2	4
nieprzekładalność	1	2	2
zamkniętość klas	-1	3	-3
ekspresywność / obrazowość	2	2	4
Razem			**17**

Tabela II.2: Ocena wyrażenia ***ręce do góry!***

Wyrażenie uzyskało 17 punktów, jest więc pełnoprawną jednostką leksykalną.

2.2.6. Zalety i wady proponowanego podejścia do klasyfikacji

Przedstawiona w tej książce definicja nieciągłej jednostki leksykalnej ma w zamierzeniu jak najlepiej oddawać jej istotę. Chociaż w propozycjach niektórych badaczy pojawiały się pomysły w pewnym stopniu zbliżone (zob. podpunkt 2.2.1), wydaje się, że w całości jest to koncepcja wyraźnie odróżniająca się od innych. W tym miejscu przedstawiam plusy i minusy jej przyjęcia. Wśród zalet można wskazać:

1) stopniowalność (zarówno ogólnego pojęcia jednostkowości, jak i w obrębie poszczególnych kryteriów) lepiej oddaje istotę tego zjawiska językowego: wiele przytoczonych przykładów potwierdza fakt, że niektóre wyrażenia są w silniejszym stopniu jednostkowe (frazeologiczne) niż inne,
2) wielość kryteriów pozwala na pełniejsze uchwycenie i opisanie całości zjawiska, w obrębie którego znajduje się wiele klas wyraźnie różniących się od siebie, a mimo to zaliczanych w poczet jednostek wieloczłonowych,
3) liczbowa ocena pozwala na wymierne porównywanie wyrażeń, a także na dzielenie zbioru na klasy (binarne: jednostka / nie-jednostka, trójkowe: jednostka pewna / jednostka niepewna / nie-jednostka itp.) w zależności od pozycji na liście,
4) arbitralność rozstrzygnięć, która jest nieodłącznym elementem wyróżniania jednostek, ma mniejsze konsekwencje: oceniane jest wiele kryteriów, a niezgodność anotatorów (problem omawiany w punkcie 2.1) dotyczy najczęściej jednego (lub kilku, lecz najczęściej niezbyt wielu) z nich, zatem niepewność dotyczy tylko pewnej części całościowej oceny. Wpływa to na zwiększenie obiektywności anotacji,

5) metoda jest elastyczna: w prosty sposób można ją dostosowywać lub nawet istotnie modyfikować za pomocą odpowiedniego doboru kryteriów i ich współczynników istotności, a także ustalania progu (progów) dzielących wyrażenia na klasy,
6) z punktu widzenia lingwistyki komputerowej wyrażenie stopnia frazeologiczności za pomocą liczb stanowi wygodne narzędzie do komputerowej reprezentacji jednostek.

Omawiana metoda posiada jednak również swoiste wady:
1) wielość kryteriów koniecznych do brania pod uwagę powoduje, że ocena poszczególnych wyrażeń jest żmudniejsza i bardziej czasochłonna,
2) konieczność podejmowania arbitralnych decyzji nie jest wyeliminowana: można wręcz powiedzieć, że takich decyzji do podjęcia jest więcej; wydaje się jednak, że wskazane wcześniej zalety takiego „rozczłonkowania" przewyższają jego wady: niepewność przy ocenie niektórych kryteriów jest kompensowana przez pewność przy ocenie pozostałych,
3) przyjęta trzystopniowa skala oceny (kryterium spełniane silnie / zauważalnie / wcale) może nasuwać wątpliwości – jako zbyt mało precyzyjna. Należy wszakże zauważyć, że im szersza jest stosowana skala, tym większe stają się różnice w wypadku rozbieżności zdań między anotatorami. Proponowane rozwiązanie jest kompromisem między dążeniem do dokładności a podatnością na rozbieżną interpretację.

3. Opis jednostek wielosegmentowych

Zaprezentowana wyżej definicja jednostki wielosegmentowej staje się podstawą do wyróżnienia pewnej grupy wyrażeń wielosegmentowych, które w tekście pełnią funkcję analogiczną do pojedynczych wyrazów i powinny być traktowane jako jednostki. Wyrażenia takie nie stanowią jednorodnej grupy, ale można wyróżnić pewne, wspólne dla nich wszystkich cechy. W poniższej sekcji przedstawiam praktyczną klasyfikację jednostek wielosegmentowych (podpunkt 3.1), a także omawiam zagadnienia dotyczące istoty takich jednostek (podpunkty 3.2-3.8). Ze względu na cel i zakres książki opisane tu ich cechy ograniczają się do zjawisk dotyczących definicji i aspektów związanych z NLP. Z tego powodu nie poruszam tu między innymi problemu motywacji jednostek wielosegmentowych czy ich roli w systemie językowym.

3.1. Typologia jednostek przyjęta w książce

Analiza zjawisk językowych, które mogą być uznane za jednostki, wskazuje, że da się w ich obrębie wyróżnić kilka charakterystycznych klas, różniących się własnościami. Były one – bardziej lub mniej dogłębnie – opisywane w cytowanych wyżej pracach. Poniżej przedstawiam główne typy, do których mogą być zaklasyfikowane poszczególne wyrażenia, na podstawie zaprezentowanej w tej książce definicji (omawiam także kategorie przysparzające trudności). Poszczególne typy różnią się cechami semantycznymi, składniowymi lub pragmatycznymi. Mają również różny status z punktu widzenia lingwistyki komputerowej. Ze względu na niejednorodność kryteriów, które posłużyły przedstawionej klasyfikacji, można ten podział uznać za mało konsekwentny – powstał jednak nie w celu systematycznego opisu, a raczej dla uchwycenia praktycznych różnic.

1. **Wyrażenia nieregularne.** Stanowią trzon tradycyjnie rozumianej frazeologii, największą grupę wyrażeń, jakie można znaleźć w słownikach frazeologicznych. Co do ich jednostkowości właściwie nie ma wątpliwości. Nieregularność wynika najczęściej (choć nie tylko – patrz podpunkt 2.2.3.1) z obecności metafory – często charakteryzującej się jednocześnie nierozkładalnością, co jeszcze bardziej zwiększa specyfikę wyrażenia. Uwzględnianie wyrażeń należących do tej kategorii jest z punktu widzenia lingwistyki komputerowej kluczowe. Nieregularność powoduje, że automatyczny system przetwarzający język naturalny musi mieć szczegółowe informacje o tym, jak taką jednostkę rozpoznawać i traktować.
2. **Rzeczowniki wielowyrazowe**. Są to jednostki typu *książka telefoniczna*, *owczarek niemiecki* czy *list polecony*. Nazywane bywają wyrażeniami gatunkującymi, ustalonymi połączeniami nieidiomatycznymi, zestawieniami, związkami terminologicznymi lub rzeczownikami wielowyrazowymi – i ten ostatni termin jest wykorzystywany w niniejszej książce jako nazwa zjawiska. Status wyrażeń należących do tej kategorii w porównaniu z klasą wyrażeń nieregularnych jest zdecydowanie bardziej kontrowersyjny; większość koncepcji wywodzących się z tradycyjnej frazeologii nie uznaje ich za jednostki lub traktuje je jako kategorię niepewną. Bogusławski (1976, 357) na przykładzie wyrażenia *zwierzę domowe* wskazuje, że powinny one być uznawane za jednostki języka[38], ale np. Bańko (2001, 160) krytykuje umieszczenie w SJPSzym połączeń tego typu, nazywając je pozornymi frazeologizmami. Lewicki (1986) uznaje je za elementy z pogranicza frazeologii. Mimo to fakt, że wskazują one na klasy wyrażeń, a nie na konkretne obiekty (posiadają denotat, ale niekoniecznie desygnat), wydaje się potwierdzać, że należy je traktować jako osobne jednostki leksykalne. Ten punkt widzenia przyjmuje także najczęściej lingwistyka komputerowa (zob. Baldwin i Kim 2010, Sag i in. 2002). Choć mogłoby się wydawać, że ze względu na niewielki stopień asumaryczności znaczeniowej, bądź całkowity jej brak, pominięcie tego typu wyrażeń w słownikach komputerowych nie powinno przynosić poważnych konsekwencji, jednak stanowią one bardzo dużą część leksykonu – ich obecność wydatnie poprawia skuteczność różnego typu automatycznych systemów przetwarzania języka.

   Rzeczowniki wielowyrazowe są bardzo zbliżone do wielowyrazowych terminów. Według Bańki (2001, 160) podobieństwo to opiera się na ich rzeczownikowym charakterze oraz na tym, że nie są ekspresywne ani obrazowe, a ich zakres znaczeniowy mieści się w obrębie znaczenia pierwszego składnika. Ten tok rozumowania wydaje się w pewnym stopniu słuszny, choć nie do końca precyzyjny: po pierwsze terminologia obejmuje nie tylko wyrażenia o charakterze rzeczownikowym, po drugie nie zawsze pierwszy składnik jest podstawowy semantycznie, jak na przykład w przypadku wyrażeń *absolutna większość*, *niezidentyfikowany obiekt latający*. Przede wszystkim jednak istota tego podobieństwa leży chyba gdzie indziej: wyrażenia z obu sfer reprezentują abstrakcyjne klasy obiektów, będące „ideami" (w rozumieniu platońskim). Rzeczowniki wielowyrazowe i terminy (te mające charakter rzeczow-

[38] Argumentacja Bogusławskiego nie opiera się na konwencjonalizacji, która jest podstawą wyróżniania rzeczowników wielowyrazowych w tej pracy: według niego „domowość", z jaką mamy do czynienia w *zwierzęciu domowym* jest semantycznie unikalna. Wydaje się jednak, że można te dwa spojrzenia pogodzić, interpretując ograniczenie wieloznaczności związane z konwencjonalizacją jako przejaw unikalności semantycznej.

nikowy) łączy zatem budowa i funkcja, różnią się właściwie wyłącznie obszarem językowym, w którym są wykorzystywane. Przy tym granica oddzielająca oba pojęcia nie jest ostra (por. Kosek 2008, 70), czasem nie ma pewności, do której grupy zakwalifikować jednostkę. Dla przykładu sformułowanie *szyba przednia* z jednej strony funkcjonuje w kodzie ogólnym, określając dosyć pospolity przedmiot, z drugiej – w branży motoryzacyjnej ma wyraźnie określoną definicję i przeznaczenie. Podobnie *szkoła podstawowa* jest frazą używaną na co dzień, ale jednocześnie ma dokładnie określoną pozycję w terminologii administracyjno-urzędniczej.

Same terminy różnią się od siebie również stopniem specjalizacji: przytoczone wyżej wyrażenia – rozumiane przez większość użytkowników języka – są na jednej krawędzi skali, na której drugim końcu lokują się terminy wybitnie specjalistyczne, jak np. *zastawka mitralna* (termin medyczny) czy *strobilizacja monodyskoidalna* (termin biologiczny), które znają niemal wyłącznie osoby zajmujące się daną dziedziną. Wydaje się zatem, że różnica między rzeczownikami i terminami wielowyrazowymi może być widziana jako różnica między specjalizacją wyrażenia: te pierwsze charakteryzuje (niemal) zerowa specjalistyczność, w przeciwieństwie do tych drugich, w których stopień specjalizacji wyrażenia jest wyższy[39]. Podobieństwo tych dwóch zjawisk językowych jest często zauważane w lingwistyce komputerowej (por. Ramisch 2012, 24; Manning i Schütze 1999, 152). Wiele prac dotyczących ekstrakcji wyrażeń wielowyrazowych można z powodzeniem stosować zarówno do wyodrębnienia jednostek leksykalnych stanowiących część kodu ogólnego, jak i wyrażeń terminologicznych (choć istnieją metody zaprojektowane wyłącznie w celu ekstrakcji terminów – zob. Frantzi i in. 2000).

3. **Pragmatemy**. Terminem tym – zaczerpniętym z koncepcji Mielczuka – określam tu wyrażenia wielowyrazowe, których użycie związane jest z konkretną sytuacją. Mogą to być formuły mocno skodyfikowane, używane w oficjalnych sytuacjach, np. *proszę o ciszę*, powitania / pożegnania, zwroty używane w korespondencji, komendy wojskowe, formuły rytuałów religijnych itp.; mogą to być także wyrażenia ustalone w mniejszym stopniu – np. *nie parkować / parkowanie wzbronione / zakaz parkowania*. Powiązanie ze szczególną sytuacją powoduje, że – z punktu widzenia automatycznego przetwarzania języka – istnieje konieczność uwzględniania pragmatemów w leksykonie (a więc traktowania ich jako jednostki). W przeciwnym wypadku zarówno analiza, jak i synteza tekstu byłyby narażone na błędy i niedokładności – traktowałyby pragmatemy jako połączenia regularne, niewyróżniające się niczym szczególnym.
4. **Wielosegmentowe nazwy własne**. Nazwy własne są kategorią specyficzną. Z jednej strony na pewno nie są swobodnym połączeniem wyrazów, z drugiej – mają szczególną, określoną referencję powiązaną z konkretnym rzeczywistym obiektem i – co za tym idzie – w celu prawidłowej interpretacji wyrażenia wymagają od użytkownika języka wiedzy pozajęzykowej. Tradycyjnie tę grupę wyrażeń traktuje się jako niezależną od frazeologii. Nazwy własne w języku polskim mają szereg cech charakterystycznych: rządzą się własnymi regułami ortograficznymi, bardzo rzadko używa się ich w liczbie mnogiej, często cechuje je specyficzna odmiana. Długość nazwy własnej jest ograniczana tylko przez względy praktyczne i może być mocno

[39] Opisywane podobieństwo dotyczy oczywiście wyłącznie terminów o charakterze rzeczownikowym (które stanowią jednak przytłaczającą większość wyrażeń terminologicznych).

zróżnicowana (por. jednowyrazowa nazwa góry *Giewont* i znacznie dłuższa nazwa krakowskiej kawiarni *Pierwszy Lokal Na Stolarskiej Po Lewej Stronie Idąc Od Małego Rynku*). Wiele nazw własnych tworzonych jest według składniowego schematu GRUPA NOMINALNA W MIANOWNIKU + CIĄG GRUP NOMINALNYCH W DOPEŁNIACZU, przy czym poszczególne grupy dopełniaczowe oddzielane są przecinkiem lub spójnikiem *i* (należy zauważyć, że ciąg dopełniaczowych grup nominalnych w wielu wypadkach ogranicza się tylko do jednej takiej grupy), np. *Polski Związek Piłki Nożnej*, *Ministerstwo Spraw Wewnętrznych i Administracji*. Te własności, będące źródłem pewnej regularności w obrębie nazw własnych, ułatwiają automatyczne ich rozpoznawanie.

Duże podobieństwo do nazw własnych wykazują deskrypcje określone. Są to wyrażenia typu *najwyższa góra świata*, które – podobnie jak nazwy – odnoszą się do konkretnego wycinka rzeczywistości, najczęściej nazwę zastępując (w podanym przykładzie chodzi oczywiście o *Mount Everest*), nie dzieląc jednak specyficznych dla nazw cech formalnych.

Istnieją również pewne wyrażenia, które Kosek (2008, 63-66) klasyfikuje jako coś pośredniego między nazwami własnymi i deskrypcjami określonymi, np. *lista Wildsteina*, *system boloński* czy *powstanie listopadowe.* One również odnoszą się do konkretnego obiektu istniejącego w rzeczywistości, wymagają zatem wiedzy pozajęzykowej, natomiast formalnie bardzo przypominają konwencjonalne nieciągłe jednostki leksykalne.

W lingwistyce komputerowej rozpoznawanie nazw własnych jest bardzo ważnym zadaniem, istnieje cały obszar badań poświęcony wyłącznie temu zagadnieniu (Named Entity Recognition, w skrócie NER). Dysponuje on własnymi metodami pozwalającymi rozpoznawać i przetwarzać jednostki tego typu w tekście. Z tego względu w niniejszej książce nazwy własne (a także deskrypcje określone i wyrażenia typu *powstanie listopadowe*) traktowane będą jako zagadnienia pokrewne, niebędące jednak częścią przedmiotu badań.

3.2. Perspektywa odbiorcy / nadawcy

To samo wyrażenie może być różnie traktowane w zależności od przyjętej perspektywy. Jeżeli na akt komunikacji patrzymy od strony odbiorcy komunikatu, istotną kwestią jest interpretacja wypowiedzi. Jeżeli natomiast na wypowiedzenie spojrzeć od strony nadawcy komunikatu, wtedy na pierwszy plan wysuwa się proces tworzenia wypowiedzi, w którym istotny jest sposób, w jaki ona powstaje. Bańko (2001, 152) pokazuje, że uznanie wyrażenia za jednostkę lub swobodne połączenie może zależeć od przyjęcia odpowiedniej perspektywy: *płatki owsiane* są z punktu widzenia odbiorcy tekstu regularne – znaczenie komponentów jest jasne i spotykane w innych połączeniach, jak np. *mąka owsiana* czy *płatki kukurydziane*. Jednak z perspektywy nadawcy komunikatu jest nieco inaczej: autor wypowiedzi nie może wybrać sekwencji semantycznie równoznacznej, jak np. *płatki z owsa*. Przyjęcie punktu widzenia nadawcy wypowiedzi jest fundamentem frazematyki.

W tradycji anglosaskiej odpowiednikiem tej koncepcji jest pojęcie idiomów dekodowania i kodowania (ang. *idioms of decoding / encoding* – zob. np. Moon 1998, Baldwin 2006). Jeżeli nieciągła jednostka jest idiomem dekodowania, to oznacza, że nie da się jej uznać za regularne połączenie składników – żeby poprawnie ją zinterpretować, trzeba dysponować odpowiednią wiedzą. Idiom kodowania to z kolei jednostka, której użycie wymuszone jest

nie przez konkretne reguły języka, lecz przez zwyczaj językowy: osoba biegle władająca danym językiem jest w stanie wskazać sformułowania, których można w danej sytuacji użyć, w odróżnieniu od innych – niepoprawnych, mimo że niełamiących żadnych reguł języka. Idiomami dekodowania są wyrażenia cechujące się asumarycznością znaczenia – np. *cztery kąty*, idiomami kodowania – wyrażenia skonwencjonalizowane, np. wspomniane *płatki owsiane*. Warto zauważyć, że jednostka będąca idiomem dekodowania jest jednocześnie idiomem kodowania, lecz ta relacja nie musi być odwrotna (nie każdy idiom kodowania jest jednocześnie idiomem dekodowania).

Istnieje subtelna różnica między opisanymi wyżej podejściami. Opozycja nadawca – odbiorca służy bardziej określeniu zakresu nieciągłych jednostek, wpływa na definicję jednostki. Z perspektywy koncepcji idiomów kodowania / dekodowania mamy do czynienia raczej z opisem jednostki – wyróżnieniem pewnej cechy charakterystycznej. Ten drugi punkt widzenia jest odpowiedniejszy z perspektywy niniejszej książki: niektóre kryteria składające się na definicję jednostki wielowyrazowej odpowiadają bowiem za wyodrębnienie idiomów dekodowania (np. nieregularność), inne – kodowania (jak konwencjonalizacja).

3.3. Wieloznaczność

Wieloznaczność nieciągłych leksemów może przejawiać się na dwa sposoby. Pierwszy z nich jest odpowiednikiem wieloznaczności pojedynczych wyrazów: dana jednostka ma dwie lub więcej interpretacji semantycznych, jej właściwa interpretacja zależy od kontekstu. Dla przykładu *czarna magia* może oznaczać ‘przywoływanie i wykorzystywanie złych mocy’ lub ‘coś bardzo skomplikowanego i trudnego do zrozumienia’. Drugi rodzaj wieloznaczności zachodzi, gdy dane wyrażenie może stanowić jednostkę lub zwykłe połączenie: *bułka z masłem* w zależności od sposobu użycia może oznaczać ‘bardzo proste zadanie’ lub odpowiednio przygotowane pieczywo.

Wieloznaczność dotyczy najczęściej pragmatemów i – w mniejszym stopniu – wyrażeń nieregularnych znaczeniowo. W przypadku terminów i rzeczowników wielowyrazowych jest dużo rzadsza.

Z punktu widzenia lingwistyki komputerowej wieloznaczność jest kłopotliwą cechą. Moduł analizujący i rozpoznający nieciągłe jednostki w tekście powinien radzić sobie z obomа rodzajami niejednoznaczności. Nie jest to łatwe, ze względu na kształtową identyczność wariantów. Podstawowe techniki odróżniania jednakowo wyglądających jednostek i swobodnych połączeń opierają się na analizie otoczenia: kontekst metaforycznej *bułki z masłem* przeważnie w zauważalny sposób różni się od kontekstu tego wyrażenia użytego w dosłownym znaczeniu. Czasem występują też dodatkowe czynniki ułatwiające rozróżnienie, najczęściej związane z ustaleniem morfosyntaktycznym jednostki. Często niemożliwe bądź mało prawdopodobne jest wystąpienie jednostki w innej liczbie niż podstawowa: *wybory prezydenckie* nie występują w liczbie pojedynczej, w związku z czym można mieć pewność, że sekwencja *wybór prezydencki* jest zwykłym syntaktycznym połączeniem. Innym czynnikiem ułatwiającym rozróżnienie bywa rodzaj gramatyczny wyrażenia, który można określić na podstawie kontekstu: w przykładzie „Dzikie linie przywożą tu różnych *niebieskich ptaszków*” (Bąba, Liberek 2001, cyt. za Kosek 2008, 116) rodzaj wyrażenia *niebieski ptaszek* jednoznacznie wskazuje na to, że mamy do czynienia z jednostką.

Efektywność wymienionych metod rozstrzygania wieloznaczności nie jest imponująca, w wielu przypadkach informacji zawartych w tekście jest za mało, by zapewniały wystarcza-

jącą skuteczność. Problem jest jeszcze większy, gdy różne warianty znaczeniowe wyrażenia są jednostkami wielosegmentowymi: różnice w takim wypadku są jeszcze subtelniejsze.

3.4. Jednostki wielosegmentowe a części mowy

Lewicki (1986), omawiając kwestie składniowe związków frazeologicznych, wprowadza pojęcia składni wewnętrznej i zewnętrznej. Ta pierwsza odnosi się do relacji między członami wyrażenia, druga – do relacji między wyrażeniem traktowanym całościowo a innymi elementami wypowiedzi. Dla składni zewnętrznej kluczowy jest aspekt przynależności wyrażenia do określonych klas funkcjonalnych, w odniesieniu do tradycyjnych wyrazów nazywanych częściami mowy.

Trudno jednoznacznie odpowiedzieć na pytanie, czy jednostkom wielosegmentowym można przypisywać tradycyjnie rozumiane pojęcie części mowy. Kosek (2008, 178-187) omawia wątpliwości dotyczące tego, czy np. *biały kruk* jest rzeczownikiem. Największym problemem wydaje się być tu nieciągłość składniowa niektórych jednostek. Pomimo tego powszechnie uznaje się, że jednostki wielosegmentowe – nawet jeśli nie do końca można je określić jako rzeczowniki, czasowniki itp. – spełniają w tekście role funkcjonalnie odpowiadające częściom mowy. H. Wróbel pisze: „Skoro przyjmujemy, że podstawową właściwością jednostek leksykalnych, dzielącą je na klasy funkcjonalne, jest ich zróżnicowany udział w konstruowaniu wypowiedzeń, a co za tym idzie, możliwość ew. niemożliwość tworzenia związków składniowych z jednostkami innych klas, to własność ta przysługuje niewątpliwie również wszystkim jednostkom wielosegmentowym" (Wróbel 1995, 16). Zatem *czerwone wino* z punktu widzenia składni zachowuje się jak rzeczownik, *kłapać językiem* jak czasownik, *na wskroś* jak przysłówek itp. Wart zauważenia jest fakt, że wyrażenie pełniące funkcję określonej części mowy nie musi zawierać w sobie ani jednego członu, który tę część mowy realizuje: przykładowo *w nogi!* pełni funkcję czasownika (mimo braku komponentu werbalnego), a *ni to ni sio* – rzeczownika. Określenie części mowy jednostki złożonej ważne jest dla lingwistyki komputerowej, w której kwestie formalne odgrywają mniejszą rolę niż praktyczna użyteczność. Z tego powodu w słowniku komputerowym jednostek wielosegmentowych (zob. punkt 4) kategoria gramatyczna jest jedną z cech uwzględnianych przy opisie.

3.5. Jednostki wyższego rzędu

Fakt, że jednostki wielowyrazowe zachowują się pod wieloma względami jak pojedyncze wyrazy, oznacza między innymi, że mogą brać udział w tworzeniu jednostek złożonych. Pojedyncze leksemy *rada* i *minister* składają się na jednostkę *rada ministrów*. Z tak powstałej jednostki i leksemu *prezes* można otrzymać jednostkę wyższego rzędu: *prezes rady ministrów*. Proces taki może być iteracyjny: podany przykład można dalej rozszerzyć, otrzymując jednostkę *kancelaria prezesa rady ministrów*[40].

Ta obserwacja, wraz z relacją wyraz – jednostka wielosegmentowa, prowadzi do wniosku, że leksykon języka można opisać za pomocą modelu warstwowego. Pierwszą, podstawową warstwą leksykalną byłyby pojedyncze wyrazy, drugą – podstawowe (minimalne)

[40] Mogą pojawić się wątpliwości, czy podane w przykładzie jednostki nie są nazwami własnymi. Wydaje się wszakże, iż można w tym wypadku dokonać rozróżnienia: ogólne określenie organu administracyjnego, będące jednostką pospolitą, i konkretna instancja takiego organu w danym państwie, która reprezentowana jest w tekście przez odpowiednią nazwę własną.

jednostki wielosegmentowe, trzecią – jednostki złożone z kombinacji elementów warstwy pierwszej i drugiej itd.

Szczególnym przypadkiem połączenia jednostek wielowyrazowych jest sytuacja, w której dwie (potencjalne) jednostki nakładają się na siebie, np. *widmo śmierci głodowej*, *koniec świata mody*, *generał Armii Czerwonej* itp. Należy zauważyć, że tylko niektóre połączenia to kontaminacja dwóch rzeczywistych jednostek: spośród wymienionych przykładów tylko w pierwszym dochodzi do nałożenia się faktycznych jednostek wielosegmentowych (*widmo śmierci* i *śmierć głodowa*), skutkującego powstaniem jednostki wyższego rzędu, w pozostałych przykładach mamy do czynienia z regularnym połączeniem wyrazu i jednostki (*koniec* i *świat mody*) lub wyrazu i nazwy własnej (*generał* i *Armia Czerwona*).

Odróżnianie jednostek wyższego rzędu i luźnych połączeń zawierających jednostki można traktować jak przypadek homonimiczności, opisywanej wyżej (podpunkt 3.3). Jednak w przypadku połączeń typu *koniec świata mody* zachodzi konieczność wskazania, która część połączenia stanowi jednostkę, a która – wyraz (wyrazy). Ten problem określany jest w NLP terminem *bracketing* (co można by przetłumaczyć jako „nawiasowanie"). Ma to źródło w analogii do symboliki matematycznej, w której kolejność działań ustala się za pomocą nawiasów. Wspomniany wyżej przykład można interpretować na dwa sposoby: (*koniec świata*) *mody* lub *koniec* (*świata mody*) – zapis nawiasowy informuje o tym, którą część wyrażenia interpretuje się jako jednostkę. W omawianym przypadku poprawna będzie oczywiście druga interpretacja. Techniki automatycznego rozstrzygania problemu bracketingu (opisywane w Baldwin i Kim 2010, 25) opierają się na analizie frekwencji poszczególnych partii wyrażenia, analizie zależnościowej (patrz Rozdział I, podpunkt 5.2.3) bądź łączeniu obu tych podejść. Wydaje się jednak, że można wskazać również komplementarną metodę, wykorzystującą cechy językowe wyrażeń. Łatwo zauważyć, że składnikiem *generała Armii Czerwonej* jest nazwa własna i to właśnie ten fakt determinuje interpretację. W przypadku wyrażenia *koniec świata mody* mamy do czynienia z jednostkami, z których jedna (*świat mody*) ma tylko jedną, asumaryczną znaczeniowo, interpretację, natomiast druga (*koniec świata*) może być rozumiana na kilka rozmaitych sposobów, również na taki, w myśl którego stanowi doraźnie tworzoną regularną konstrukcję *koniec czegoś*. Podobna sytuacja zachodzi w przypadku wyrażenia *mieć kota domowego*, w którym potencjalnymi jednostkami są wyrażenia *mieć kota* (mogące być asumaryczną jednostką lub zwykłym połączeniem) i rzeczownik wielowyrazowy *kot domowy*. W obu przypadkach możliwość interpretacji jednej z jednostek jako swobodnej konstrukcji składniowej determinuje sposób interpretacji. Wstępna analiza zagadnienia pozwoliła na wyróżnienie hierarchii, która ułatwia odpowiedni podział wyrażeń. Hierarchia ta, wykorzystująca omówioną wyżej typologię jednostek, przedstawia się następująco:

1) nazwy własne,
2) wyrażenia nieregularne znaczeniowo (nieposiadające homonimu o znaczeniu dosłownym),
3) rzeczowniki wielowyrazowe,
4) wyrażenia nieregularne znaczeniowo (posiadające homonim o znaczeniu dosłownym),
5) pragmatemy**.**

Według tej reguły w analizowanym wyrażeniu za jednostkę uznawany byłby fragment, który w hierarchii ma wyższą pozycję. W przypadku, gdy wszystkie elementy należą do tej

samej kategorii, o interpretacji mogłyby decydować wspomniane wyżej techniki (frekwencja, analiza zależności).

Należy tu podkreślić, że przedstawiona koncepcja dotycząca bracketingu nie jest poparta systematycznymi badaniami – jest to zaledwie zarys koncepcji, niewątpliwie wymagający dalszych analiz i eksperymentów.

3.6. Wariantywność

Wiele nieciągłych jednostek leksykalnych posiada jedno- lub wielowyrazowe odpowiedniki. Istnienie wariantów wyrażenia może być konsekwencją reguł języka – np. kategoria aspektu w wyrażeniach werbalnych często jest przyczyną formułowania się dwóch form kształtowych różniących się czasownikiem, przy czym dokonany odpowiednik czasownika niedokonanego bywa nieregularny, jak w przypadku par *bije na alarm – uderzy na alarm*, *bije godzina – wybije godzin*a, *bije na głowę – pobije na głowę* itp. (por. Lewicki i Pajdzińska 2001, 317). Warianty mogą także powstawać przez wymianę któregoś segmentu wyrażenia na synonim lub wyraz bliskoznaczny, często nieco innej wartości ekspresywnej, np. *zakuta głowa* / *zakuty łeb*, *nie wystawiać* / *wyściubiać nosa*. Istnieją także warianty semantyczne – powstałe w wyniku ujęcia tej samej idei na różne sposoby, np. *mieć nie po kolei w głowie – mieć nierówno pod sufitem*. Warianty semantyczne mogą być także pojedynczymi wyrazami, w których źródłem opozycji jest metonimia (*cztery kółka* – ‘samochód’), metafora (*pies morski* – ‘foka’), peryfraza (*najlepszy przyjaciel człowieka* – ‘pies’), eufemizm (*mijać się z prawdą* – ‘kłamać’) albo relacja wyrażenie oficjalne – potoczne (*łódź motorowa* – ‘motorówka’). Wiedza o wariantach, ważna dla analizy semantycznej, jest uwzględniana w słowniku komputerowym opisywanym w punkcie 4.

3.7. Forma hasłowa

Przy tworzeniu leksykonu jednostek nieciągłych istotne jest ustalenie formy hasłowej wielowyrazowych wyrażeń. Forma podstawowa ważna jest także z punktu widzenia przetwarzania języka: służy najczęściej jako wartość kluczowa bazy danych, za pomocą której można odwoływać się do danego leksemu.

Ustalenie formy hasłowej jednostki nieciągłej nie sprowadza się zwykle do znalezienia form hasłowych składników wyrażenia. Jednostki często mają ustalone lub preferowane formy morfosyntaktyczne – pominięcie tego faktu prowadziłoby do powstawania form dewiacyjnych w rodzaju **wysoki czas* (zamiast *najwyższy czas*) czy **pies ogrodnik*. Ograniczenia paradygmatu fleksyjnego, blokada stopniowalności przymiotnika, defektywność liczby i inne nieregularności muszą być zatem brane pod uwagę w procesie tworzenia formy hasłowej. Przykładowo w rzeczownikach wielowyrazowych o postaci RZECZOWNIK – RZECZOWNIK W DOPEŁNIACZU (np. *rak piersi*, *mistrz świata*) forma hasłowa pierwszego członu to mianownik, drugiego – dopełniacz. W przypadku członów, które nie podlegają ograniczeniom, formę podstawową ustala się tak samo jak w przypadku zwykłego leksemu (por. Żmigrodzki 2009, 56-57).

Nie da się arbitralnie określić liczby charakteryzującej formę hasłową, gdyż zależy ona od konkretnego wyrażenia: w przytoczonych wyżej przykładach będzie to liczba pojedyncza, w przypadku *poszukiwacza skarbów* – pojedyncza dla pierwszego członu, mnoga dla drugiego, w *mistrzostwach świata* – mnoga dla pierwszego, pojedyncza dla drugiego,

w *godzinach przyjęć* – mnoga dla obu członów. Pewną trudność koncepcyjną sprawiają wyrażenia, które w ogromnej większości przypadków używane są w liczbie mnogiej, natomiast dopuszczalna jest forma operująca liczbą pojedynczą, np. *skrzydlate słowa*[41]. Według przyjętej zasady podstawową formą powinno być wyrażenie *skrzydlate słowo* (odstępstwo od konwencjonalnie przyjętej formy hasłowej powinno być motywowane absolutną niemożnością jej użycia) – co w rezultacie skutkuje wyrażeniem brzmiącym nieco niezręcznie. Jednak przyjęcie stanowiska przeciwnego ma dwie poważne wady: po pierwsze nie jest łatwo określić, jak dalece forma pluralna przeważa nad pojedynczą i ustalić granicę, której przekroczenie pozwalałoby na zmianę liczby w formie hasłowej na mnogą; po drugie zaburzona zostaje wtedy jasna konwencja, według której wyrażenie może być odmieniane przez liczbę, jeśli w formie hasłowej występuje w liczbie pojedynczej, w przeciwnym wypadku odmiana jest niemożliwa[42]. W związku z powyższym, w koncepcji słownika jednostek wielosegmentowych przedstawionej w niniejszej książce przyjęta jest pierwsza konwencja.

3.8. Defektywność kategorii gramatycznych

Jednostki nieciągłe niejednokrotnie cechowane są przez ograniczenia dotyczące jednego lub wielu aspektów gramatycznych, co – między innymi – odróżnia je od zwykłych połączeń składniowych. Często te ograniczenia wynikają z ich składni wewnętrznej – relacji między poszczególnymi członami wyrażenia, jak np. w przypadku frazy *pani domu* i innych realizujących schemat składniowy RZECZOWNIK + RZECZOWNIK W DOPEŁNIACZU. W wyrażeniach o takiej konstrukcji drugi człon ma ustaloną i niemożliwą do zmiany formę. W podpunkcie 2.2.3.6 opisane zostały przypadki ograniczeń innego typu, wynikające z „jednostkowości" – blokada stopniowalności przymiotnika, niemożność zmiany liczby czy pojedyncze przypadki ograniczeń dotyczących paradygmatu odmiany niewynikające z konstrukcji syntaktycznej (np. *dieta cud*). Ograniczenia w odmianie mogą dotyczyć także innych części mowy, np. liczebników (*trzy cztery*) lub czasowników (*umarł w butach*).

4. Komputerowy słownik nieciągłych jednostek leksykalnych

Pojęcie *słownik komputerowy* może być rozumiane na dwa sposoby. W pierwszym znaczeniu chodzi o tradycyjny słownik przystosowany do użytkowania przez człowieka, ale w wersji cyfrowej, jako program komputerowy czy serwis internetowy. W drugim – o zbiór danych językowych, zapisanych w odpowiedniej postaci, z których korzystać ma system komputerowy. W tej książce słownik komputerowy przywoływany jest wyłącznie w drugim znaczeniu[43].

Z uwagi na to, że słownik jest tworzony z myślą o komputerze, warto wskazać na pewne istotne cechy odróżniające go od słowników przeznaczonych do użytkowania przez człowieka.

Najistotniejszą różnicą jest brak praktycznych ograniczeń co do wielkości słownika. W leksykografii tradycyjnej restrykcyjna selekcja materiału stanowi istotny proces nie tylko

[41] Warto tu zauważyć, że występowanie któregoś z członów wyłącznie w liczbie mnogiej nie musi determinować liczby całego wyrażenia.

[42] Ewentualny brak formy mnogiej (jak np. w jednostce *czwarta władza*) jest konsekwencją ograniczeń semantycznych, a nie systemowych, zatem formalnie jest dopuszczalna – zob. Kosek (2008, 93).

[43] Więcej uwag na temat elektronicznych słowników kolokacji, w angielskiej wersji A*utomatic Collocation Dictionaries*, można znaleźć w pracy Pęzik (2014).

ze względów teoretycznych, ale również praktycznych. Ograniczona wielkość słownika narzuca zazwyczaj konieczność pominięcia części mniej istotnego materiału, skłania także do stosowania ostrzejszych kryteriów przy wyborze jednostek, które mają się znaleźć w słowniku. Słownik, którego adresatem jest komputer, nie ma takiego ograniczenia. Współczesna technika umożliwia łatwe magazynowanie i przetwarzanie nawet największego zbioru haseł.

Kolejna istotna różnica jest związana z brakiem jakiejkolwiek wewnętrznej kompetencji językowej komputera. Podczas gdy człowiek może opierać się na swojej – większej lub mniejszej – znajomości języka i rzeczywistości pozajęzykowej, dla komputera istnieje tylko to, co zostało do niego *explicite* wprowadzone.

Wymienione cechy powodują, że selekcja materiału do słownika komputerowego może i powinna być znacznie mniej restrykcyjna. Lepiej w tym przypadku zgrzeszyć nadmiarem niż zbytnią wstrzemięźliwością. Jednocześnie należy zauważyć, że w przypadku, gdy ma się do czynienia z maszyną, niezbędna jest duża precyzja i uwaga przy wprowadzaniu danych. Błędy logiczne czy nawet typograficzne są bezkrytycznie przez komputer przyjmowane i niejednokrotnie trudne później do zauważenia. Dlatego istotne jest, by struktura danych była względnie elastyczna, łatwa do dostosowywania i poprawiania błędów.

Właściwością odróżniającą słownik komputerowy od tradycyjnego jest też odmienność struktury – zarówno słownika jako całości, jak i pojedynczego hasła. W przypadku „ludzkiego" słownika frazeologicznego pojawia się problem ułożenia materiału w sposób, który z jednej strony oddaje semantyczne i składniowe relacje między jednostkami słownika, a z drugiej – ułatwia użytkownikowi odnalezienie odpowiedniej informacji. Bańko (2001, 173-177) szerzej omawia szczegóły tego nietrywialnego zagadnienia. W słowniku komputerowym ten problem jest właściwie nieistotny – lista haseł może być szeregowana w dowolny, arbitralnie ustalony sposób[44]. Również struktura hasła jest odmienna – w słowniku konwencjonalnym spotykana jest najczęściej postać: forma hasłowa wyrażenia + definicja i przykłady użycia; w słowniku komputerowym obok formy hasłowej niezbędna jest informacja o wszystkich formach danego wyrażenia, jego cechach składniowych itp.

Powstałe w ramach niniejszej książki, w oparciu o przedstawioną tu definicję jednostki nieciągłej, zasoby leksykalne: zbiór porównawczy (wspominany w podpunkcie 2.2.3.2, szerzej omówiony w Rozdziale IV) i listy jednostek wielosegmentowych, będące wynikiem działania algorytmu do ekstrakcji, stanowią zbiór, który utworzony został przede wszystkim dla wykorzystania go przy komputerowym przetwarzaniu języka polskiego. Jest to zatem materiał do słownika komputerowego.

[44] Nawet zupełnie przypadkowe ułożenie haseł jest w zasadzie możliwe, choć oczywiście zmniejszałoby to efektywność działania.

III
Automatyczna ekstrakcja nieciągłych jednostek leksykalnych

1. Potrzeba automatycznej ekstrakcji

Jak odnotowano w poprzednich rozdziałach, nieciągłe jednostki leksykalne stanowią istotny element języka, którego nie sposób pominąć bez narażania się na wymierne straty w jakości – zarówno w badaniach naukowych, jak i zastosowaniach praktycznych. Istniejące zasoby – takie jak słowniki związków frazeologicznych czy kolokacji (w znaczeniu typowych połączeń wyrazowych) – nie zapewniają jednak kompletnej reprezentacji zjawiska. Potrzebne jest zatem stworzenie wyczerpującej bazy jednostek nieciągłych. Jednakże ilość jednostek tego typu w języku naturalnym jest na tyle przytłaczająca, że ręczne tworzenie takiego zbioru byłoby zadaniem niezwykle czasochłonnym i kosztownym. Dużą pomocą w osiągnięciu celu mogą stać się systemy komputerowe. Automatyczne przetwarzanie tekstu pozwala na wyodrębnianie jednostek nieciągłych na skalę dużo szerszą, niż byłoby to realne bez pomocy komputerów.

Należy zaznaczyć, że maszynowa ekstrakcja nie jest pozbawiona wad. Przede wszystkim dokładność algorytmu komputerowego nie jest w stanie dorównać wynikom pracy specjalisty: jak w przypadku właściwie każdego systemu przetwarzania języka naturalnego, tak i tutaj program wyodrębniający popełnia błędy, zarówno poprzez zaliczenie zwykłych połączeń w poczet jednostek, jak i przez pominięcie jednostek faktycznych. Jednak zalety automatycznego systemu wyraźnie przewyższają jego wady, zwłaszcza jeśli wyniki jego działania potraktować jako punkt wyjścia dla dalszej pracy językoznawców (co jest najczęstszą praktyką w tej dziedzinie).

2. Metody automatycznej ekstrakcji

Ogólny schemat i najczęstsze techniki używane do ekstrakcji jednostek wielowyrazowych zostały przedstawione w punkcie 5.2. Rozdziału I. Rozmaite podejścia do ekstrakcji opisano tam z punktu widzenia historycznego. W tym miejscu przedstawione zostaną w sposób pogłębiony wybrane koncepcje, mające podstawowe znaczenie dla niniejszej książki.

Metody wyodrębniania jednostek można podzielić na dwie (szerzej omówione w następnym podpunkcie) grupy:

1) motywowane lingwistycznie,
2) motywowane statystycznie.

Można by oczekiwać, że w przypadku jednostek wielosegmentowych – pojęcia ściśle lingwistycznego – najlepiej sprawdzać się będzie podejście oparte na metodach językoznawczych. Okazuje się jednak, że skomplikowanie materii problemu idzie w parze z niewielką specyfiką formalną. Jak ujmuje to Lewicki (Wstęp do Lewicki i in. 1987, 9): „Nie istnieją żadne właściwości kształtowe, które by jednostkę frazeologiczną w tekście wyodrębniały i delimitowały". Rzeczywiście, jeżeli pominąć semantykę, jednostki wieloczłonowe są najczęściej nie do odróżnienia od swobodnych połączeń: *błędne koło* i *błędny wybór* czy *dać nogę* i *dać książkę* mają identyczną strukturę, jednak pierwsze z wymienionych są jednostkami, a drugie – zwykłymi połączeniami. Z tego powodu intuicyjne podejście oparte na systemie reguł, które pozwalałyby na odróżnienie jednostek wielosegmentowych od swobodnych połączeń, praktycznie nie jest stosowane. Warto przy tym zauważyć, że metody regułowe są często stosowane w dziedzinie pokrewnej: identyfikacji jednostek nazewniczych (*Named Entity Recognition*), wśród których dużą grupę stanowią właśnie jednostki wieloczłonowe. W ich przypadku łatwiej o charakterystyczne schematy, np. omawianą w Rozdziale II strukturę GRUPA NOMINALNA + CIĄG GRUP NOMINALNYCH W DOPEŁNIACZU, reprezentującą nazwy typu *Urząd Ubezpieczeń Zdrowotnych* (por. Piskorski 2004).

Powyższe zastrzeżenia nie oznaczają, że metody motywowane lingwistycznie są nieprzydatne w ekstrakcji: pełnią rolę pomocniczą, ale istotną. Metody statystyczne bez ich wsparcia borykają się z licznymi problemami – na przykład z faktem, że dużą część tekstu tworzą zazwyczaj wyrazy funkcyjne o niewielkiej lub wręcz zerowej wartości semantycznej, co mocno wpływa na poprawność wyników. Z tego powodu najskuteczniejsze i najczęściej stosowane jest podejście hybrydowe, łączące narzędzia z obu grup.

Proces automatycznej identyfikacji jednostek nieciągłych można podzielić na trzy ogólne etapy:

1) Przygotowanie tekstu do przetwarzania przez algorytm wyodrębniający. Rezultatem tego etapu jest tekst podzielony na segmenty (wyrazy, zdania itp.), które najczęściej są opisane morfosyntaktycznie;
2) Selekcja kandydatów na jednostki leksykalne. Na tym etapie dokonywana jest wstępna ocena sekwencji wyrazowych i wybierane są te, które spełniają odpowiednie kryteria (np. realizują określony schemat składniowy);
3) Ocena i sortowanie kandydatów. Każdy z kandydatów z poprzedniej fazy jest oceniany za pomocą odpowiednich miar statystycznych (lub ich kombinacji). Lista ocenionych wyrażeń jest następnie sortowana w porządku od najlepiej do najgorzej ocenionych.

Na każdym z tych etapów stosuje się odpowiednie narzędzia: w dwóch pierwszych lingwistyczne, w trzecim statystyczne[45].

2.1. Narzędzia językowe

Tekst, na którym pracować będzie program do ekstrakcji, powinien być wcześniej odpowiednio przygotowany. Istnieje kilka poziomów wstępnego przetwarzania (ang. *pre-pro-*

[45] W procesie selekcji kandydatów stosowany jest często również filtr frekwencyjny, który powoduje odrzucenie wyrażeń, których liczba wystąpień w korpusie jest mniejsza od arbitralnie ustalonego progu. Nie jest to oczywiście reguła o motywacji lingwistycznej – wynika ze specyfiki analizy statystycznej, trudno jednak określić ją mianem „narzędzia statystycznego".

cessing) tekstu, za które odpowiedzialne są najczęściej odrębne moduły. Poniżej omówione zostaną poszczególne poziomy, poczynając od najbardziej podstawowych.

2.1.1. Podział na segmenty (tokenizacja)

Segmentacja jest procesem, w trakcie którego tekst, traktowany jako jeden (najczęściej długi) ciąg znaków, zostaje podzielony na mniejsze elementy. Takimi elementami mogą być m.in. akapity, zdania, frazy i – przede wszystkim – wyrazy. O ile wyróżnianie segmentów złożonych może być pomocne, ale nie jest konieczne, o tyle bez podziału na wyrazy trudno sobie wyobrazić jakiekolwiek automatyczne przetwarzanie tekstu.

Proces dzielenia ciągu znaków na wyrazy nie jest wbrew pozorom zadaniem trywialnym. Znakiem delimitującym wyraz jest najczęściej spacja, ale może to być również większość znaków interpunkcyjnych. Należy zdecydować, jak będą traktowane ciągi zawierające znaki nieliterowe (np. apostrof lub dywiz, mogący być łącznikiem między dwoma wyrazami albo znakiem przeniesienia wyrazu do nowej linii). Bardziej zaawansowana segmentacja może dzielić wyrazy złożone na segmenty podstawowe (np. ciąg *Tom* w zdaniu *Tom się ubawił* jest w rzeczywistości złożeniem leksemu *to* i formy aglutynacyjnej leksemu *być*); przykłady praktycznych rozwiązań tego typu kwestii można znaleźć w pracy omawiającej tworzenie Narodowego Korpusu Języka Polskiego (Przepiórkowski i in. 2012, 61). Osobnym problemem, który należy rozwiązać, jest kwestia niestarannej, niepoprawnej pisowni (np. brak spacji w ciągu *ul.Szeroka* powoduje, że ciąg taki według zwykłych reguł ortograficznych powinien stanowić jeden wyraz).

2.1.2. Lematyzacja i dezambiguacja

Lematyzacja (ang. *lemmatisation* lub *lemmatization*), zwana niekiedy hasłowaniem, polega na przypisaniu formie wyrazowej napotkanej w przetwarzanym tekście leksemu (reprezentowanego przez formę podstawową – lemat), którego ta forma jest reprezentantem. Jest to proces – z punktu widzenia systemu automatycznego – dosyć złożony. Składa się z dwóch zasadniczych etapów: 1) wyszukanie wszystkich leksemów, które mogą przyjąć badaną formę, 2) identyfikacja właściwego leksemu. Etap pierwszy jest zbliżony do tzw. stemmingu (od ang. *stem* – ‘rdzeń’), automatycznego wyszukiwania tematu słowotwórczego formy wyrazowej. Istnieją różne podejścia do stemmingu, jednak najpopularniejszą metodą jest usuwanie sufiksów wyrazu przy wykorzystaniu szeregu lingwistycznie motywowanych reguł. Takie podejście stosuje m.in. stemmer Portera (Porter 1980) – jeden z najszerzej wykorzystywanych algorytmów tego typu dotyczących języka angielskiego. Przykładowo, w przypadku wyrazów *pisać i pisarz* stemmer powinien odrzucić morfemy *-ać* i *-arz* i zidentyfikować *pis-* jako rdzeń obu wyrazów.

Ustalenie tematu wyrazu najczęściej jednak nie wystarcza, aby znaleźć odpowiedni lemat. Lematyzacja wykorzystuje zatem inne dane o wyrazie, np. jego część mowy (jej ustalenie stanowi osobne wyzwanie – zob. następny podpunkt), cechy morfologiczne, informacje o nieregularności (takiej jak *dobry-lepszy*) itp.

Lematyzacja pozwala na ustalenie formy hasłowej badanego wyrazu, ale w przypadku (częstym, zwłaszcza w języku polskim), gdy daną formę może przyjmować więcej niż jeden leksem, należy również ustalić, o który z potencjalnych leksemów chodzi

w danym przypadku. Ten etap nazywany jest ujednoznacznianiem bądź dezambiguacją (ang. *disambiguation*)[46].

W procesie ekstrakcji jednostek wielosegmentowych lematyzacja i dezambiguacja odgrywają istotną rolę: pozwalają na zliczanie wystąpień wyrażenia niezależnie od tego, w jakiej formie jego człony pojawiają się w tekście. Przykładowo, napotkane sekwencje wyrazów *lwa salonowego* i *lwem salonowym* zostaną dzięki temu potraktowane jako dwukrotna realizacja jednostki *lew salonowy*. Należy tu zauważyć, że stosowanie lematyzacji ma też pewne wady: po pierwsze łączy się często z utratą danych (konkretna forma tekstowa jest „zapominana", pozostaje tylko lemat), po drugie niektóre jednostki wieloczłonowe występują tylko w określonych, niemożliwych do zmiany formach, zatem zliczanie wszystkich form, które mają określony lemat, prowadzi do utraty dokładności. Za przykład może tu służyć niemała grupa wyrażeń wielosegmentowych, będących formalnie połączeniem rzeczownika i przymiotnika: w języku polskim takie konstrukcje połączone są związkiem zgody, zatem przymiotnik może występować wyłącznie w formie zgodnej z rodzajem rzeczownika, który określa[47]. Mimo tych zastrzeżeń zdecydowana większość systemów automatycznej ekstrakcji preferuje – o ile dostępne są odpowiednie narzędzia – działanie na lematach.

2.1.3. Znakowanie morfosyntaktyczne (tagowanie)

Lematyzacja jest szczególnym przypadkiem procesu ogólniejszego, polegającego na przypisaniu danej formie tekstowej jej cech morfosyntaktycznych, takich jak część mowy, cechy fleksyjne itp. Wymaga to zastosowania odpowiedniego narzędzia zwanego tagerem, dostosowanego oczywiście do konkretnego języka.

Z punktu widzenia ekstrakcji jednostek znakowanie morfosyntaktyczne oddaje istotne usługi na drugim etapie ekstrakcji: selekcji kandydatów. Dzięki informacji o części mowy możliwe jest wyszukiwanie wyłącznie wyrażeń spełniających określone schematy części mowy, co wydatnie poprawia jakość ekstrakcji (zob. punkt 5.2.3. Rozdziału I).

Ideę schematów opierających się na częściach mowy można rozwinąć w oparciu o dalsze informacje uzyskane w trakcie tagowania. Oparcie wzorców na dodatkowych danych takich jak przypadek, liczba, rodzaj czy innych kategorii gramatycznych pozwala na stworzenie schematów syntaktycznych, dzięki którym możliwe jest tzw. płytkie parsowanie (chunking) – wyodrębnianie z tekstu niewielkich wyrażeń, które pozostają ze sobą w określonej relacji składniowej, np. czasownik-dopełnienie, czy rzeczownik-przydawka. Pozwala to na jeszcze dokładniejszą selekcję kandydatów na jednostki.

Rezultatem drugiego etapu wyodrębniania jednostek (selekcji kandydatów) jest lista wszystkich wyrażeń, które – z formalnego punktu widzenia – mogą stanowić jednostki. Selekcja taka najczęściej polega na wyszukiwaniu sekwencji wyrazów powiązanych składniowo (za pomocą opisywanych wyżej schematów składniowych), zastosowaniu filtra frekwencyjnego i – ewentualnie – dodatkowych zasad heurystycznych (np. odrzucenie wyrażeń, których człony znajdują się na liście wyrazów niepożądanych – tzw. stop-liście).

[46] Dezambiguacja jest niekiedy uznawana za część lematyzacji.

[47] Przypadki, gdy w tekście pojawią się obok siebie rzeczownik i przymiotnik niepołączone związkiem zgody są na tyle rzadkie, że nie odgrywają istotnej roli, jednak wiele miar asocjacji bierze pod uwagę oprócz frekwencji samego wyrażenia również frekwencję jego członów. Z tego powodu zliczanie wszystkich przypadków użycia leksemu *biały*, gdy analizowane jest połączenie wyrazów *biały kruk*, daje zawyżony wynik (liczone są również formy *biała*, *białe* itp.).

2.2. Narzędzia statystyczne – miary asocjacji

Istnieje szereg technik statystycznych, które można wykorzystać przy ekstrakcji jednostek wielosegmentowych: miary asocjacji (wykorzystujące frekwencję wyrazów), miary restrykcji leksykalnych i składniowych, metody oparte na ocenie nieregularności znaczeniowej wyrażenia lub porównujące tłumaczenia wyrażeń (opisane są w Rozdziale I). Spośród wymienionych metod, najefektywniejsza i najczęściej używana jest pierwsza z nich: miary asocjacji. Poniższa sekcja opisuje szczegółowo to zagadnienie, jako że jest ono wykorzystywane przez prezentowany w książce algorytm ekstrakcji.

Wielosegmentowe jednostki leksykalne charakteryzuje bardzo ważna – z punktu widzenia automatycznego wyodrębniania – cecha: występują w tekście w takiej samej (lub bardzo zbliżonej) postaci[48]. Ta cecha, którą można nazwać powtarzalnością, połączona z faktem, że jednostki wielowyrazowe są często wykorzystywane przy tworzeniu wypowiedzi, powoduje, że w odpowiednio dużym zbiorze tekstów występują one statystycznie częściej niż wynikałoby to z czystego przypadku. Ta obserwacja jest fundamentalna dla procesu automatycznego wyodrębniania jednostek: istnieje wiele statystycznych testów, zwanych miarami asocjacji, które pozwalają na wychwycenie sekwencji wyrażeń, których częstość współwystępowania sugeruje, że są one od siebie zależne.

Trzeci etap ekstrakcji, a więc stosowanie metod statystycznych (w szczególności miar asocjacji) wygląda następująco: dla każdego kandydata na jednostkę (z listy utworzonej w etapie drugim) wyliczana jest odpowiednia wartość miary. Posortowanie takiej listy w kolejności od sekwencji, która uzyskała najwyższy wynik, daje w rezultacie ranking wyrażeń, w którym na szczycie znajdują się połączenia wyrazowe o największym prawdopodobieństwie (z punktu widzenia danej miary) tego, że stanowią jednostkę leksykalną. Prawdopodobieństwo to zmniejsza się w miarę schodzenia na dół listy. Ostateczna selekcja wyrażeń uznawanych przez algorytm za jednostki polega na odrzuceniu wszystkich, dla których wartość nie przekracza ustalonego (wyznaczonego eksperymentalnie) progu. Warto zauważyć, że takie rozumowanie dobrze koresponduje z definicją wielosegmentowej jednostki leksykalnej, przyjętą w książce (zob. punkt 2 Rozdziału II).

2.2.1. Wstępne założenia

Badając język za pomocą metod statystycznych, należy pamiętać, że nie dysponujemy rzeczywistym modelem języka, a jedynie korpusem, który jest do tego modelu zbliżony. Korpus taki można traktować jako losową próbę uzyskaną z nieskończonej populacji odpowiadającej danemu językowi (więcej na ten temat Evert 2005).

Ilekroć mowa jest o wystąpieniu wyrazu w korpusie, chodzi o wystąpienie leksemu (reprezentowanego przez lemat), nie zaś konkretnej formy tekstowej. Jako że jednym z celów tej książki jest stworzenie algorytmu ekstrahującego wielosegmentowe jednostki leksykalne, a nie wyszukującego tekstowe realizacje tych jednostek, każdy wyraz tekstowy poddawany jest lematyzacji.

Frekwencja (liczba wystąpień w korpusie) leksemu u wyrażona jest symbolem $f(u)$.

Miary asocjacji zaprojektowane są do badania korelacji między dwoma elementami (w tym przypadku wyrazami), zatem rozumowanie przedstawione poniżej zakłada, że brane

[48] Kwestię różnic wynikających z faktycznej formy gramatycznej rozwiązuje lematyzacja.

są pod uwagę wyłącznie pary wyrazów – bigramy[49], np. *masło roślinne* czy *głos sumienia*. W niektórych przypadkach możliwe jest stosunkowo proste uogólnienie miar tak, by można je było zastosować do dłuższych sekwencji, np. sekwencji trzech wyrazów – trigramów (takich jak *niezidentyfikowany obiekt latający*), lecz metody takie należy uznać za heurystyczne (zob. Ramisch 2012, 47).

W niniejszej książce przyjęto, że bigramy (i ogólniej: n-gramy) to sekwencje wyrazów następujące bezpośrednio po sobie. Nie jest to założenie oczywiste: w zdaniu *Wilk morski miał już wtedy mocno w czubie* wyrazy składowe jednostki *wilk morski* sąsiadują ze sobą, ale już człony wyrażenia *mieć w czubie* są od siebie oddalone. Evert (2005, 18-20) wyróżnia dwa typy współwystępowania: pozycyjny, w którym elementy brane pod uwagę znajdują się w ustalonej odległości od siebie (w szczególności odległość może być zerowa, co oznacza, że elementy bezpośrednio ze sobą sąsiadują) bądź w obrębie określonej jednostki tekstu (np. zdania lub akapitu), i relacyjny, gdzie elementy znajdują się w określonej relacji (najczęściej składniowej). Przykład relacyjnego podejścia można znaleźć w (Seretan 2011b). Oba podejścia mają zalety i wady. Model pozycyjny jest dużo prostszy koncepcyjnie i obliczeniowo, a fakt, że większość jednostek wielosegmentowych występuje w tekście w bezpośredniej odległości, pozwala na wyodrębnienie większości interesującego materiału. Z drugiej strony nie jest w stanie wykryć bardziej złożonych relacji (dotyczy to zwłaszcza wyrażeń o funkcji czasownikowej – tak jak w podanym wyżej przykładzie *mieć w czubie*). Model relacyjny z kolei potrafi wychwycić więcej zależności (jest więc ogólniejszy), n-gramy uzyskane w ten sposób na pewno łączy bezpośrednia relacja składniowa, co niekoniecznie zachodzi pomiędzy wyrazami sąsiadującymi ze sobą. Wadami tego podejścia jest konieczność wykorzystania dodatkowego narzędzia, co po pierwsze zwiększa stopień złożoności całego systemu (i wydłuża – często istotnie – czas jego działania), po drugie jest obciążone nieuniknionymi błędami parsowania. Niniejsza książka pozostaje przy najprostszym, pozycyjnym modelu, który w połączeniu z dodatkowymi lingwistycznymi ograniczeniami (zob. Rozdział IV) jest jednocześnie funkcjonalny i skuteczny. Architektura przedstawionego tam systemu ekstrakcji nie wyklucza przy tym zastosowania modelu relacyjnego.

2.2.2. Testowanie hipotezy

Sposobem na określenie, czy dany bigram pojawia się w korpusie częściej, niż wynikałoby to z przypadku, jest metoda testowania hipotezy. Polega ona na przyjęciu tzw. hipotezy zerowej (oznaczanej jako H_0), zakładającej, że żadna korelacja między komponentami bigramu nie występuje, a następnie ocenie, na ile hipoteza taka jest prawdopodobna w odniesieniu do obserwowanych danych (uzyskanych z korpusu). Jeżeli prawdopodobieństwo to jest mniejsze od ustalonego progu (nazywanego poziomem istotności), hipotezę zerową można odrzucić, przyjmując tym samym, że człony badanego wyrażenia są powiązane. Typowymi wartościami poziomu istotności są 0.05, 0.01, 0.005. W przypadku pierwszego z progów oznacza to, że błędne odrzucenie hipotezy zerowej (a więc niesłuszne uznanie, że człony badanego wyrażenia są powiązane) zdarza się nie częściej niż w 5 przypadkach na sto, w przypadku drugiego – nie częściej niż raz na sto itd.

Błędne odrzucenie hipotezy zerowej (gdy mamy do czynienia z wynikiem fałszywie pozytywnym) stanowi jeden z dwóch typów błędów, które mogą wystąpić w procesie te-

[49] Z tego względu lwia część prac poświęconych ekstrakcji wyrażeń wielowyrazowych dotyczy wyłącznie bigramów.

stowania hipotezy. Drugi typ to nieodrzucenie hipotezy zerowej, choć ta jest faktycznie fałszywa (wynik fałszywie negatywny). Zwyczajowo mówi się o błędach odpowiednio typu I i II. Błędy typu pierwszego wpływają na dokładność ekstrakcji jednostek wielosegmentowych (proporcję wyników pozytywnych – to jest wykrytych przez algorytm ekstrahujący faktycznych jednostek wielosegmentowych – w stosunku do wszystkich wyników uznanych przez algorytm za pozytywne). Błędy typu drugiego wpływają na jej kompletność (stosunek wyników pozytywnych, które udało się wykryć algorytmowi, i wszystkich wyników pozytywnych w korpusie)[50]. Manipulacja poziomem istotności powoduje zmianę stosunku dokładności i kompletności: jeśli poziom istotności zwiększymy, dokładność się zmniejszy, a kompletność wzrośnie – i na odwrót (szersze omówienie tych kwestii i ich konsekwencji znajduje się w Rozdziale IV).

W przypadku rozważanego problemu – zależności między wyrazami – hipotezą zerową jest założenie, że wyrazy występują od siebie niezależnie. Jeżeli weźmiemy pod uwagę dwa wyrazy *u* i *v* i ich prawdopodobieństwo wystąpienia w korpusie *P* (*u*) i *P* (*v*), to hipoteza zerowa oznacza, że spełniają one następującą zależność:

$$P(uv) = P(u)\,P(v)$$

gdzie *P (uv)* oznacza prawdopodobieństwo wystąpienia bigramu składającego się z wyrazów *u* i *v*.

Należy zauważyć, że w niniejszej książce standardowa procedura testowania hipotezy nie jest faktycznie stosowana. Właściwie każda para wyrazów sąsiadujących ze sobą w obrębie jednej frazy jest od siebie zależna – jeśli zatem wszystkie wyrażenia, dla których hipotezę zerową udałoby się odrzucić, uznawane byłyby za jednostki leksykalne – niemal każde badane wyrażenie trzeba by zaliczyć w poczet jednostek. Z tego względu stosuje się nieco inne podejście: wartości liczbowe uzyskane dla wyrażenia na podstawie statysk testowych wykorzystywane są do porównywania wyrażeń między sobą. Zakłada się, że bigram, który otrzymał wyższą punktację, ma większą szansę na bycie jednostką niż bigram z punktacją niższą.

2.2.3. Tablice dwudzielcze

Pomocnym narzędziem przy obliczaniu miar asocjacji dla par wyrazów są tzw. tablice dwudzielcze (ang. *contingency tables*). Zestawiają one (w formie macierzy o wymiarach 2 x 2) dane frekwencyjne dotyczące interesującego bigramu (kandydata na jednostkę) i obu jego członów.

Na korpus można patrzyć jako na rezultat wielokrotnie powtarzanych prób polegających na losowaniu z nieznanej populacji (języka) par wyrazów (leksemów). Korpus składa się zatem z par zmiennych losowych *U* i *V*, z których pierwsza odpowiada pierwszemu członowi bigramu, druga – drugiemu. Dla par wyrazów *u* i *v* zmienna *U* przyjmuje wartość 1, gdy w próbie na pierwszym miejscu wystąpił wyraz *u*, i wartość 0, gdy wyraz *u* n i e wystąpił.

[50] Dokładność i kompletność to tłumaczenie angielskich terminów *precision* i *recall*, oznaczających dwie miary mierzące jakość ekstrakcji. Używane powszechnie w dziedzinie wyszukiwania informacji (*information retrieval*) stały się również standardowymi miarami ewaluacji w NLP. W literaturze polskojęzycznej spotyka się również inne tłumaczenie *recall*: przywołanie – jednak wydaje się, że „kompletność" jest intuicyjnie bardziej zrozumiała, a jej semantyczne konotacje lepiej oddają istotę pojęcia.

Zmienna V przyjmuje analogiczne wartości, dla wyrazu v na pozycji drugiej.

Dla każdej pary wyrazów u, v można utworzyć tablicę dwudzielczą wartości obserwowanych w korpusie w następujący sposób:

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$R_1 (u \cdot)$
$U \neq u$	O_{21}	O_{22}	$R_2 (\neg u \cdot)$
	$C_1 (v \cdot)$	$C_2 (\neg v \cdot)$	N

Tabela III.1: Tablica dwudzielcza dla wartości obserwowanych bigramu uv

Zawartości poszczególnych komórek oznaczają frekwencje w korpusie odpowiedniej kombinacji U i V:

$O_{11} - f(uv)$, frekwencja bigramów składających się z u i v, nazywana frekwencją współwystępowania lub frekwencją wspólną

$O_{12} - f(u\neg v)$, frekwencja bigramów, których pierwszym członem jest u, a drugim – dowolny wyraz oprócz v

$O_{21} - f(\neg uv)$, frekwencja bigramów, których pierwszy człon stanowi dowolny wyraz z wyjątkiem u, a drugi to v

$O_{22} - f(\neg u\neg v)$, frekwencja bigramów, w których pierwszy człon to nie u, a drugi to nie v

Wartości R_1 i C_1 to tzw. wartości brzegowe (ponieważ znajdują się na brzegach tabeli), których wielkość jest równa ilości wystąpień w korpusie bigramów, w których pierwszym wyrazem jest u, a drugim dowolny wyraz (w przypadku R_1) lub pierwszy wyraz jest dowolny, a drugim wyrazem jest v (w przypadku C_1). Jak łatwo zauważyć, $R_1 = O_{11} + O_{12}$, a $C_1 = O_{11} + O_{21}$.

Wartość N oznacza ilość wszystkich bigramów w korpusie i jest równa $O_{11} + O_{12} + O_{21} + O_{22}$.

Poniższy przykład przedstawia tablicę dwudzielczą dla rzeczywistych danych, konkretnie dla wyrażenia *piłka nożna* w korpusie *sample* (podkorpusie korpusu IPI PAN).

	$V =$ *nożny*	$V \neq$ *nożny*	
$U =$ *piłka*	312	3167	3479
$U \neq$ *piłka*	310	12356471	12356781
	622	12359638	12360260

Tabela III.2: Tablica dwudzielcza dla wartości obserwowanych bigramu „piłka nożna”

Aby móc dokonać testowania hipotezy, należy – oprócz przedstawionych wyżej wartości obserwowanych w próbie – znać również wartości oczekiwane, których należałoby się spodziewać, jeśli hipoteza zerowa byłaby prawdziwa. Oczekiwana wartość wspólna (oznaczona jako E_{11}) dla bigramu uv to (zgodnie z hipotezą zerową):

$$E_{11} = P(u)\,P(v)\,N = \frac{f(u)}{N}\,\frac{f(v)}{N}\,N = \frac{f(u)\,f(v)}{N}.$$

Wartości oczekiwane można zgromadzić w tablicy dwudzielczej, analogicznie jak w przypadku wartości obserwowanych. Ogólny wzór pozwalający obliczyć wartość każdej komórki w tabeli jest następujący:

$$E_{ij} = \frac{R_i C_j}{N},$$

dla $i, j \in \{1,2\}$

R_i oznacza sumę komórek w *i-tym* wierszu tablicy dwudzielczej zawierającej wartości obserwowane, C_j – sumę komórek w *j-tej* kolumnie tabeli. Tablica dwudzielcza wartości oczekiwanych bigramu *uv* wygląda następująco:

	$V = v$	$V \neq v$
$U = u$	$\frac{R_1 C_1}{N}$	$\frac{R_1 C_2}{N}$
$U \neq u$	$\frac{R_2 C_1}{N}$	$\frac{R_2 C_2}{N}$

Tabela III.3: Tablica dwudzielcza dla wartości oczekiwanych bigramu uv

2.2.4. Zastrzeżenia dotyczące miar asocjacji

Część miar asocjacji stosowanych przy ekstrakcji wyrażeń wielowyrazowych zakłada, że dane w populacji podlegają rozkładowi normalnemu. Rozkład normalny charakteryzuje się tym, że przeważająca część danych ma wartość zbliżoną do średniej w danej populacji. Wartości te układają się na wykresie w tzw. krzywą dzwonową (zob. Wykres 1). Rozkładowi normalnemu podlega wiele zagadnień (np. wzrost w grupie ludzi), jednak w przypadku języka naturalnego jest inaczej. Rozkład normalny zakłada, że większość danych w niewielkim stopniu różni się od średniej danej populacji, a wartości skrajne zdarzają się rzadko. Tymczasem, jeśli elementy leksykonu języka naturalnego ułożymy w kolejności od najczęściej występującego do najrzadszego, okaże się, że ich frekwencja jest odwrotnie proporcjonalna do pozycji w otrzymanym rankingu. Zatem najpopularniejszy wyraz będzie występował mniej więcej dwukrotnie częściej niż następny na liście. Zależność tę opisuje prawo Zipfa (nazwane od nazwiska odkrywcy, G. Zipfa):

$$f(k;N,s) = \frac{1/k^s}{\sum_{n=1}^{N} (1/n^s)},$$

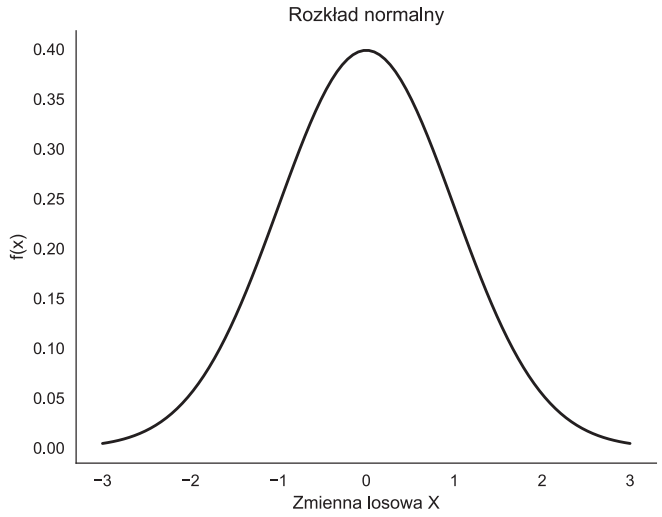
gdzie k oznacza pozycję na liście rankingowej, N – rozmiar korpusu, s – parametr danego rozkładu.

Rozkład statystyczny spełniający powyższą zależność nazywany jest rozkładem Zipfa.

Ponieważ część miar asocjacji zakłada rozkład normalny danych, wydawać się może, że stosowanie ich do ekstrakcji n-gramów (podlegających prawu Zipfa) prowadzić będzie do błędnych wyników. W rzeczywistości miary takie z powodzeniem były i są wykorzy-

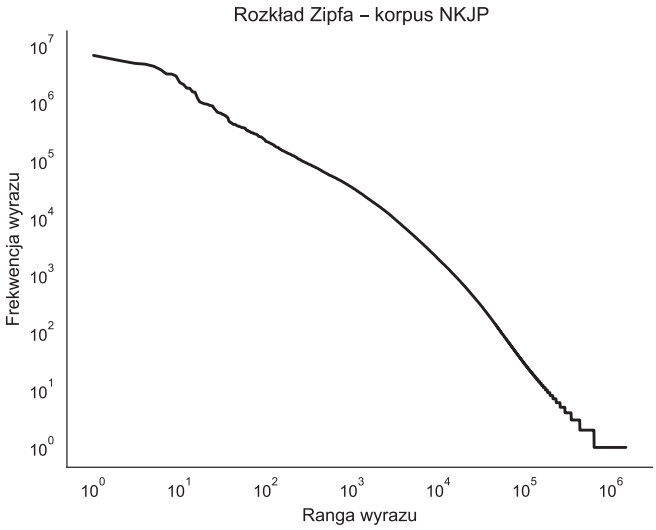
stywane. Jest to konsekwencją faktu, że – jak już wspomniano wyżej – metoda testowania hipotezy w przypadku tego zagadnienia stanowi tylko punkt wyjścia: w praktyce najczęściej pomija się poziom istotności (a więc to, czy hipotezę zerową można odrzucić), skupiając się na punktacji, którą badany kandydat na jednostkę otrzymał dla danej miary, i porównując ten wynik z wynikami pozostałych kandydatów. Rezultatem takiego podejścia jest lista rankingowa, na której szczycie znajdują się n-gramy, które zdobyły najwyższą punktację (w domyśle: mają największe prawdopodobieństwo okazać się rzeczywistymi jednostkami). Okazuje się, że nieścisłości wynikające z niespełnienia założenia o rozkładzie normalnym nie wpływają w bardzo istotny sposób na jakość procesu wyodrębniania. Niemniej, należy pamiętać, że matematyczne fundamenty stosowania niektórych miar są chwiejne.

Rozkład Zipfa ma to do siebie, że na dużą część populacji składają się elementy występujące bardzo rzadko. Oznacza to, że zauważalną część korpusu tekstów tworzą wyrazy (lub jednostki wielowyrazowe), które występują w nim tylko raz lub dwa. Jednostki takie (zwane odpowiednio hapaks i dis legomena[51]) stanowią problem z punktu widzenia analizy statystycznej: ich frekwencja w korpusie jest zbyt niska, by na jej podstawie można było otrzymać wiarygodne wyniki. Dlatego też najczęściej przy selekcji kandydatów odrzuca się wszystkie, których frekwencja nie przekracza określonego progu. Stosowane są różne progi, np. $f \geqslant 5, f \geqslant 10, f \geqslant 30$. Evert (2004, 133) wskazuje, że – jakkolwiek hapaks i dis legomena powinny być zawsze wykluczane z analizy statystycznej – próg 5 wystąpień w korpusie jest wystarczający do osiągnięcia wiarygodnych wyników. Zasada ta (nazwana filtrem frekwencyjnym) była omawiana wcześniej jako część procesu selekcji kandydatów. Poniższe wykresy zestawiają graficzną ilustrację rozkładu normalnego i rozkładu Zipfa.



Wykres 1: Rozkład normalny (tzw. standardowy, gdzie średnia = 0, wariancja = 1)

[51] Z greckiego: *hapax* – ‘jeden raz’, *di* – ‘dwa, dwu’ i *logomenon* (pl. *logomena*) – ‘rzecz mówiona’.



Wykres 2: Rozkład Zipfa dla podkorpusu sample korpusu IPI PAN. Oś x przedstawia miejsce wyrazu na liście, oś y – frekwencję wyrazu. Wykres jest sporządzony w układzie logarytmicznym, by wyniki były czytelniejsze

2.2.5. Najpopularniejsze miary asocjacji

Możliwych do wykorzystania przy ekstrakcji jednostek wielosegmentowych miar asocjacji jest dużo (Pecina i Schlesinger 2006 wymieniają ich 82). W tym miejscu omówione zostaną najczęściej wykorzystywane i najskuteczniejsze z nich. Podane poniżej techniki i wzory mogą wydawać się suche i trudne do zastosowania w praktyce, dlatego czytelnikowi zainsteresowanemu samodzielnym stosowaniem miar asocjacji polecam przejrzenie Dodatku C, w którym cała procedura obliczeniowa zilustrowana jest przykładami.

2.2.5.1. Statystyka testowa t Studenta (t-score)

Jest to bardzo popularna miara, opierająca się na porównaniu średniej z próby (danych korpusowych) i średniej z populacji, przy założeniu (hipotezie zerowej), że próba została wzięta z populacji, i skalowaniu za pomocą błędu standardowego średniej. Otrzymujemy w ten sposób statystykę, która pokazuje, w jakim stopniu próba różni się od populacji, a co za tym idzie – jak duże jest prawdopodobieństwo, że hipoteza zerowa jest niespełniona. Jest to wyrażone poniższym wzorem:

$$t - score = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}},$$

gdzie $\bar{x}$ oznacza średnią z próby, μ – średnią z populacji, s^2 – wariancję próby, N – liczność próby. Aby móc określić wartości $\bar{x}$, μ i s^2 zazwyczaj stosuje się następujące rozumowanie: przyjmujemy (jak było to opisywane wyżej), że korpus jest rezultatem N prób polegających na losowaniu par wyrazów. Weźmy pod uwagę wyrazy u i v, tworzące rozpatrywany bigram. Prawdopodobieństwo wylosowania pary $(u, \cdot)$, a więc bigramu, w którym pierwszym

elementem jest *u*, drugim – dowolny wyraz, można obliczyć, używając estymatora największej wiarygodności:

$$P(u) = \frac{f(u,\cdot)}{N} = \frac{R_1}{N},$$

analogicznie dla pary (·, *v*). Jeżeli każdemu zdarzeniu polegającemu na wylosowaniu pary wyrazów z populacji przypiszemy wartość 1 w przypadku, gdy zostanie wylosowana para (*u*, *v*), i 0, gdy wylosowana zostanie dowolna inna kombinacja, proces taki można utożsamić z próbą Bernoulliego. Otrzymujemy w ten sposób dystrybucję zero-jedynkową, dla której średnia jest równa prawdopodobieństwu sukcesu w jednej próbie oznaczanemu jako *p*, a wariancja wynosi *p(1 – p)*. Dla całej populacji średnia μ przybliżona za pomocą estymatora największej wiarygodności wynosi (zgodnie z hipotezą zerową):

$$\mu = p_1 = P(u,\cdot)\, P(\cdot,v).$$

Dla badanej próby prawdopodobieństwo p_2, a zarazem średnią $\bar{x}$ można wyliczyć na podstawie obserwowanej frekwencji *uv:*

$$p_2 = \bar{x} = f(uv) = \frac{O_{11}}{N}.$$

Wariancja s^2 próby, równa *p(1 – p)*, jest w przybliżeniu równa *p* dla małych wartości *p* (co jest zazwyczaj spełnione, jeśli korpus jest odpowiednio duży):

$$s^2 = p_2\,(1 - p_2) \approx p_2 = \frac{O_{11}}{N}.$$

W rezultacie otrzymujemy formułę:

$$t - score = \frac{\frac{O_{11}}{N} - \frac{R_1}{N}\frac{C_1}{N}}{\sqrt{\frac{O_{11}}{N^2}}} = \frac{\left(\frac{O_{11}}{N} - \frac{R_1}{N}\frac{C_1}{N}\right)N}{\sqrt{O_{11}}} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}.$$

Wskaźnik t przy całej swej popularności bywa krytykowany. Evert (2005, 82-83) kwestionuje zasadność użycia testu, ze względu na nieprawidłowe założenie: t-score wymaga rozkładu normalnego; w trakcie obliczania wariancji próby stosuje się jako przybliżenie rozkładu normalnego rozkład zero-jedynkowy, co jest działaniem mocno wątpliwym.

2.2.5.2. Statystyka testowa z (z-score)

Jest to jedna z najpopularniejszych statystyk testowych mierząca oddalenie obserwowanej wartości *x* od średniej populacji μ (mierzonej za pomocą odchylenia standardowego σ), co można widzieć jako stopień niepewności, że próba została wzięta z populacji:

$$z = \frac{x - \mu}{\sigma}.$$

Standaryzacja z zakłada rozkład normalny danych. Zgodnie z Evert (2005, 80) powyższy wzór można wykorzystać przy mierzeniu korelacji między wyrazami. Jeżeli E_{11} jest wystar-

czająco duże, to – przy przyjęciu hipotezy zerowej – rozkład O_{11} (uzyskany w myśl podobnego rozumowania jak w przypadku t-score) stanowi przybliżenie rozkładu normalnego, ze średnią E_{11} i standardowym odchyleniem $\sqrt{E_{11}}$. Otrzymujemy wzór:

$$z - score = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}.$$

Jak łatwo zauważyć, t-score i z-score różnią się od siebie tylko mianownikiem. Niektórzy w związku z tym traktują wskaźnik t jako heurystyczny wariant wskaźnika z. Pierwszy z nich ma przy tym tę zaletę, że unika niepożądanego przeszacowania w przypadku, gdy E_{11} jest niewielkie (a więc przybliżenie rozkładu dwumianowego do normalnego zaczyna być coraz luźniejsze). Jeżeli E_{11} jest mniejsze od 1, wartość mianownika w z-score powoduje faktyczny wzrost różnicy $O_{11} - E_{11}$, w przypadku t-score mianownik nigdy nie spada poniżej 1.

2.2.5.3. Statystyka testowa chi-kwadrat Pearsona (χ^2)

Jest to statystyka niewymagająca założenia o rozkładzie normalnym populacji, zatem z teoretycznego punktu widzenia odpowiedniejsza niż dwie poprzednie. Sumuje ona kwadraty różnic pomiędzy wartościami obserwowanymi i oczekiwanymi we wszystkich komórkach tablicy dwudzielczej, normalizowane przez wartości oczekiwane:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

$i, j \in \{1,2\}$.

Alternatywnymi postaciami powyższej formuły są wzory (zob. Evert 2005, 81):

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{R_1 R_2 C_1 C_2} \quad \text{i} \quad \chi^2 = \frac{N(O_{11} - E_{11})^2}{E_{11} E_{22}}.$$

2.2.5.4. Logarytm wskaźnika wiarygodności (*log-likehood ratio*)

Dla hipotezy statystycznej można wyznaczyć funkcję wiarygodności, mierzącą, na ile obserwowane dane są prawdopodobne, jeśli przyjmiemy tę hipotezę. Wskaźnik wiarygodności jest to stosunek funkcji wiarygodności dwóch hipotez:

$$\lambda = \frac{L(H_1)}{L(H_2)},$$

gdzie H_1 i H_2 to hipotezy, a *L (H)* – funkcja wiarygodności hipotezy. Jeżeli λ jest większe od 1, oznacza to, że wiarygodność hipotezy H_1 jest odpowiednio większa od wiarygodności H_2.

Tę ideę można wykorzystać do wyodrębniania jednostek wielosegmentowych, porównując hipotezę zerową, zakładającą niezależność członów wyrażenia, z hipotezą przeciwną – zakładającą korelację między wyrazami składowymi. Pierwszą z hipotez formułujemy następująco:

$$H_1 : P(v \mid u) = P(v \mid \neg u) = p.$$

Według hipotezy H_1 prawdopodobieństwo wystąpienia wyrazu *v* pod warunkiem wystąpienia wyrazu *u* jest taka sama jak wystąpienie *v*, gdy wyraz *u* nie wystąpił – zatem *v* jest niezależne od *u*. Alternatywna hipoteza ma postać:

$$H_2 : P(v \mid u) = p_1 \neq p_2 = P(v \mid \neg u).$$

Zgodnie z hipotezą H_2 prawdopodobieństwa wystąpienia *v* w przypadku wystąpienia bądź niewystąpienia *u* są różne, zatem *u* i *v* są od siebie zależne. Jeżeli zatem – w świetle obserwowanych danych – hipoteza o zależności jest bardziej prawdopodobna od hipotezy przyjmującej niezależność, stanowi to podstawę do uznania pary wyrazów *u*, *v* za jednostkę (oczywiście o mocy tego dowodu przesądza wielkość λ).

Prawdopodobieństwa p, p_1, p_2 można wyliczyć, używając estymatora największej wiarygodności:

$$p = P(\cdot,v) = \frac{C_1}{N},$$

$$p_1 = \frac{P(uv)}{P(u,\cdot)} = \frac{O_{11}}{R_1},$$

$$p_1 = \frac{P(\neg uv)}{P(\neg u,\cdot)} = \frac{O_{21}}{R_2}.$$

Funkcja wiarygodności obliczana jest jako iloczyn prawdopodobieństw dwóch przypadków:

1) wystąpienie wyrazu *v*, gdy wystąpił wyraz *u*,
2) wystąpienie *v*, gdy nie wystąpił *u*.

Prawdopodobieństwa te można wyliczyć, przyjmując, jak w przypadku wcześniejszych miar, rozkład dwumianowy: $b\,(k;n,x) = \binom{n}{k} x^k\,(1-x)^{(n-k)}$, gdzie *k* oznacza liczbę sukcesów, *n* liczbę prób, a *x* prawdopodobieństwo sukcesu w pojedynczej próbie. Dla hipotezy H_1 prawdopodobieństwo pierwszego przypadku jest równe $b\,(O_{11}; R_1, p)$, drugiego $b\,(O_{21}; R_2, p)$. Dla hipotezy H_2 w pierwszym przypadku mamy $b\,(O_{11};R_1, p_1)$, w drugim $b\,(O_{21}; R_2, p_2)$. Zatem:

$$L(H_1) = b\,(O_{11}; R_1; p)\; b\,(O_{21}; R_2, p),$$
$$L(H_2) = b\,(O_{11}; R_1; p_1)\; b\,(O_{21}; R_2, p_2).$$

Z powyższych równości łatwo policzyć wskaźnik wiarygodności λ. Lepiej jest jednak zastosować logarytm wskaźnika wiarygodności, który osiąga lepsze wyniki dla przypadków, gdy relatywna frekwencja (mierzona w odniesieniu do wielkości korpusu) bigramu jest niewielka:

$$\begin{aligned} LLR = -2\log\lambda = 2(&O_{11}\log O_{11} + O_{12}\log O_{12} + O_{21}\log O_{21} + O_{22}\log O_{22} - R_1\log R_1 \\ &- C_1\log C_1 - C_2\log C_2 - R_2\log R_2 \\ &+ (O_{11} + O_{12} + O_{21} + O_{22})\log(O_{11} + O_{12} + O_{21} + O_{22})). \end{aligned}$$

Szczegółowe obliczenia prowadzące do uzyskania końcowego wzoru można znaleźć w Seretan (2011b, 134). Evert (2005, 83) przedstawia równoważny wzór:

$$LLR = 2\sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} .$$

2.2.5.5. Informacja wzajemna (*Mutual Information*)

Jest to bardzo często stosowana miara, o tyle charakterystyczna, że wywodzi się z teorii informacji. W literaturze panuje pewien nieład dotyczący nazewnictwa, związany z faktem, że istnieją dwie koncepcje, powiązane ze sobą teoretycznie, które nazywane są informacją wzajemną. Pierwsza z nich czasem zwana jest średnią informacją wzajemną (ang. *Average Mutual Information*), druga – punktową informacją wzajemną (*Pointwise Mutual Information*). W celu uniknięcia nieporozumienia w niniejszej książce obie będą określane za pomocą dłuższej wersji nazwy.

Średnia informacja wzajemna jest rozumiana jako ilość wspólnej informacji, jaką zawierają dwie zmienne losowe X i Y (lub jako zmniejszenie entropii – miary niepewności – Y, gdy dana jest X). Definiowana jest następująco:

$$AMI = \sum_{x,y} p(x, y) \log_2 \frac{p(x,y)}{p(x)\, p(y)} .$$

Warto zauważyć, że jest to miara symetryczna („wzajemna"), a jej wartość wynosi 0 tylko w przypadku, gdy X i Y są niezależne – dlatego też niektórzy mówią o niej jako o „mierze niezależności". Jak można się przekonać (zob. Evert 2005, 89), AMI jest równoważna LLR (logarytmowi wskaźnika wiarygodności).

Punktowa informacja wzajemna dotyczy nie zmiennych losowych, ale konkretnych punktów należących do rozkładów X i Y, i może być rozumiana jako różnica między obserwowanym prawdopodobieństwem wspólnym zdarzeń *x*, *y* a prawdopodobieństwem spodziewanym na podstawie prawdopodobieństw poszczególnych wyrazów, przy założeniu ich niezależności:

$$PMI = \log_2 \frac{p(x,y)}{p(x)\, p(y)} .$$

Prawdopodobieństwa używane we wzorze wylicza się w poniższy sposób:

$$p(x,y) = \frac{O_{11}}{N} , p(x) = \frac{R_1}{N} , p(y) = \frac{C_1}{N} ,$$

mamy zatem:

$$PMI = \log_2 \frac{p(x,y)}{p(x)\, p(y)} = \log_2 \frac{\frac{O_{11}}{N}}{\frac{R_1}{N}\frac{C_1}{N}} = \log_2 \frac{O_{11} N}{R_1 C_1} = \log_2 \frac{O_{11}}{E_{11}} .$$

PMI pozostaje jedną z najpopularniejszych miar używanych do ekstrakcji jednostek wielosegmentowych, jest jednak znana z przeszacowywania wagi rzadkich wydarzeń – wyrażeń, których relatywna frekwencja jest niewielka. Daille (1994) proponuje heurystyczne

warianty PMI – PMI^n, gdzie licznik wzoru podnoszony jest do potęgi *n*. Spośród przetestowanych wartości *n* za najskuteczniejszą uznaje PMI^3:

$$PMI^3 = \log_2 \frac{(O_{11} N)^3}{R_1 C_1} .$$

Bouma (2009) podaje alternatywną wersję PMI, która pomaga zwiększyć użyteczność miary w procesie ekstrakcji:

$$NPMI = \frac{\log_2 \frac{p(x,y)}{p(x)\,p(y)}}{-\log_2 p(x,y)} .$$

W tym wariancie miara jest normalizowana za pomocą logarytmu prawdopodobieństwa wspólnego *x* i *y*[52].

2.2.5.6. Współczynnik Dice'a

Miara popularna w dziedzinie wyszukiwania informacji do porównywania podobieństwa dwóch zbiorów. Jak można się przekonać (zob. Smadja i in. 1996), współczynnik Dice'a zależy wyłącznie od prawdopodobieństw warunkowych zmiennych losowych X i Y, wartość prawdopodobieństw brzegowych obu zmiennych jest nieistotna (inaczej niż w przypadku PMI). Według tej samej pracy współczynnik Dice'a okazał się najefektywniejszy w dziedzinie tłumaczenia kolokacji. Miarę tę można policzyć za pomocą poniższego wzoru:

$$Dice = \frac{2O_{11}}{R_1 + C_1} .$$

2.3. Kwestia języka

Podobnie jak w przypadku rozważań dotyczących definicji jednostek wielowyrazowych, należy się zastanowić, czy język, w obrębie którego pragniemy wodrębnić takie jednostki, wpływa na dobór i skuteczność metod ekstrakcji. Odpowiedź na to pytanie nie jest jednoznaczna. Omawiane wyżej metody lingwistyczne z natury rzeczy muszą być dostosowane do właściwości danego języka – zarówno narzędzia do lematyzacji i dezambiguacji, jak i zawartość stop-listy i brane pod uwagę struktury składniowe. Nie znaczy to jednak, że nie ma tu punktów wspólnych: przykładowo, dla większości języków europejskich duża część wyrażeń wielowyrazowych będzie realizowała schemat składniowy przymiotnik + rzeczownik.

Nieco inaczej sprawa ma się w przypadku technik statystycznych. Ich zasada działania opiera się na współwystępowaniu słów, a jest to cecha, która dotyczy wszystkich języków – dlatego różnice między językami nie mają tak widocznego wpływu na rezultaty działania miar asocjacji. Trzeba jednak pamiętać, że wpływ – choć niewielki – pozostaje. Dla przykładu języki o stabilniejszym szyku (jak np. angielski) mogą zachowywać się nieco inaczej niż te, w których pozycja wyrazów nie odgrywa tak znaczącej roli. Statystyczne metody ekstrakcji, oparte na miarach asocjacji – choć pierwotnie zaprojektowane i używane dla ję-

[52] Bouma w swoim artykule używa logarytmu naturalnego, co nie zmienia niczego z punktu widzenia ekstrakcji – zmianie ulega wyłącznie skala.

zyka angielskiego – mogą być więc bez specjalnych modyfikacji stosowane do wielu innych języków, np. niemieckiego, francuskiego, czeskiego, rosyjskiego itp. Opisane dalej badania, przeprowadzone oczywiście na języku polskim, nie są więc tu wyjątkiem.

2.4. Uwagi dodatkowe

Chociaż dokładna interpretacja wyników poszczególnych miar nie jest istotna z punktu widzenia procesu ekstrakcji, warto wskazać na pewne zależności. Testy statystyczne – wśród nich miary asocjacji – dzielą się na jedno- i dwustronne. Główną (praktyczną) różnicą między nimi jest możliwość – bądź jej brak – rozróżniania między pozytywną a negatywną korelacją. W przypadku testów jednostronnych wysoka wartość miary oznacza silną pozytywną korelację, niska – niezależność lub asocjację negatywną (tendencję do „unikania się" dwóch wyrazów). Wysoka punktacja zdobyta w teście dwustronnym oznacza natomiast istnienie korelacji, przy czym nie da się powiedzieć, czy jest ona pozytywna czy negatywna. Niska wartość miary sugeruje brak powiązania między wyrazami. Do pierwszej grupy należą m.in. z-score, t-score, do drugiej LLR i χ^2.

Spośród licznych miar asocjacji w niniejszym rozdziale przedstawione zostały te, które według wielu opinii sprawdzają się najlepiej w wyodrębnianiu jednostek wielosegmentowych. Ciekawy jest fakt, niejednokrotnie zauważany przez badaczy (np. Evert i Krenn 2005), że nie ma jednej, uniwersalnej, najkorzystniej wypadającej miary: to, która sprawdza się najlepiej, zależy od typu składniowego jednostek, języka, rodzaju korpusu i innych czynników. Wybór najskuteczniejszej metody jest w dużej mierze kwestią eksperymentów.

IV
Algorytm automatycznej ekstrakcji nieciągłych jednostek leksykalnych

Teoretyczne zagadnienia dotyczące definiowania i wyszukiwania nieciągłych jednostek słownika są bardzo zajmujące z naukowego punktu widzenia, jednak lingwistyka komputerowa skupia się w większym stopniu na praktycznych aspektach przetwarzania języka. Jeśli automatyczny system ma być „świadomy" istnienia wielosegmentowych jednostek i mieć możliwość odpowiedniego ich traktowania, należy je uprzednio zgromadzić i odpowiednio opisać (w sposób zrozumiały dla komputera). Ze względu na fakt, że w każdym języku naturalnym istnieje ogromna ilość nieciągłych jednostek leksykalnych (zob. Rozdział II), zadanie takie powinno być w jak największym stopniu zautomatyzowane. Niniejszy rozdział przedstawia algorytm umożliwiający automatyczne wyodrębnianie pożądanych jednostek z tekstu w języku polskim.

1. Cel i profil algorytmu

Zadaniem opisywanego algorytmu jest wyodrębnienie wielosegmentowych jednostek leksykalnych z korpusu polskich tekstów, w celu utworzenia bądź rozwinięcia komputerowego słownika takich jednostek. Należy tu zauważyć, że istnieją obecnie automatyczne systemy spełniające takie właśnie zadanie, jednak zdecydowana większość z nich opracowana jest dla języków innych niż polski (a więc nie wykorzystuje jego specyficznych cech gramatycznych), natomiast polskie próby dotyczą wyłącznie kolokacji (zob. Rozdział I, podpunkt 5.3).

Przed przystąpieniem do realizacji powyższego celu należy najpierw dokonać kilku wstępnych założeń dotyczących algorytmu. Niektóre z nich są wymuszone przez specyfikę dziedziny, przyjęcie innych jest kwestią arbitralnej decyzji.

1) Istnieje wiele różnych typów jednostek wieloczłonowych, często dosyć odmiennych pod względem budowy i funkcji. Algorytm powinien mieć możliwość identyfikowania możliwie szerokiego spektrum jednostek. Nie oznacza to, że każdy z typów musi być wykrywany z jednakową skutecznością (ze względu na specyfikę problemu główny nacisk położony powinien być na bigramy);
2) Algorytm powinna cechować względna elastyczność, tak by mógł pracować z tekstami o różnej zawartości lingwistycznych metadanych, wliczając w to takie, które nie zawierają żadnych dodatkowych informacji;

3) Wynik działania algorytmu powinien być traktowany jako punkt wyjścia do dalszej pracy dokonywanej przez ludzi. Jest to wymuszone przez stosunkowo niską precyzję, jaką osiągają automatyczne systemy wyodrębniające jednostki wielowyrazowe[53]. Różne zadania z dziedziny lingwistyki komputerowej charakteryzuje różny stopień złożoności. Przekłada się to na ogólną skuteczność algorytmów próbujących dany problem rozwiązać. Przykładowo, zgodnie z Manning i Schütze (1999, 233), dosyć łatwo jest opracować algorytm przypisujący angielskim wyrazom odpowiednie części mowy, który osiągałby 90% skuteczności, z kolei skuteczność taka jest nieosiągalna dla systemów tłumaczenia maszynowego. Automatyczna ekstrakcja jednostek wielosegmentowych jest zadaniem bardzo trudnym – kwestię złożoności problemu (zarówno teoretycznej, związanej z definiowaniem jednostek, jak i praktycznej, dotyczącej rozpoznawania ich w tekście) poruszały poprzednie rozdziały. Przyjęcie powyższego założenia powoduje, że istotniejszym błędem staje się pominięcie przez algorytm faktycznej jednostki niż błędne uznanie za jednostkę swobodnego połączenia. Innymi słowy algorytm powinien cechować się przede wszystkim wysoką kompletnością, natomiast co do dokładności wymagania są nieco mniejsze[54];
4) Algorytm powinien mieć formę hybrydową: łączyć zarówno metody lingwistyczne, jak i statystyczne w celu zmaksymalizowania skuteczności.

2. Opis algorytmu

2.1. Zasada działania

Algorytm przeszukuje tekst i wyodrębnia z niego sekwencje wyrazów, które odpowiadają zdefiniowanym uprzednio wzorcom syntaktycznym. Przykładem takiego wzorca może być np. schemat składniowy RZECZOWNIK ^ PRZYMIOTNIK, gdzie symbol ^ oznacza związek zgody pomiędzy oboma członami. W dalszej części rozdziału zamiast nazw części mowy w schematach składniowych będą używane symbole powszechnie używane w literaturze przedmiotu. Przykładowo, wspomniany wyżej schemat będzie oznaczany jako N^A. Tabela IV.1 prezentuje zestaw używanych symboli.

Dla każdego elementu tak powstałej listy zliczana jest frekwencja wystąpienia całej sekwencji, a także każdego z jej wyrazów składowych. W przypadku bigramów obliczane są także miary asocjacji dla kandydata na jednostkę. Dla wszystkich członów wyrażenia szacowany jest również stopień jego łączliwości leksykalnej i stopień homonimiczności. Za każdą z tych danych stoi odpowiednie założenie o podłożu lingwistycznym:

[53] Należy tu zauważyć, że samo zagadnienie oceny skuteczności algorytmu nie jest oczywiste. O ile w przypadku niektórych zadań skuteczność można określić bardzo łatwo (np. w przypadku oznaczania części mowy skuteczność wyznacza procent prawidłowych decyzji), o tyle w innych przypadkach o jakości algorytmu mogą też świadczyć inne czynniki. Więcej na ten temat w podrozdziale 3.

[54] Nie znaczy to oczywiście, że dokładność jest nieistotna: nie stanowi dużego problemu opracowanie algorytmu, który potrafiłby wyodrębnić prawie wszystkie pożądane jednostki z tekstu, ale kosztem tego, że tylko jedno na dziesięć wyrażeń uznanych przez algorytm za jednostkę faktycznie by jednostką było. Z oczywistych względów taki algorytm byłby niemal bezużyteczny.

Część mowy	Znaczenie	Przypadek	Znaczenie	Symbol	Znaczenie
N	rzeczownik	N	mianownik	^	związek zgody
A	przymiotnik	G	dopełniacz	(X)	przypadek X
V	czasownik	D	celownik		
Adv	przysłówek	A	biernik		
P	zaimek	L	miejscownik		
Pr	przyimek	I	narzędnik		
(N)	mianownik	V	wołacz		

Tabela IV.1: Symbole wykorzystywane we wzorcach syntaktycznych

1) Wysoka frekwencja i miary asocjacji mogą świadczyć o częstym wykorzystywaniu danego połączenia w takim a nie innym kształcie, co sugeruje jego ustalenie. Jedna z miar asocjacji – informacja wzajemna – mierzy z kolei stopień prawdopodobieństwa, z jakim wystąpienie wyrazu $\mathbf{w}_1$ pociąga za sobą wystąpienie wyrazu $\mathbf{w}_2$;
2) Mała łączliwość leksykalna wyrazu wchodzącego w skład badanego połączenia (a więc niewielka liczba różnych połączeń, które taki wyraz może tworzyć) świadczy o tym, że jest to wyraz o znaczeniu czy zastosowaniu w jakiś sposób specyficznym, mniej ogólnym. Może to sugerować, że połączenie, w którego skład wchodzi, jest faktyczną jednostką leksykalną. Pozostaje to także w ścisłym związku z ideą zamkniętości klasy substytucyjnej proponowanej przez Bogusławskiego i będącej jednym z kryterium uznawania połączeń za jednostki w myśl niniejszej książce (patrz Rozdział II, podpunkt 2.2.3.9).
   Stopień łączliwości wyrazu szacowany jest przez algorytm za pomocą liczby substytucji prawostronnych (rozumianych jako bigramy występujące w tekście źródłowym, w których pierwszym członem jest badany wyraz) i lewostronnych (bigramy, w których drugim członem jest badany wyraz). Dla przykładu, człony wyrażenia *misterium paschalne* cechuje niewielka łączliwość (oba wyrazy łączą się z niewielką grupą innych wyrazów, najczęściej o tematyce religijnej, np. *misterium Męki Pańskiej*, *rok paschalny*). Z kolei połączenie *nowy stół* składa się z dwóch wyrazów o ewidentnie wysokiej łączliwości;
3) Wysoka homonimiczność wyrazu (rozumiana przez algorytm jako liczba różnych wyrazów mających taką samą postać w tekście: np. *mam* – forma czasownika *mieć* i *mam* – forma rzeczownika *mama*) oznacza, że istnieje wiele interpretacji semantycznych lub składniowych wyrażenia, w którego skład wchodzi ten wyraz, co zwiększa podatność wyrażenia na błędną klasyfikację.

Należy zauważyć, że żadna z powyższych cech nie daje możliwości klasyfikacji badanych wyrażeń w sposób pewny. Przykładowo swobodne połączenie *ostatni rok* w korpusie IPI PAN cechuje wysoka frekwencja i miary asocjacji, a przymiotnik *biały* w jednostce leksykalnej *biały kruk* ma bardzo dużą łączliwość[55]. Z tego powodu wymienione wyżej ce-

[55] Przy założeniu, że unikalne w przypadku *białego kruka* znaczenie komponentu przymiotnikowego jest ignorowane. Jako że system komputerowy nie ma możliwości rozróżnienia znaczenia konwencjonalnego od specyficznego (dysponuje tylko informacjami o tekstowej formie wyrazu, jego lemacie i własnościach gramatycznych), jest to założenie uprawnione.

chy (frekwencja, ewentualne miary asocjacji, stopień łączliwości i homonimiczności) są podstawą do wyliczenia zbiorczej oceny dla danego kandydata za pomocą klasyfikatora wykorzystującego technikę maszynowego uczenia. Klasyfikator taki na podstawie danych treningowych uczy się, w jaki sposób zestawić wartości poszczególnych cech, aby uzyskać najbardziej miarodajną ocenę.

Rezultatem działania algorytmu jest lista zawierająca sekwencje wyrazów wraz z ich oceną, która w założeniu ma odzwierciedlać stopień, w jakim mogą być uznane za jednostkę leksykalną: im wyższa ocena, tym większe prawdopodobieństwo, że badane połączenie wyrazów jest jednostką wielowyrazową. Ocenę taką można również rozumieć jako sugerowany przez algorytm „stopień jednostkowości" wyrażenia.

2.2. Źródła danych tekstowych

Algorytm pracuje na tekście w języku polskim, który – aby jakość ekstrakcji była wystarczająco wysoka – powinien być odpowiednio długi (co najmniej kilkadziesiąt tysięcy wyrazów). W związku z tym najbardziej oczywistym źródłem danych tekstowych są korpusy tekstów. Korpusy mogą zawierać różną ilość dodatkowych informacji (metadanych). Najprostsze pod tym względem mają postać zwykłego pliku tekstowego, niezawierającego żadnych dodatkowych informacji. Bardziej złożone dysponują metadanymi, które mogą odnosić się do tekstów składowych lub poszczególnych jednostek tekstu (rozdziałów, akapitów, zdań czy wyrazów). Z punktu widzenia algorytmu istotne są zwłaszcza dane morfosyntaktyczne opisujące wyrazy. Korpusy mogą wymagać specjalnych narzędzi pozwalających na ich przeglądanie. W szczególności korpusy opracowane przez Instytut Podstaw Informatyki PAN (m.in. korpus IPI PAN, jego próbka *sample*, Narodowy Korpus Języka Polskiego) dysponują własnym zestawem znaczników morfosyntaktycznych, a korzystanie z nich wymaga użycia narzędzia Poliqarp (Janus i Przepiórkowski 2007).

Algorytm zasadniczo nie jest uzależniony od typu korpusu (choć w przypadku, gdy dysponuje odpowiednimi metadanymi, będzie działać lepiej), jednak jeśli źródłem danych jest korpus anotowany (posiadający metadane), musi mieć dostęp do interfejsu (zestawu funkcji pozwalających na odczyt danych z korpusu). Podobnie jest w przypadku, gdy interakcja z korpusem wymaga specyficznych narzędzi. W prezentowanym w niniejszej książce programie zaimplementowany został interfejs umożliwiający komunikację z korpusami wymagającymi korzystania z Poliqarpa.

2.3. Schemat działania algorytmu

Proces wyszukiwania jednostek wielosegmentowych odbywa się w pięciu zasadniczych etapach:

1) procedury wstępne,
2) wyszukiwanie kandydatów na jednostki,
3) ocena jednostek za pomocą miar asocjacji,
4) klasyfikacja jednostek metodą maszynowego uczenia,
5) procedury końcowe.

2.3.1. Procedury wstępne

Na tym etapie algorytm dokonuje wczytania i utworzenia wszystkich potrzebnych danych. Dane, z których korzysta program, są następujące:

a) tekst źródłowy (najczęściej korpus tekstów), w którym przeprowadzane jest wyszukiwanie jednostek nieciągłych,
b) słownik frekwencyjny wyrazów w tekście źródłowym (tworzony przez program),
c) słownik form języka polskiego: może to być CLP – biblioteka języka C umożliwiająca programistyczny dostęp do Słownika Fleksyjnego Języka Polskiego (Gajęcki 2009) lub słownik wykorzystywany przez tagger Morfeusz (Woliński 2006).

Na tym etapie dokonywana jest także segmentacja tekstu wejściowego (o ile jest potrzebna) na wyrazy. W przypadku korpusu dostępnego za pomocą Poliqarpa tworzone są pliki zawierające sekwencje wyrazów, które realizują określone schematy syntaktyczne, np. zawierające wszystkie wystąpienia par AN (PRZYMIOTNIK + RZECZOWNIK). Te pliki (w obrębie których zapewniona jest przez Poliqarpa segmentacja) stanowią źródło danych tekstowych.

2.3.2. Wyszukiwanie kandydatów na jednostki

Na tym etapie algorytm przeszukuje korpus w poszukiwaniu kandydatów na jednostki leksykalne. Procedura ta różni się w szczegółach dla korpusu anotowanego i nieanotowanego.

2.3.2.1. Przeszukiwanie korpusu

Korpus nie zawierający anotacji przeszukiwany jest za pomocą tzw. techniki okienkowej. Polega ona na wyodrębnieniu z tekstu sekwencji wyrazów o potrzebnej długości (np. dla bigramów długość wynosi 2) i dokonaniu na takim segmencie odpowiednich operacji (w przypadku omawianego algorytmu jest to sprawdzenie, czy segment jest zgodny z wzorcem syntaktycznym i czy spełnia ewentualne dodatkowe wymagania). Następnie „okno" jest przesuwane o jeden wyraz w prawo, o ile nie napotkany zostanie znak interpunkcyjny świadczący o końcu frazy (w takim przypadku okno jest czyszczone i proces jest ponawiany od początku następnej frazy).

W przypadku korpusów obsługiwanych przez Poliqarpa ta technika jest niepotrzebna: Poliqarp w odpowiedzi na odpowiednio przygotowane zapytanie zwraca listę wyników odpowiadających zadanym warunkom. Przykładowo, zapytanie [pos=adj][pos=noun] informuje Poliqarpa, że powinien zwrócić obecne w korpusie sekwencje NA. Poniżej przedstawiony jest przykładowy rezultat takiego zapytania:

””,” długowłosy [długowłosy:adj:sg:nom:m1:pos] brodacz [brodacz:subst:sg:nom:m1]”,””
””,” swoim [swój:adj:sg:loc:m3:pos] gabinecie [gabinet:subst:sg:loc:m3]”,””
””,” nasze [nasz:adj:pl:acc:n:pos] zaniedbania [zaniedbanie:subst:pl:acc:n]”,””
””,” ofiarnych [ofiarny:adj:pl:gen:f:pos] jałówek [jałówka:subst:pl:gen:f]”,””

2.3.2.2. Wzorce syntaktyczne

Algorytm dysponuje zestawem predefiniowanych lub określonych przez użytkownika wzorców syntaktycznych, na podstawie których wyszukiwane są odpowiednie sekwencje wyrazów. Takimi wzorcami mogą być np. rzeczownik + przymiotnik tworzące związek zgo-

dy (N^A) lub VN(A) – czyli czasownik + rzeczownik w bierniku. Pierwsza z tych struktur pozwala na identyfikację wielosegmentowych jednostek typu *kwas siarkowy*, *panna młoda*, za pomocą drugiej wyodrębnić można jednostki takie jak *bić pianę* czy *łapać okazję*. Tabela IV.2 zawiera listę predefiniowanych wzorców syntaktycznych i procentowy udział sekwencji tego typu w podkorpusie *sample*. Elementem wzorca może być część mowy, część mowy + przypadek, związek zgody.

Wzorzec	**Udział w korpusie *sample***
AN	9.9 %
NA	6.8 %
NN(G)	7.5 %
PrN(G)	3.7 %
PrN(A)	2.5 %
PrN(L)	4.7 %
PrN(I)	1.5 %

Tabela IV.2: Predefiniowane w programie wzorce syntaktyczne i ich procentowy udział w korpusie sample

Wzorce syntaktyczne pełnią rolę powierzchniowego (płytkiego) parsowania syntaktycznego (zwanego w terminologii angielskiej *chunkingiem*). Jest to ważny proces, gdyż w dużym stopniu wyklucza pary wyrazów niepozostające ze sobą w bezpośredniej relacji.

Sprawdzenie, czy badany segment spełnia zadane wymagania składniowe, wymaga analizy morfologicznej wyrazów wchodzących w jego skład. Analizę taką w przypadku korpusów niezawierających odpowiednich metadanych umożliwia biblioteka CLP. Należy zaznaczyć, że nie zapewnia ona dezambiguacji: dostarcza tylko kompletną listę wszystkich leksemów, których badany ciąg liter może być przedstawicielem (wraz z kompletnymi informacjami o kategoriach gramatycznych danej formy). Przykładowo, ciąg liter *bez* zostanie rozpoznany jako potencjalna forma trzech leksemów: *bez* (rzeczownik), *bez* (przyimek) i *beza* (rzeczownik). Konsekwencją tego jest problem w przypadku form homonimicznych: jeżeli przynajmniej jedna z możliwych form spełnia warunki składniowe i jednocześnie przynajmniej jedna takich warunków nie spełnia, powstaje ryzyko błędu. Problem ten można rozwiązać dwojako: albo brać pod uwagę tylko takie wyrażenia, które na pewno spełniają reguły, albo takie, które mogą mieć interpretację zgodną z regułami. Jako że pierwsze rozwiązanie zwiększa precyzję, zmniejszając kompletność, a drugie odwrotnie, w algorytmie (zgodnie z przyjętymi założeniami) zaimplementowano rozwiązanie drugie.

Algorytm wyszukuje w tekście źródłowym wszystkie przypadki wystąpienia sekwencji wyrazów spełniającej dany wzorzec. Uzyskana w ten sposób zostaje lista kandydatów na jednostki – a więc połączeń wyrazowych, które spełniają formalne kryterium, aby móc stanowić jednostkę wielosegmentową odpowiedniego typu.

2.3.2.3. Pozyskiwanie danych liczbowych

Na etapie drugim zbierane są także informacje o frekwencji danego związku wyrazowego (frekwencji całego wyrażenia i poszczególnych wyrazów składowych), a także o łączliwości leksykalnej i stopniu homonimiczności każdego wyrazu składowego. Łączliwość leksykalna jest opisywana przez dwie dane: liczbę substytucji prawostronnych danego wyrazu (liczbę wszystkich bigramów, których pierwszym wyrazem składowym jest badany wyraz) i liczbę substytucji lewostronnych (wyliczaną analogicznie). Stopień homonimiczności opisywany jest za pomocą liczby identycznych pod względem tekstowym form występujących w słowniku.

2.3.2.4. Filtrowanie wyników

Lista kandydatów jest poddawana procesowi filtrowania, w trakcie którego stosowane są dodatkowe kryteria wykluczające niektórych kandydatów. Pierwszym z takich filtrów, mającym kluczowe znaczenie dla poprawnego działania algorytmu, jest odrzucenie wszystkich kandydatów, których frekwencja nie przekracza 4 wystąpień. Poniżej tego progu wyniki kolejnego etapu, opierającego się na działaniach statystycznych, tracą wiarygodność (zob. Rozdział III, podpunkt 2.2.4). Drugi rodzaj filtra ma za zadanie odrzucić te sekwencje, które są uznawane za fragmenty dłuższych jednostek. Jeżeli rozpatrujemy dwa wyrażenia, z których dłuższe zawiera w sobie krótsze, nie da się wybrać właściwego za pomocą prostego porównania frekwencji, gdyż frekwencja wyrażenia krótszego będzie zawsze równa lub większa od frekwencji wyrażenia dłuższego. Przykładowo, wyrażenie *wojna światowa* będzie miało wyższą frekwencję niż faktyczne jednostki *I wojna światowa* i *II wojna światowa*. W algorytmie wykorzystywane jest zatem kryterium kosztu stosowane w pracy Kita i in. (1994)[56]. Wyraża się ono następującym wzorem:

$$K(A)=(|A| - 1) * (f(A) - f(B)),$$

gdzie A oznacza badane wyrażenie, B – dłuższe wyrażenie, zawierające w sobie wyrażenie A, $|A|$ – długość wyrażenia A, $f(X)$ – frekwencję wyrażenia X.

Istota kryterium polega na tym, że jeśli rzeczywistą jednostką jest B, a nie A, to wartość $f(B)$ będzie zbliżona do $f(A)$, zatem kryterium kosztu dla A będzie stosunkowo niskie, natomiast kryterium kosztu dla B (wyliczane analogicznie) będzie wyższe.

Działanie kryterium najlepiej zobrazować na przykładzie. Wyrażenie *pewien sens* występuje w korpusie *sample* 287 razy. Natomiast frekwencja wyrażenia szerszego, *w pewnym sensie* wynosi 284. Kryterium kosztu dla wyrażenia *pewien czas* wynosi:

$$K(\textit{pewien sens}) = (2\text{-}1) * (287 - 284) = 3.$$

W kolejnym etapie liczone jest kryterium kosztu dla wyrażenia *w pewnym sensie*. Frekwencja najpopularniejszego wyrażenia szerszego *jest w pewnym sensie* wynosi 21. Mamy zatem:

$$K(\textit{w pewnym sensie}) = (3\text{-}1) * (284 - 21) = 525.$$

[56] Należy dodać, że kryterium to w cytowanej pracy służy wyszukiwaniu kolokacji, a nie odrzucaniu kandydatów.

Jak widać, przewaga sekwencji trójwyrazowej jest bardzo wyraźna, zatem bigram *pewien sens* zostanie przez algorytm odrzucony.

Ostatnim rodzajem filtra jest lista wyrazów odrzucanych (tzw. *stop-lista*). Lista ta zawiera słowa często pojawiające się korpusie, które nigdy albo prawie nigdy nie wchodzą w skład wieloczłonowych jednostek. Tabela IV.3 ilustruje sposób działania tego kryterium.

wyraz na stop-liście	przykłady odrzuconych wyrażeń
jakiś	czas jakiś; jakiś cud; jakaś prawda
cały	całe życie; cała sprawa; cała historia
sam	sam czas; sam początek; sam dzień
inny	coś innego; inny sposób; inna osoba
następny	rok następny; następny dzień;

Tabela IV.3: Wyrażenia odrzucone na podstawie wyrazów-kluczy znajdujących się na stop-liście

2.3.2.5. Sprowadzanie do formy hasłowej

Na etapie drugim odbywa się także sprowadzanie potencjalnej jednostki do formy hasłowej. Przykładowo, wyrażenie *szarego człowieka* zostaje przekształcone do formy *szary człowiek*. Zasady tego procesu zostały omówione w Rozdziale II (podpunkt 3.7). Normalizacja tego typu jest ważna z dwóch powodów: po pierwsze, lista wyników działania algorytmu powinna być jak najbardziej przyjazna dla ludzkiego odbiorcy, po drugie i ważniejsze, umożliwia to wspólne zliczanie wszystkich form danego wyrażenia (a więc połączenia *szarego człowieka* i *szaremu człowiekowi* uznane zostaną za dwa wystąpienia tego samego wyrażenia *szary człowiek*, a nie za dwa odrębne hasła). Normalizacja nie zawsze przebiega prawidłowo: problem jest zwłaszcza z ustalaniem, w jakiej liczbie powinien być reprezentowany dany wyraz. Algorytm przyjmuje, że jest to liczba pojedyncza, chyba że ponad 90% wszystkich wystąpień odnotowano w liczbie mnogiej (np. w jednostce *wybory prezydenckie*).

2.3.3. Ocena jednostek za pomocą miar asocjacji

Na tym etapie odbywa się obliczanie danych statystycznych. Jeżeli lista kandydatów na jednostki składa się z bigramów, dla każdego kandydata wyliczane są wartości miar asocjacji: informacja wzajemna (PMI) i jej dwa warianty: PMI^3 i NPMI, statystyka z (z-score), statystyka t Studenta (t-score), logarytm wskaźnika wiarygodności (LLR) i współczynnik Dice'a (Dice). Tabela IV.4 ilustruje wyniki uzyskane dla 5 najczęstszych bigramów typu NA w korpusie *sample*. Warto zauważyć, że posortowanie listy według różnych miar da inne wyniki.

Rezultatem tego etapu jest – w przypadku bigramów – lista kandydatów z dodanymi ocenami poszczególnych miar asocjacji. W przypadku gdy poszukiwane są dłuższe sekwencje, wszystkim miarom przyporządkowywana jest wartość 0.

wyrażenie	f1	f2	f3	PMI	PMI^3	NPMI	z-score	t-score	LLR	Dice
działalność gospodarcza	1650	10270	8224	8.74	30.12	0.63	839.1	40.52	226.8	0.178
szkoła podstawowa	1599	22340	8148	7.59	28.87	0.55	552.3	39.78	6868.3	0.104
samorząd terytorialny	1265	5623	2084	11.21	31.82	0.79	1730.5	35.55	1687.6	0.328
piłka nożna	1211	8355	1598	10.95	31.44	0.77	1551.8	34.78	5027.4	0.243
pomoc społeczna	876	12698	10371	7.18	26.73	0.49	355.17	29.39	6494.2	0.075

Tabela IV.4: Frekwencja i miary asocjacji dla pięciu najpopularniejszych wyrażeń typu NA w korpusie sample. Kolumny f1, f2, f2 oznaczają odpowiednio: frekwencję całości, frekwencję pierwszego wyrazu, frekwencję drugiego wyrazu

2.3.4. Klasyfikacja jednostek metodą maszynowego uczenia się

Jak zostało to opisane wyżej, każdy z branych pod uwagę czynników (frekwencja, miary asocjacji, łączliwość leksykalna, homonimiczność) zapewnia pewną dozę informacji na temat stopnia, w jakim można dane wyrażenie uznać za jednostkę, jednak skuteczność klasyfikacji opartej na pojedynczych czynnikach pozostawia wiele do życzenia. Naturalnym krokiem umożliwiającym lepszą klasyfikację byłoby zebranie wszystkich czynników i połączenie ich w jeden mechanizm klasyfikujący. Można tego dokonać na różne sposoby, np. przez „głosowanie reguł", w którym każdy z czynników mógłby klasyfikować wyrażenie osobno, następnie zaś wszystkie głosy „za" i „przeciw" byłyby sumowane. Z drugiej strony warto zauważyć, że ocena badanego wyrażenia jest przez każdy z czynników przedstawiona na skali liczbowej – zatem lepszym pomysłem wydaje się zsumowanie tych ocen, przy czym każdy z czynników charakteryzowałby się własną wagą, odzwierciedlającą jego istotność. Powstaje jednak pytanie: jak znaleźć odpowiednie wagi? Arbitralne ich ustalanie i testowanie wydaje się procesem żmudnym i narażonym na rozliczne błędy. Inną kwestią jest to, czy optymalne jest sumowanie ocen, czy może istnieje jakaś bardziej skuteczna funkcja pozwalająca na ich połączenie. Z pomocą w takiej sytuacji przychodzą metody maszynowego uczenia (ang. *machine learning*).

2.3.4.1. Podstawowe pojęcia maszynowego uczenia

Człowiek, próbując nauczyć się odpowiedniego postępowania w nieznanej sobie sytuacji, często próbuje pewnych działań i obserwuje, jakie są ich konsekwencje, a następnie stara się znaleźć ogólną zasadę postępowania. Może również obserwować działania kogoś, kto wie, jak sobie z problemem poradzić. Podobna zasada przyświeca idei maszynowego uczenia: algorytm zapoznaje się z przykładami pewnego problemu, który ma za zadanie rozwiązać, a następnie stara się znaleźć odpowiednie uogólnienie pozwalające na podejmowanie decyzji. Problemem do rozwiązania może być np. rozpoznawanie twarzy na zdjęciu albo – jak w opisywanym przypadku – znalezienie funkcji, która najlepiej klasyfikuje pewne obiekty.

Zestaw przykładów, na którym algorytm się uczy, nazywany jest zbiorem treningowym. Każdy przykład reprezentowany jest przez zestaw własności, które możliwie dokładnie

opisują problem, nazywane cechami lub atrybutami. Przykładowo, jeśli chcemy zbudować prosty algorytm przewidujący cenę mieszkania na podstawie kilku cech, fragment zbioru treningowego mógłby wyglądać następująco (jest to oczywiście przykład skrajnie uproszczony, mający wyłącznie ilustrować koncepcję):

powierzchnia	**liczba pokoi**	**odległość od centrum**	**cena**
30 m^2	1	5 km	150
44 m^2	2	12 km	230
80 m^2	3	3 km	450

Tabela IV.5: Fragment prostego zbioru treningowego

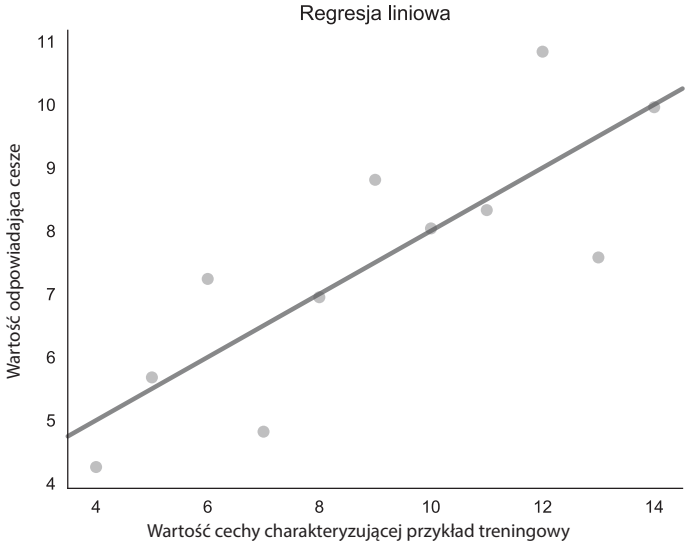
W powyższym przykładzie pierwsze trzy kolumny zawierają wartości poszczególnych cech, ostatnia – wartość, którą algorytm ma za zadanie przewidywać.

Metody maszynowego uczenia można podzielić na dwie grupy: uczenie z nadzorem (ang. *supervised learning*) i uczenie bez nadzoru (ang. *unsupervised learning*). Algorytmy uczenia z nadzorem wymagają zbioru treningowego, w którym przykłady są prawidłowo ocenione (najczęściej przez człowieka). W powyższym przykładzie rolę oceny pełnią realne ceny mieszkań charakteryzujących się odpowiednimi parametrami.

Algorytmy uczenia bez nadzoru nie wymagają, by przykłady w zbiorze treningowym były ocenione, dają też lepsze możliwości uogólnienia problemu, jednak ich skuteczność jest przeważnie niższa niż metod z nadzorem. W dalszym ciągu opisu przyjmuje się, że mamy do czynienia z algorytmem uczenia z nadzorem.

Algorytmy uczące się można podzielić również ze względu na rodzaj problemu, który rozwiązują. Algorytmy klasyfikacyjne jako wynik dla każdego przykładu podają wartość dyskretną (tj. przyjmującą wartości ze skończonego zbioru, np. {0, 1}), reprezentującą klasę, do której przypisany zostaje dany przykład. Algorytmy regresyjne jako wynik podają wartość ciągłą (np. liczbę rzeczywistą). Przykładowo, algorytm rozpoznający odręcznie pisane cyfry jest algorytmem klasyfikacyjnym (wynikiem jest zakwalifikowanie przykładu do jednej z dziesięciu grup), z kolei algorytm próbujący przewidzieć cenę mieszkania jest algorytmem regresyjnym.

Wartość wynikowa przewidywana przez algorytm dla każdego przykładu jest uzyskiwana na podstawie funkcji (zależnej od typu algorytmu), zwanej hipotezą, której zmiennymi są cechy i która każdej cesze przypisuje odpowiedni parametr (wagę). Zadaniem algorytmu uczącego jest takie dobranie parametrów, aby hipoteza minimalizowała błąd przewidywania. Wykres 1 ilustruje działanie algorytmu regresji liniowej: punkty osi X odpowiadają przykładom w zbiorze treningowym (wykres dwuwymiarowy odnosi się do sytuacji, w której każdy punkt opisany jest tylko za pomocą jednej cechy – np. powierzchnia mieszkania z przykładu powyżej), oś Y odpowiada wartościom funkcji, którą algorytm stara się przybliżyć (np. cena mieszkania). Algorytm stara się znaleźć taką funkcję liniową, która przechodzi możliwie najbliżej wszystkich punktów wyznaczanych przez wartości x i y. Kiedy taką funkcję znajdzie, może na jej podstawie wyznaczać wartości dla nowych przykładów, które nie znalazły się w zbiorze treningowym.



Wykres 1: Regresja liniowa

2.3.4.2. Wykorzystanie maszynowego uczenia do ekstrakcji jednostek nieciągłych

2.3.4.2.1. Algorytm: maszyna wektorów nośnych

Metodą maszynowego uczenia zastosowaną w przypadku programu do ekstrakcji wielosegmentowych jednostek leksykalnych jest maszyna wektorów nośnych (ang. *Support Vector Machine* – SVM). Jest to dosyć zaawansowana i stosunkowo młoda metoda, w przypadku wielu problemów dająca bardzo dobre wyniki.

Mówiąc najprościej, metoda ta znajduje taką płaszczyznę (ściślej rzecz biorąc, w przypadku gdy cech jest więcej niż dwie, hiperpłaszczyznę), która rozgranicza przykłady należące do różnych kategorii z maksymalnie dużym marginesem.

Metoda SVM została pierwotnie opracowana do rozwiązywania problemów klasyfikacji, jednak istnieje jej rozszerzenie umożliwiające regresję. Jest to ważne z punktu widzenia ekstrakcji wielosegmentowych jednostek, gdyż można dzięki tej metodzie otrzymać funkcję rankingową, która umożliwia posortowanie zidentyfikowanych jednostek od najbardziej do najmniej prawdopodobnych.

2.3.4.2.2. Zestaw cech

Cechy opisujące każdy przykład (w tym przypadku zbiorem przykładów jest lista wyrażeń – kandydatów na jednostki) zaprezentowane zostały w punkcie 2.1. Dla przypomnienia są to: frekwencja całości wyrażenia, frekwencja poszczególnych wyrazów składowych, miary asocjacji (PMI, PMI3, NPMI, z-score, t-score, LLR i Dice), liczba substytucji prawostronnych i lewostronnych wszystkich wyrazów składowych, liczba form homonimicznych wszystkich wyrazów, przy czym liczba substytucji i homonimów poszczególnych wyrazów są uśredniane. Uśrednienie jest konieczne, ponieważ algorytm ekstrakcji nie stawia ograniczeń co do długości badanego wyrażenia, natomiast klasyfikator SVM wymaga, by liczba cech opisujących przykłady była zawsze taka sama.

Wartości cech są normalizowane za pomocą wzoru: $f = \frac{f - \mu}{max(f) - min(f)}$,

gdzie *f* oznacza wartość cechy, μ – średnią cechy dla wszystkich przykładów ze zbioru treningowego, *max (f)* i *min (f)* odpowiednio najwyższą i najniższą wartość cechy, jaka występuje w zbiorze.

Normalizacja powoduje, że średnia danej cechy w obrębie wszystkich przykładów ma wartość oczekiwaną równą 0, a wartość rozpiętości różnych cech jest zbliżona. Proces ten jest konieczny ze względu na to, że klasyfikator SVM jest bardzo wrażliwy na różnice w skali cech (cechy przyjmujące generalnie wyższe wartości zdominowałyby pozostałe, o niższym zakresie wartości).

Należy zauważyć, że algorytm oceniający na tym etapie nie posiada żadnej wiedzy na temat tego, jak brzmi oceniane wyrażenie – wszystko, czym dysponuje, to opisane wyżej dane liczbowe.

2.3.4.2.3. Zbiór treningowy

Jako zbiór treningowy, na którym trenowany jest klasyfikator SVM, wykorzystany jest zbiór porównawczy, używany do ewaluacji całego algorytmu, szerzej opisany w podpunkcie 3.2. Składa się on z 2067 ręcznie ocenionych wyrażeń, w tym 822 (39%) stanowią rzeczywiste jednostki wielosegmentowe, a 1245 – zwykłe połączenia. Należy tu podkreślić, że informacje zawarte w zbiorze treningowym na temat poprawnej klasyfikacji poszczególnych przykładów (czy stanowią jednostki, czy zwykłe połączenia) są dostępne dla algorytmu tylko podczas treningu. Przy samodzielnym ocenianiu przykładów po zakończeniu procesu uczenia (w tym takich, na których się uczył) algorytm nie zna ich prawidłowej klasyfikacji.

2.3.5. Etap końcowy

Rezultatem działania ostatniego etapu jest lista wyrażeń spełniających odpowiednie warunki co do struktury, z których każdemu przypisana jest ocena (wyrażona liczbą rzeczywistą), interpretowana jako wyznacznik prawdopodobieństwa, że wyrażenie jest jednostką. Taka lista jest sortowana w kolejności od najwyżej ocenionych do tych z najniższą oceną. Wyniki są następnie zapisywane w pliku tekstowym.

3. Wyniki i ewaluacja algorytmu

W dawniejszych pracach poświęconych ekstrakcji wyrażeń wielowyrazowych nie przykładano wiele wagi do ewaluacji algorytmów, skupiając się przede wszystkim na opisie konkretnych technik ekstrakcji. W późniejszych pracach problemowi ewaluacji zaczęto poświęcać więcej uwagi, proponowano różne techniki. Wykształciły się pewne zwyczaje, które obecnie w mniejszym lub większym stopniu są powielane w kolejnych pracach. Niniejsza książka trzyma się tej tradycji.

3.1. Metodologia

Proces ekstrakcji jednostek nieciągłych daje w rezultacie listę wyników posortowanych od najlepiej ocenionych do ocenionych najgorzej. Listę taką można interpretować na dwa sposoby. W myśl pierwszej ustalany jest próg, powyżej którego wyrażenia traktowane są

jako jednostki, a poniżej – jako zwykłe połączenia. Otrzymujemy w ten sposób binarną klasyfikację. Drugą możliwością interpretacji jest wzięcie pod uwagę uporządkowania listy, możemy wtedy przyjąć, że te na szczycie listy są „bardziej jednostkami", a wraz z przesuwaniem się w dół listy cecha „jednostkowości" jest coraz mniej wyraźna. Oba podejścia mają swoje zalety i wady, w obu przypadkach można także mierzyć skuteczność algorytmu.

Wyniki pracy algorytmu można podzielić na cztery grupy: prawidłowe pozytywne (ang. *true positives*) – jednostki wielosegmentowe prawidłowo wyodrębnione przez algorytm, nieprawidłowe pozytywne (ang. *false positives*) – swobodne połączenia uznane przez algorytm za jednostki, prawidłowe negatywne (ang. *true negatives*) – swobodne połączenia prawidłowo oznaczone przez algorytm i nieprawidłowe negatywne (ang. *false negatives*) – swobodne połączenia sklasyfikowane przez algorytm jako jednostki. Kategorie te będą odpowiednio oznaczane jako PP, NP, PN, NN. Poniższa tabela ilustruje to zestawienie:

		stan faktyczny	
		jednostki	połączenia
klasyfikacja algorytmu	jednostki	**PP**	**NP**
	połączenia	**NN**	**PN**

Tabela IV.6: Zestawienie możliwych kategorii klasyfikacji algorytmu. PP – prawidłowo pozytywne, NP – nieprawidłowo pozytywne, NN – nieprawidłowo negatywne, PN – prawidłowo negatywne

3.1.1. Zbiór porównawczy

Standardową metodą ewaluacji jest porównanie wyników algorytmu do zbioru porównawczego (ang. *reference set* / *gold standard*). Zbiór taki oparty jest na konkretnym korpusie i zawiera wszystkie wyrażenia znajdujące się w korpusie, które w myśl danej definicji traktowane są jako wyrażenia wielowyrazowe, pozyskane w drodze ręcznej anotacji.

Należy zauważyć, że przygotowanie takiego zbioru jest dosyć pracochłonne, jako że eksperci dokonujący klasyfikacji muszą ocenić duże ilości danych (często sięgające kilku tysięcy wyrażeń), przy czym należy pamiętać, że sama klasyfikacja nie jest zadaniem prostym, co automatycznie przekłada się na dłuższy czas opracowywania zbioru. Dodatkowo, wiele ze zbiorów porównawczych jest tworzonych przez dwóch lub więcej anotatorów w celu zwiększenia obiektywności klasyfikacji.

W języku polskim jak do tej pory nie powstał żaden zbiór porównawczy, co trzeba złożyć na karb niewielkiej popularności tematu na rodzimym gruncie i dużych wymagań, jeśli chodzi o nakład pracy. Z tego powodu powstała konieczność, by w ramach niniejszej książce taki zbiór został opracowany.

Wybór odpowiedniego korpusu, na podstawie którego opracowany został później zbiór porównawczy, podyktowany był następującymi czynnikami:

1) Korpus powinien być odpowiednio duży, tak by zapewniał ilość danych wystarczającą do statystycznego przetwarzania;
2) Korpus powinien być anotowany morfosyntaktycznie – jest to ważne ze względu na bogatą warstwę morfologiczną języka polskiego;
3) Korpus powinien być ogólnie dostępny do przetwarzania.

Wybór padł na korpus *sample*, który spełnia powyższe warunki. Jest to zróżnicowana pod względem typów tekstów15-milionowa próbka korpusu IPI PAN. Podział tekstów w korpusie jest następujący:

- proza współczesna: 10%,
- proza dawna: 10%,
- teksty naukowe: 10%,
- teksty prasowe: 50%,
- stenogramy sejmowe i senackie: 15%,
- ustawy: 5%.

Korpus jest anotowany morfosyntaktycznie, do anotacji użyty jest zestaw znaczników opisywany w pracy Woliński (2003). Tabela IV.7 zawiera zestaw najczęstszych znaczników używanych w korpusie (pełną listę można znaleźć w pracy Przepiórkowski i in. 2012).

Pewnym utrudnieniem jest fakt, że korpus *sample* nie jest dostępny w postaci źródłowej – dostęp do niego jest możliwy za pomocą narzędzia Poliqarp. W ramach tego narzędzia dostarczany jest jednak moduł pozwalający na sterowanie Poliqarpem z poziomu linii komend, co jest wystarczające do zamierzonych celów.

kategoria gramatyczna	**symbol**	**przykłady**
CZĘŚCI MOWY		
rzeczownik	subst	*kot*, *profesorowie*
liczebnik	num	*sześć*, *dużo*
przymiotnik	adj	*polski*
przysłówek	adv	*bardziej*
zaimek nietrzecioosobowy	ppron12	*ja*, *tobie*
zaimek trzecioosobowy	ppron3	*on*, *jemu*
czasownik – forma nieprzeszła	fin	*jadam*
czasownik – aglutynat czasownika *być*	aglt	*-śmy*
czasownik – bezokolicznik	inf	*jadać*
imiesłów przysłówkowy współczesny	pcon	*jadając*
imiesłów przysłówkowy uprzedni	pant	*zjadłszy*
imiesłów przymiotnikowy czynny	pact	*jadający*
imiesłów przymiotnikowy bierny	ppas	*jadany*
predykatyw	pred	*trzeba, słychać*
przyimek	prep	*pod*, *we*
spójnik współrzędny	conj	*oraz*, *lub*
spójnik podrzędny	comp	*że*, *aby*
wykrzyknik	interj	*ach*, *psiakrew*
kublik	qub	*nie*, *-ż*
skrót	brev	*dr*, *np*
interpunkcja	interp	*;,.*

PRZYPADKI		
mianownik	nom	*woda*
dopełniacz	gen	*wody*
celownik	dat	*wodzie*
biernik	acc	*wodę*
narzędnik	inst	*wodą*
miejscownik	loc	*wodzie*
wołacz	voc	*wodo*
RODZAJ		
męski osobowy	m1	*papież*
męski zwierzęcy	m2	*baranek*, *walc*
męski rzeczowy	m3	*stół*
żeński	f	*stuła*
nijaki	n	*okno*, *co*
LICZBA		
pojedyncza	sing	*oko*
mnoga	pl	*oczy*
OSOBA		
pierwsza	pri	*bredzę*
druga	sec	*bredzisz*
trzecia	ter	*bredzi*
STOPIEŃ		
równy	pos	*cudny*
wyższy	com	*cudniejszy*
najwyższy	sup	*najcudniejszy*

Tabela IV.7: Zestaw znaczników używanych w korpusie ***sample***

3.1.1.1. Proces tworzenia zbioru porównawczego

Rozmiar korpusu *sample* jest na tyle duży, że ręczna anotacja całości byłaby zadaniem zbyt czasochłonnym. W związku z tym pojawiła się konieczność ograniczenia ilości danych przeznaczonych do oceny. Osiągnięto to za pomocą ograniczenia wielkości korpusu do początkowych 5 milionów segmentów oraz wybrania spośród możliwych składniowych wzorców wielosegmentowych jednostek trzech najpopularniejszych i najstabilniejszych pod względem składniowym: NA, AN i NN(G).

W pierwszym etapie wyodrębniono z powyższego fragmentu korpusu wszystkie bigramy spełniające powyższe warunki składniowe. Otrzymano w ten sposób ok. 243 tysięcy bigramów typu AN, 147 tysięcy bigramów typu NA i 138 tysięcy bigramów typu NN(G), łącznie około 530 tysięcy wyrażeń. Wszystkie zostały sprowadzone do postaci hasłowej, następnie utworzono listę wszystkich haseł. Powstała lista liczy 172 934 odrębnych potencjalnych jednostek, w tym 64 124 typu AN, 37 267 typu NA i 71 543 typu NN(G).

W kolejnym etapie odrzucono wszystkie hasła, których frekwencja wynosiła 10 lub mniej, dzięki czemu wyniki można uznać za statystycznie znaczące. W ten sposób lista została ograniczona do 683 potencjalnych jednostek typu AN, 773 haseł typu NA i 611 haseł

typu NN(G)[57]. W sumie cała lista liczy 2067 potencjalnych bigramów wielosegmentowych. Tabela IV.8 zestawia wyniki uzyskane na poszczególnych etapach wraz z procentowym udziałem każdego typu w całości.

Etap	AN	%	NA	%	NN(G)	%	Łącznie
bigramy	243422	46	147472	28	138680	26	529574
formy hasłowe	64124	37	37267	22	71543	41	172934
frekwencja > 10	683	33	773	37	611	30	**2067**

Tabela IV.8: Ilość wyników na poszczególnych etapach selekcji

Każde wyrażenie w zbiorze zostało ręcznie opracowane według zasad podanych w rozdziale II. Zgodnie z definicją przyjętą w książce za jednostki uznane zostały wszystkie wyrażenia, które uzyskały co najmniej 8 punktów. Tabela IV.9 zestawia przykładowe wyrażenia znajdujące się w tak utworzonym zbiorze w trzech grupach: wyrażenia ocenione najwyżej, wyrażenia o ocenie oscylującej w okolicach progu (czyli 8 punktów) i wyrażenia o najniższej punktacji. Znacznie więcej (500 najwyżej ocenionych) wyrażeń wielosegmentowych zaprezentowanych jest w Dodatku A.

	wyrażenie	ocena
najlepiej ocenione	duży ekran	38
	czarne chmury	30
	imię własne	30
	gaz cieplarniany	27
	ślepa próba	27
okolica progu	dziennik budowy	9
	komora grobowa	8
	szpital wojskowy	8
	para taneczna	8
	koncert skrzypcowy	7
najniżej ocenione	ofiara nazizmu	-6
	interpretacja prawa	-6
	względy komunikacyjne	-6
	możliwość kupna	-6
	czteropiętrowy blok	-6

Tabela IV.9: Zestawienie kandydatów na jednostki wraz z ocenami. Poniżej 8 punktów kandydat uznawany jest za zwykłe połączenie

[57] Warto zwrócić uwagę na fakt, że jeśli weźmie się pod uwagę wszystkie formy hasłowe, typ AN jest blisko dwukrotnie liczniejszy od typu NA, jednak ta proporcja wyraźnie się zmienia, gdy zastosowany zostanie filtr frekwencyjny (typ NA zawiera więcej form od AN). Sugeruje to, że typ NA jest sztywniejszy pod względem składniowym: jednostki o takiej konstrukcji są bardziej ustalone w języku, co powoduje, że więcej z nich przekracza barierę 10 wystąpień.

W wyniku opisanej wyżej procedury uzyskano listę jednostek wielosegmentowych, z których do typu AN zalicza się 144 (21% spośród wszystkich kandydatów tego typu), do typu NA 465 (60%), do typu NN(G) 213 (35%). Powyższe wyniki zawarto w Tabeli IV.10.

	AN	**%$_1$**	**%$_2$**	**NA**	**%$_1$**	**%$_2$**	**NN(G)**	**%$_1$**	**%$_2$**	**Łącznie**
Liczba jednostek	144	18	21	465	56	60	213	26	35	822

Tabela IV.10. Liczba faktycznych jednostek wielosegmentowych w zależności od typu. W kolumnie oznaczonej jako %$_1$ podany jest procentowy udział typu w całym zbiorze faktycznych jednostek; w kolumnie %$_2$ procent kandydatów uznanych za faktyczne jednostki w stosunku do wszystkich kandydatów danego typu (zatem jednostki typu AN stanowią 18% wszystkich jednostek, a z wszystkich kandydatów typu AN 21% okazuje się faktycznymi jednostkami leksykalnymi)

W Tabeli IV.10 warto zwrócić uwagę na bardzo wyraźną przewagę typu NA nad pozostałymi strukturami.

3.2. Rezultaty działania algorytmu

Przy ocenie skuteczności algorytmu ważne jest, aby pamiętać, że problem, z którym się on mierzy, jest – z punktu widzenia systemu automatycznego – jednym z najtrudniejszych zadań. Klasyfikacja wyrażenia jako jednostki albo swobodnego połączenia w dużej mierze nie zależy od jego kształtu czy cech lingwistycznych, decydują o tym głównie względy semantyczne i pozajęzykowe, do których algorytm nie ma dostępu. Dlatego też algorytmy tego typu niejako „z definicji" mają relatywnie niską skuteczność, przez co muszą być traktowane jako materiał do późniejszej pracy specjalistów.

W celu dokonania ewaluacji działania algorytmu zbadano rezultaty jego pracy na tym samym fragmencie korpusu *sample*, który posłużył jako podstawa do stworzenia zbioru porównawczego. Algorytm dokonał wyodrębnienia bigramów dla trzech struktur syntaktycznych: AN, NA i NN(G), przy czym w przypadku dwóch pierwszych struktur brał pod uwagę wyłącznie wyrażenia, których elementy łączył związek zgody. Na tym etapie dokonana została wstępna korekta związana z zastosowaną w korpusie źródłowym konwencją: korpus uznaje większość zaimków i liczebników za przymiotniki, co z punktu widzenia algorytmu jest zachowaniem niepożądanym.

Otrzymane bigramy zostały znormalizowane (sprowadzone do postaci hasłowej) i na tej podstawie opracowana została lista wyników zawierająca wszystkie unikalne formy hasłowe wraz z ich frekwencją i frekwencją poszczególnych członów.

W kolejnym kroku wyniki zostały poddane działaniu trzech filtrów: frekwencyjnego (odrzucone zostały wszystkie wyrażenia o frekwencji mniejszej niż 5), odrzucającego bigramy, które algorytm uznaje za będące częścią dłuższej jednostki, oraz stop-listy, zawierającej słowa, które nigdy lub niemal nigdy nie tworzą wielosegmentowych jednostek leksykalnych.

W ostatniej fazie działania algorytmu przeprowadzona została ocena kandydatów za pomocą klasyfikatora maszynowego uczenia (techniką SVM – *Support Vector Machines*).

Z uwagi na dużą „techniczność" ewaluacji wyników działania algorytmu duża część tabel i wykresów została umieszczona w Dodatku B. W zasadniczej części pracy akcent położony jest na prezentację materiału językowego.

Tabela IV.11 demonstruje fragment wyników uzyskanych przez algorytm dla typu składniowego NA (RZECZOWNIK + PRZYMIOTNIK). Tabela IV.12 zestawia 25 najlepszych wyników (uzyskanych przy posortowaniu list wynikowych za pomocą miary PMI) dla wszystkich trzech typów składniowych.

w_1	w_2	formy podstawowe	$f(w_1+w_2)$	$f(w_1)$	$f(w_2)$	PMI	PMI^3	NPMI	z	t	LLR	Dice
nawa	główna	nawa główny	56	180	8839	9.59	21.2	0.51	208	7.4	612	0.012
własne	sumienie	własny sumienie	44	12692	840	6.5	17.4	0.34	62.43	6.56	280	0.0065
forma	życia	forma życie	31	7598	21234	2.1	11.9	0.10	8.72	4.24	6714	0.002
młody	człowiek	młody człowiek	1319	13106	26555	6.3	27.1	0.45	327	35.8	8307	0.066
siły	zbrojne	siła zbrojny	236	7458	842	9.68	25.45	0.59	440	15.3	2607	0.057
dystych	elegijny	dystych elegijny	11	20	20	19.2	26.12	0.92	2578	3.31	278	0.55

Tabela IV.11: Przykładowe wyniki działania algorytmu dla typu NA, posortowane według miary PMI. Oznaczenia: w_1 – pierwszy wyraz; w_2 – drugi wyraz; f(x) – frekwencja wyrażenia x; PMI – informacja wzajemna; PMI^3 – informacja wzajemna sześcienna; NPMI – znormalizowawa informacja wzajemna; z – wskaźnik z; t – test t-Studenta; LLR – logarytm wskaźnika wiarygodności; Dice – współczynnik Dice'a

Typ AN		Typ NA		Typ NN(G)	
Hasło	**PMI**	**Hasło**	**PMI**	**Hasło**	**PMI**
przyszywana ciotka	15.05	przepuklina przeponowa	19.58	gęstość zaludnienia	14.78
papieska konfirmacja	13.39	krwawienie śródczaszkowe	19.32	gęstość elektronów	14.71
średniowieczna hymnografia	13.22	dystych elegijny	19.2	nietolerancja aspiryny	14.71
naoczny świadek	13.12	odma opłucnowa	18.62	obrzęk krtani	14.23
tylne siedzenie	13.05	plamka katodowa	17.15	jony argonu	14.21
niepełna grubość	13.00	symetria cylindryczna	17.06	tlenek azotu	13.97

Typ AN		Typ NA		Typ NN(G)	
Hasło	**PMI**	**Hasło**	**PMI**	**Hasło**	**PMI**
skórzana kurtka	12.96	blizny przerostowe	16.91	regeneracja naskórka	13.96
żelazna kurtyna	12.85	błona podstawna	16.6	zapalenie spojówek	13.91
płócienna czapka	12.63	endotoksyna bakteryjna	16.28	rak sutka	13.45
bezwarunkowa akceptacja	12.20	mediatorzy lipidowi	16.08	wyrzut sumienia	13.28
zdrowy rozsądek	12.15	regestra kancelaryjne	15.98	czubek katody	13.04
świetliste litery	11.85	wstrząs septyczny	15.94	skurcz oskrzeli	13
przewlekłe zapalenie	11.79	wiązka sygnałowa	15.92	częstość występowania	12.97
afgańscy mudżahedini	11.77	spektroskopia emisyjna	15.76	dwutlenek węgla	12.94
nieprzespana noc	11.52	astma oskrzelowa	15.72	niedrożność nosa	12.79
swobodne zwierciadło	11.48	jama brzuszna	15.51	komórki śródbłonka	12.74
próżna chwała	11.42	pęcherzyki płucne	15.3	mediatory zapalenia	12.69
wolne rodniki	11.38	alergeny wziewne	15.27	ułamek sekundy	12.59
intensywna terapia	11.38	tkanka ziarninowa	15.17	konserwator zabytków	12.55
fałszywe zeznanie	11.29	śródbłonek naczyniowy	15.11	zapalenie oskrzeli	12.43
swoiste przeciwciała	11.27	rozmnażanie płciowe	15.1	współczynnik tarcia	12.38
niniejsza monografia	11.26	zaimek dzierżawczy	15.07	interfejs użytkownika	12.35
wąska uliczka	11.26	naciek zapalny	15.07	spis abonentów	12.3
szkodliwy pokarm	11.21	łożysko naczyniowe	15	diagnostyka plazmy	12.19
wczesny ranek	11.19	strofa trójdzielna	14.99	współczynnik korelacji	12.08

Tabela IV.12: Zestawienie wyników algorytmu dla wszystkich trzech typów. Tabela przedstawia 25 haseł o najwyższej punktacji PMI (informacja wzajemna)

Interesujący jest fakt, że w obrębie tej samej listy uporządkowanej według innej miary pojawiają się dosyć istotne różnice, zwłaszcza jeśli brać pod uwagę sam szczyt listy. Tabela IV.13 zestawia pierwsze 25 haseł dla typu NA uzyskane dla różnych miar.

Frekwencja	NPMI	Dice
działalność gospodarcza	odma opłucnowa	odma opłucnowa
szkoła podstawowa	przepuklina przeponowa	przepuklina przeponowa
samorząd terytorialny	krwawienie śródczaszkowe	krwawienie śródczaszkowe
piłka nożna	dystych elegijny	dystych elegijny
wojna światowa	klatka schodowa	klatka schodowa
pomoc społeczna	wiązka sygnałowa	wiązka sygnałowa
związek zawodowy	wstrząs septyczny	klatka piersiowa
szkoła średnia	klatka piersiowa	papiery wartościowe
kodeks karny	spektroskopia emisyjna	wstrząs septyczny

Frekwencja	NPMI	Dice
środki finansowe	błona podstawna	samorząd terytorialny
papiery wartościowe	blizny przerostowe	ropa naftowa
akt prawny	alergeny wziewne	alergeny wziewne
sprawy wewnętrzne	symetria cylindryczna	przewód pokarmowy
dzień dzisiejszy	plamka katodowa	regestra kancelaryjne
postępowanie karne	papiery wartościowe	arcybiskup gnieźnieński
rok szkolny	pęcherzyki płucne	spektroskopia emisyjna
opinia publiczna	regestra kancelaryjne	symetria cylindryczna
sala gimnastyczna	ropa naftowa	pęcherzyki płucne
język obcy	arcybiskup gnieźnieński	piłka nożna
gospodarstwo rolne	mediatorzy lipidowi	plamka katodowa
służba wojskowa	samorząd terytorialny	błona podstawna
język angielski	astma oskrzelowa	niewydolność oddechowa
telefon komórkowy	jama brzuszna	telefon komórkowy
polityka zagraniczna	rana oparzeniowa	karabin maszynowy
rzecz jasna	niewydolność oddechowa	tworzywo sztuczne

Tabela IV.13: Porównanie uporządkowania listy wynikowej (dla typu NA) w zależności od zastosowanej miary

3.2.1. Miary ewaluacji

3.2.1.1. Dokładność i kompletność

Dokładność[58] (ang. *precision*) mierzy ilość poprawnych decyzji, jakie algorytm podjął przy uznawaniu wyrażeń za jednostki. Im jest wyższa, tym większą można mieć pewność, że wyrażenie znajdujące się na liście wynikowej jest faktycznie jednostką. Wyraża się to poniższym wzorem:

$$P = \frac{PP}{PP + NP} .$$

Kompletność (ang. *recall*) informuje, jak dużo jednostek faktycznie znajdujących się w badanym korpusie znalazło się na liście wyników stworzonej przez algorytm, według wzoru:

$$R = \frac{PP}{PP + NN} .$$

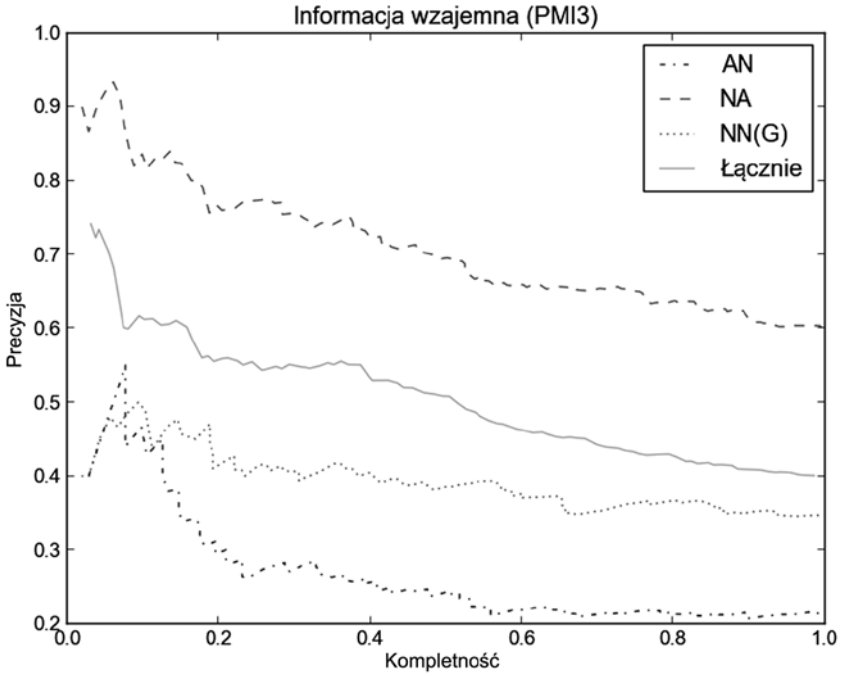
[58] Zob. uwagi dotyczące tłumaczenia w Rozdziale III.

Wyniki powyższych miar mogą się zmieniać wraz ze zmianą parametrów algorytmu (np. zmiany progu, powyżej którego wyrażenia uznawane są za jednostki). W zdecydowanej większości przypadków można zauważyć odwrotną korelację między tymi wielkościami: im większa dokładność, tym mniejsza kompletność i odwrotnie. Jest to zgodne z intuicją: precyzję można zwiększać przez uczynienie reguł wyodrębniania bardziej restrykcyjnymi, ale konsekwencją tego jest pomijanie niektórych faktycznych wyników pozytywnych, które są trudniejsze do identyfikacji. Omawiany algorytm daje w wyniku listę wszystkich połączeń, które z formalnego puntku widzenia mogą być jednostkami wielosegmentowymi. Lista ta jest uporządkowana według którejś z omawianych wyżej miar, im wyższa pozycja wyrażenia, tym wyższa – w założeniu – szansa, że wyrażenie to stanowi jednostkę. Listę tę można (i zazwyczaj daje to dobre rezultaty) skrócić poprzez ustalenie progu, poniżej którego wyrażenia są „odcinane". Należy zauważyć, że jeśli odcięcie nie zostanie dokonane (a zatem jako wynik traktowana jest cała lista), można się spodziewać kompletności równej 1 (lub zbliżonej do 1, jeżeli algorytm korzysta z filtrów typu stop-lista) i dosyć niskiej dokładności (na samym dole listy znajduje się więcej swobodnych połączeń niż jednostek). Z kolei zastosowanie progu powoduje zmniejszenie kompletności (część jednostek mieszcząca się na dole listy jest odcinana) i – najczęściej – wzrost dokładności. Typową praktyką przy omawianiu wyników jest w takim przypadku takie dobieranie progu, by otrzymać pewien z góry ustalony poziom kompletności i dla niego podawać poziom dokładności. Tabela IV.14 podaje dokładność algorytmu dla kompletności ustalonej na 0.3 (a więc fragmentu listy wynikowej, który zawiera 30% wszystkich faktycznych jednostek), 0.6 i zbliżonej do 1[59], przy uporządkowaniu listy wyników na podstawie czystej frekwencji i informacji wzajemnej (w wersji sześciennej), mierze dającej najlepsze rezultaty. Więcej szczegółowych zestawień znajduje się w Dodatku B.

	Frekwencja				**PMI³**			
	AN	**NA**	**NN(G)**	**Łącznie**	**AN**	**NA**	**NN(G)**	**Łącznie**
P (R ≈ 0.3)	0.20	0.71	0.56	0.50	0.27	0.79	0.55	0.55
P (R ≈ 0.6)	0.20	0.66	0.47	0.45	0.20	0.73	0.42	0.46
P (R ≈ 1)	0.21	0.60	0.35	0.40	0.21	0.60	0.35	0.40

Tabela IV.14: Wyniki poszczególnych miar ewaluacyjnych dla listy uporządkowanej na podstawie frekwencji i NPMI (znormalizowanej informacji wzajemnej). P – precyzja, R – kompletność

[59] Kompletność równa 1 oznacza w praktyce, że brana pod uwagę jest właściwie cała lista, zatem sposób jej uporządkowania nie wpływa na poziom dokładności – stąd identycze wartości dokładności dla frekwencji i dowolnej miary asocjacji. Uwaga ta dotyczy również miary F omawianej w następnym podpunkcie.



Wykres 2: Krzywe precyzji-kompletności dla danych posortowanych na podstawie informacji wzajemnej w wersji sześciennej (PMI3)

Dokładność i kompletność często zestawia się graficznie, za pomocą tzw. krzywych dokładności-kompletności (ang. *precision-recall curves*). Zmieniając miejsce odcięcia wyników (próg), tak by uzyskać kolejne poziomy kompletności (od najniższej możliwej do najwyższej), obliczamy jednocześnie precyzję w danym punkcie. Powstałe w ten sposób pary kompletność-dokładność możemy zobrazować na układzie współrzędnych, uzyskując łatwą do zinterpretowania linię. Wykres 2 przedstawia krzywe dokładności-kompletności dla wszystkich trzech typów składniowych z osobna i dla wszystkich razem, uzyskane na podstawie informacji wzajemnej w wersji sześciennej (PMI3) .

3.2.1.2. Miara F

Komplementarność dokładności i kompletności powoduje, że najczęściej wymieniane i omawiane są razem. Jednak to, że do oceny skuteczności używa się dwóch ocen naraz, jest niewygodne: za każdym razem trzeba przedstawiać je obie, nie do końca wiadomo, jak porównywać wyniki (czy np. algorytm o dwukrotnie mniejszej dokładności, ale dwa razy większej kompletności, jest lepszy czy gorszy?). Lepsza byłaby zatem jedna wielkość oceniająca, która łączyłaby cechy dokładności i kompletności. Taką miarą jest miara F (ang. *F-measure*). Może mieć ona różne wersje, w zależności od tego, czy precyzję i kompletność traktujemy jako równie ważne czy nie. Ogólna postać tej miary wyraża się poniższym wzorem:

$$F_{\beta} = (1 + \beta) * \frac{P * R}{\beta^2 * P + R},$$

gdzie β jest współczynnikiem wagi: dla $\beta > 1$ ważniejsza staje się kompletność, dla $\beta < 1$ bardziej liczy się dokładność. W przypadku najpopularniejszej wersji, przykładającej jednakową wagę do obu składników mamy:

$$F_1 = 2\frac{PR}{P + R}.$$

W praktyce najczęściej zamiast F_1 używa się symbolu F.

Wyniki algorytmu dla miary F_1, uzyskane na podstawie frekwencji i PMI (informacji wzajemnej) przedstawia Tabela IV.15.

	Frekwencja				**PMI³**			
	AN	**NA**	**NN(G)**	**Łącznie**	**AN**	**NA**	**NN(G)**	**Łącznie**
F (R ≈ 0.3)	0.24	0.43	0.40	0.36	0.27	0.79	0.55	0.55
F (R ≈ 0.6)	0.30	0.63	0.53	0.49	0.20	0.73	0.42	0.46
F (R ≈ 1)	0.35	0.75	0.52	0.55	0.35	0.75	0.52	0.55

Tabela IV.15: Miara F1 dla wyników algorytmu. R – kompletność, PMI³ – informacja wzajemna w wersji sześciennej

3.2.1.3. Przeciętna dokładność

Dokładność i kompletność są pomocne w przypadku, gdy listę wyników traktować jako binarną klasyfikację. Często jednak wygodnie jest przyjąć, że uporządkowanie listy jest istotne. Po pierwsze zajmujemy się wtedy całą listą, a nie tylko jej częścią znajdującą się powyżej arbitralnie ustalonego progu (który w zależności od potrzeb, a także definicji wielosegmentowej jednostki może się zmieniać), po drugie lista, w której rzeczywiste jednostki znajdują się wyżej, jest lepsza z punktu widzenia późniejszego jej opracowywania. W celu porównania dwóch uporządkowań tej samej listy można użyć krzywych dokładności-kompletności, jednak są one tylko narzędziem graficznym, które czasem jest trudne do porównywania. Z tego powodu lepiej użyć innej miary: przeciętnej dokładności (ang. *Average Precision*). Jest to średnia z dokładności uzyskanych dla każdego poziomu kompletności:

$$AP = \frac{1}{r}\sum_{i=1}^{n} x_i p_i, \quad p_m = \frac{1}{m}\sum_{k=1}^{m} x_k, \quad x_k \in \{0,1\},$$

gdzie r oznacza ilość faktycznych jednostek na liście, n oznacza długość listy, p_m to dokładność m pierwszych kandydatów na liście, x_k wskazuje, czy kandydat na pozycji k jest faktyczną jednostką ($x_k = 1$) czy nie ($x_k = 0$).

W Tabeli IV.16 zaprezentowana jest miara *AP* dla wyników algorytmu przy wykorzystaniu frekwencji i PMI (informacji wzajemnej).

	Frekwencja				PMI[3]			
	AN	**NA**	**NN(G)**	**Łącznie**	**AN**	**NA**	**NN(G)**	**Łącznie**
AP	0.23	0.69	0.49	0.48	0.28	0.74	0.50	0.52

Tabela IV.16: Miara AP (przeciętna dokładność) dla wyników algorytmu przy wykorzystaniu frekwencji i PMI (informacji wzajemnej)

3.3. Ewaluacja klasyfikatora uczącego się

Przy ocenie skuteczności algorytmu wykorzystującego maszynowe uczenie należy zwrócić uwagę na ważny szczegół. Ocena działania nie powinna być przeprowadzana na zbiorze użytym do trenowania, gdyż wtedy wynik ewaluacji byłby zawyżony. Jednocześnie należy zauważyć, że zbiory treningowy i testowy mają identyczną strukturę, co oznacza w praktyce, że część potencjalnego zbioru treningowego nie może być użyta do trenowania algorytmu. Metodą pozwalającą na ominięcie tego problemu jest tzw. walidacja krzyżowa (ang. *cross-validation*). Polega ona na podzieleniu zbioru treningowego na *n* równych podzbiorów. Jeden z nich używany jest jako zbiór testowy, pozostałe, połączone, jako zbiór treningowy. Procedurę tę powtarza się *n* razy, za każdym razem wybierając inny podzbiór. Uzyskane wyniki ewaluacji są uśredniane. Jeżeli do ewaluacji użyć przeciętnej dokładności, uśredniona miara nosi nazwę średniej przeciętnej dokładności (ang. *Mean Average Precision*).

Wyniki ewaluacji klasyfikatora SVM przedstawia Tabela IV.17.

	AN	**NA**	**NN(G)**	**Łącznie**
P (R ≈ 0.3) / (R ≈ 0.6)	0.45/ 0.33	0.84 / 0.76	0.55 / 0.49	0.74 / 0.60
F (R ≈ 0.3) / (R ≈ 0.6)	0.36 / 0.43	0.44 / 0.67	0.39 / 0.55	0.42 / 0.60
MAP	0.34	0.75	0.53	**0.65**

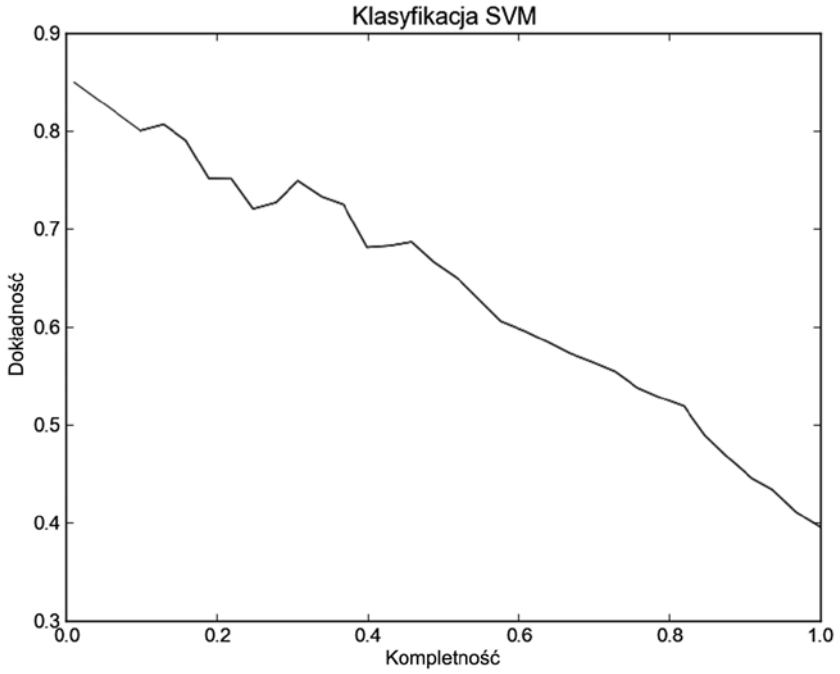
Tabela IV.17: Wyniki ewaluacji działania klasyfikatora SVM uzyskane na podstawie 5-krotnej walidacji krzyżowej

Powyższe zestawienie pokazuje bardzo wyraźną poprawę związaną z użyciem klasyfikatora wykorzystującego maszynowe uczenie. W stosunku do zwykłej frekwencji poprawa AP wynosi 17 punktów procentowych (relatywnie 35%), w stosunku do klasyfikatora opartego na pojedynczej mierze asocjacji jest to 13 punktów procentowych (relatywnie 25%). Tę poprawę można zauważyć również, analizując listę najlepiej ocenianych wyrażeń (Tabela IV.18 – kursywą zaznaczone są wyrażenia błędnie sklasyfikowane).

Hasło	Ocena	Hasło	Ocena
zdanie proste	2.058	kość słoniowa	1.11
kwiatek czysty	1.564	jama brzuszna	1.099
miejsce pracy	1.487	astma oskrzelowa	1.097
msza święta	1.430	równowaga termodynamiczna	1.077
pomoc społeczna	1.399	spektroskopia emisyjna	1.076
jelito grube	1.398	tkanka ziarninowa	1.066
przewód pokarmowy	1.384	praca zawodowa	1.061
kodeks karny	1.359	naczynia włosowate	1.059
ropa naftowa	1.349	błona podstawna	1.057
węzeł chłonny	1.292	pan młody	1.047
akt prawny	1.259	chrzest święty	1.041
komórka tuczna	1.253	mięsień sercowy	1.039
pęcherzyki płucne	1.238	telefon komórkowy	1.037
niewydolność oddechowa	1.236	szkoła podstawowa	1.01
zakład pracy	1.205	samorząd terytorialny	1.01
gospodarstwo rolne	1.178	lata pracy	1.01
wojna światowa	1.161	zaimek względny	1.01
endotoksyna bakteryjna	1.145	związek zawodowy	1.01
blizna przerostowa	1.138	praca naukowa	1.01
sala gimnastyczna	1.135	stan zapalny	1.01
dystych elegijny	1.134	dzień dobry	1.009
naciek zapalny	1.133	sakramenty święte	1.009
tworzywo sztuczne	1.131	rok liturgiczny	1.009
arcybiskup gnieźnieński	1.127	karabin maszynowy	1.009
alergeny wziewne	1.112	kość słoniowa	1.11

Tabela IV.18: 50 najwyżej ocenionych przez klasyfikator SVM wyrażeń

Jak łatwo zauważyć, przeważająca część najwyżej sklasyfikowanych wyrażeń należy do typu NA. Nie jest to zaskakujące – jako że w obrębie tego typu wyrażenia zdecydowanie częściej okazują się faktycznymi jednostkami, klasyfikator „nauczył się" przyznawać takim kandydatom wyższą punktację. Z tej przyczyny wartości łączne widniejące w Tabeli IV.19 są wyższe niż wynikałoby to ze średniej ważonej poszczególnych typów – w związku z tym, że na początku listy wynikowej jest najwięcej wyrażeń NA, łączna ocena skuteczności ekstrakcji jest zwiększona. W dodatku B można znaleźć rozszerzone listy wynikowe.



Wykres 3: Krzywa dokładności-kompletności dla klasyfikatora SVM

Ocena skuteczności klasyfikatora SVM pozwoliła także na ustalenie, jaki wpływ na jakość wyników mają poszczególne cechy brane pod uwagę przez klasyfikator. Najważniejszą cechą okazał się typ składniowy, w dalszej kolejności były to frekwencja pierwszego członu wyrazu i liczba jego substytucji – a więc cechy czysto lingwistyczne. Miary statystyczne pod względem ważności uplasowały się na dalszych miejscach. Interesujący okazał się wpływ badania homonimiczności członów wyrażenia: o ile uwzględnienie liczby homonimicznych lematów dawało pozytywne rezultaty, o tyle wzięcie pod uwagę liczby homonimicznych form powodowało lekkie obniżenie skuteczności algorytmu.

3.4. Omówienie wyników

Zasadnicza część wyników (zwłaszcza pełne porównanie rezultatów uzyskanych za pomocą zastosowania różnych miar) znajduje się w Dodatku B, dlatego uwagi zawarte w tej sekcji warto skonsultować z danymi tam zamieszczonymi.

Analiza wyników nasuwa kilka wniosków. Po pierwsze, rzuca się w oczy istotna przewaga typu NA nad pozostałymi typami (zarówno jeśli chodzi o skuteczność wykrywania, jak i liczbę wykrytych jednostek) – najlepiej to widać na Wykresie 2. Po drugie, na tle miar asocjacji zwykła frekwencja sprawuje się gorzej, ale nie jest to różnica bardzo duża (a w niektórych sytuacjach przewaga leży wręcz po stronie frekwencji). Najbardziej obiektywna miara, AP, pokazuje, że największa poprawa skuteczności algorytmu uzyskana dzięki użyciu miar asocjacji wynosi 4 punkty procentowe (relatywna poprawa wynosi 14%). Potwierdza to wnioski zawarte w niektórych pracach poświęconych automatycznej

ekstrakcji wyrażeń wielowyrazowych (np. Krenn i Evert 2001). Trzecia rzecz warta odnotowania to fakt, że w przypadku różnych typów najlepiej sprawdzają się różne miary asocjacji. Z konstrukcjami zawierającymi przymiotnik najlepiej radzą sobie różne wersje informacji wzajemnej (zwłaszcza w wersji sześciennej) i z-score, podczas gdy t-score daje zauważalnie gorsze rezultaty. Ta ostatnia miara lepiej natomiast się sprawdza w przypadku typu NN(G) (mimo to żadna z miar nie jest lepsza w tym przypadku niż zwykła frekwencja). Ta obserwacja również potwierdza wnioski zawarte w niektórych pracach (por. Krenn 2000, Evert 2005). Kolejny wniosek: PMI, PMI3, NPMI i z-score dają bardzo zbliżone wyniki.

Jak wskazują rezultaty ewaluacji, obiektywna skuteczność algorytmu jest przeciętna i raczej nie można go traktować jako narzędzia umożliwiającego pełną automatyzację. Nie stanowi to zaskoczenia: we wstępnej części rozdziału zaznaczono, że problem automatycznej ekstrakcji jednostek wielosegmentowych jest bardzo trudny do automatycznego przetwarzania; skuteczność innych współczesnych algorytmów jest porównywalna.

Ewaluacja pokazuje, że skuteczność tradycyjnych metod ekstrakcji polegająca na stosowaniu miar asocjacji nie potrafi przekroczyć pewnego, stosunkowo niskiego progu. Wyraźny potencjał tkwi natomiast w bardziej zaawansowanej metodzie, jaką jest uczenie maszynowe. Pecina (2009) odnotowuje analogiczny, ponaddwudziestoprocentowy wzrost relatywnej skuteczności metod maszynowego uczenia się w stosunku do pojedynczych miar asocjacji. Pomiędzy podejściem tam przyjętym a metodami stosowanymi w niniejszej książce, istnieje pewna koncepcyjna różnica. Algorytm proponowany przez Pecinę opiera się na łączeniu dużej ilości różnych miar asocjacji (w pierwotnej ich liczba wynosi 82, w ostatecznej 17), a więc w praktyce wyłącznie na informacjach statystycznych, podczas gdy algorytm omawiany tutaj łączy dane statystyczne z koncepcjami lingwistycznymi.

Ważną, zarówno pod względem teoretycznym, jak i praktycznym, kwestią jest odnotowana wyżej przewaga typu składniowego NA nad pozostałymi. Przewaga ta wynika z tego, że taka konstrukcja syntaktyczna – choć całkowicie poprawna z gramatycznego punktu widzenia – jest w pewien sposób nacechowana i używana przeważnie wtedy, gdy dane wyrażenie z jakiegoś powodu – najczęściej semantycznego – powinno się odróżniać od zwykłego połączenia. Jest to zatem typ niejako „naturalny" dla imiennych jednostek nieciągłych.

3.4.1. Najczęstsze przyczyny błędów

Błędy popełniane przez algorytm mają rozmaite przyczyny. Analiza rezultatów działania systemu ekstrakcji pokazuje, że istnieją dwa główne powody nieprawidłowej klasyfikacji:

a) duża ogólność połączenia wyrazów: wyrażenia typu *istotna zmiana* czy *inna możliwość* mają mało sprecyzowane znaczenie, które pozwala na wykorzystywanie ich w wielu sytuacjach, a w konsekwencji zwiększa frekwencję występowania,
b) popularność wyrażenia: połączenia takie jak *biedne dziecko* czy *własne mieszkanie* odnoszą się do tematów dosyć często poruszanych, a cechuje je przy tym pewna (choć niewielka) doza ustalenia, powodująca, że mówiącemu nasuwa się właśnie taka, a nie inna, ekwiwalentna, konstrukcja (można by mówić o odtwarzalności takich związków).

Dużo rzadziej występują błędy wynikające z nieprawidłowej anotacji segmentu w korpusie czy nieprawidłowego działania algorytmu (takim błędem jest np. ustalenie formy hasłowej wyrażenia *czas wolny* na *czas wolen*).

3.5. Testy na innym korpusie

Działanie algorytmu przetestowano również na korpusie niezawierającym anotacji. Do tego celu wybrano korpus notatek prasowych Polskiej Agencji Prasowej. Korpus zawiera około 51 tysięcy krótkich artykułów, w sumie zawiera około 3 milionów wyrazów. Ze względu na brak odpowiedniego zbioru testowego nie ma możliwości przeprowadzenia pełnej ewaluacji, policzono natomiast dokładność dla stu najwyżej sklasyfikowanych wyrażeń, wynosi ona 69%. Tabela IV.19 zestawia dwadzieścia wyrażeń z początku listy (kursywą oznaczone są wyrażenia niestanowiące jednostki). Tabela zawierająca wszystkie 100 wyrażeń znajduje się w dodatku A.

konferencja prasowa	mistrz świata
wybory prezydenckie	minister sprawiedliwości
wybory parlamentarne	strona internetowa
sekretarz generalny	kurs akcji
kampania wyborcza	komitet wyborczy
rzecznik prasowy	partia polityczna
projekt ustawy	stopa procentowa
minister obrony	związek zawodowy
wojna światowa	opinia publiczna
nowelizacja ustawy	sekretarz stanu

Tabela IV.19: 20 najwyżej ocenionych bigramów w korpusie PAP

4. Uwagi dotyczące implementacji

Pierwotna wersja algorytmu została zaimplementowana w języku programowania C++, ostateczna w języku Python. Język ten został wybrany ze względu na dużą przejrzystość, łatwość i szybkość pisania kodu oraz dostępność wielu bibliotek.

Python jest językiem skryptowym, nie wymaga kompilacji, tak więc pliki z kodem źródłowym są zwykłymi plikami tekstowymi. Każdy użytkownik może dzięki temu zmieniać parametry startowe programu.

Do uruchomienia programu niezbędny jest zainstalowany w systemie interpreter Pythona w wersji 2.7 (można go pobrać ze strony *https://www.python.org/download/releases/2.7/*), a także dodatkowe biblioteki:

- numpy (umożliwiająca zaawansowane przetwarzanie matematyczne), dostępna pod adresem *http://www.scipy.org/scipylib/download.html* ,
- sklearn (niezbędna do korzystania z klasyfikatora opartego na metodach maszynowego uczenia), dostępna pod adresem *http://sourceforge.net/projects/scikit-learn/files/* .

Uruchomienie programu do ekstrakcji następuje za pomocą pliku *szukajWsg.py*. W pliku *wsg.py* zdefiniowane są struktury składniowe, których ma poszukiwać program. Można je modyfikować i dodawać własne poprzez modyfikację zmiennej lista_struktur.

Rezultatem działania programu są pliki tekstowe, których jest tyle, ile struktur syntaktycznych zostało zdefiniowanych. Przykładowo, znalezione jednostki wielosegmentowe o strukturze AN zapisywane są do pliku AN.txt, w kolejności od najlepiej ocenionych. Pliki wynikowe zapisywane są w podkatalogu ekstrakcja/wyniki katalogu głównego.

V
Wnioski końcowe

1. Możliwości rozwoju

Przedstawiona koncepcja jednostki wielosegmentowej opracowana została z myślą o reprezentacji za jej pomocą wszystkich typów jednostek. Badania skupiały się jednak przede wszystkim na jednostkach imiennych, pełniących rolę rzeczownika lub przymiotnika, oraz na jednostkach, w których skład wchodzi przysłówek. Przydatne byłoby głębsze zbadanie – i ewentualne dopracowanie – definicji w odniesieniu do innych typów, takich jak np. czasowniki.

Proponowany algorytm ekstrakcji wyrażeń wielowyrazowych najskuteczniejszy jest dla bigramów (jednostek dwuelementowych), ze względu na stosowanie w takim przypadku miar asocjacji. Istnieją sposoby umożliwiające dostosowanie miar asocjacji do modelu ogólniejszego, w którym długość badanego wyrażenia nie jest ograniczona. Sposoby te są problematyczne zarówno z teoretycznego (nie opierają się na silnych fundamentach matematycznych), jak i praktycznego punktu widzenia (złożoność obliczeniowa algorytmu – a więc czas jego działania – gwałtownie wzrasta). Pomimo tych zastrzeżeń próba ich zastosowania byłaby godnym uwagi eksperymentem.

Możliwości rozwoju algorytmu związane są także z samymi miarami asocjacji. Zastosowanych w książce sześć miar jest powszechnie wykorzystywanych w ekstrakcji wyrażeń wielowyrazowych, jednak w literaturze można znaleźć ich znacznie więcej; istnieją również modyfikacje miar zastosowanych, jak np. ich asymetryczne wersje.

Zastosowanie w algorytmie metod maszynowego uczenia dało w rezultacie zauważalne zwiększenie skuteczności ekstrakcji. Ten postęp staje się dużo wyraźniejszy, gdy algorytm ma do dyspozycji nie tylko dane czysto statystyczne, ale również informacje wynikające z cech językowych wyrażenia, takich jak typ składniowy, ilość substytucji (będących w zamierzeniu odzwierciedleniem łączliwości wyrazów wchodzących w skład potencjalnej jednostki) i stopień homonimiczności. Uzyskane rezultaty sugerują, że dołączenie kolejnych danych (np. o stopniu asumaryczności znaczenia czy stałości składniowej) zwiększyłoby jakość algorytmu jeszcze bardziej.

Zastosowana w algorytmie technika maszynowego uczenia (*Support Vector Machine* – SVM) jest jedną z najskuteczniejszych i chętnie wykorzystywanych. Nie znaczy to jednak, że nie istnieją alternatywy: interesujące byłyby eksperymenty ze skutecznością sieci neuronowych czy klasyfikatorów Bayesowskich.

Zbiór porównawczy powstały na potrzeby niniejszej książki pozostawia spore pole do rozwoju. W obecnej postaci zawiera on tylko bigramy należące do trzech typów składniowych (PRZYMIOTNIK + RZECZOWNIK, RZECZOWNIK + PRZYMIOTNIK, RZECZOWNIK + RZECZOWNIK W DOPEŁNIACZU). Został utworzony na podstawie fragmentu (około jednej trzeciej) podkorpusu-próbki *sample*, liczącego w całości około 15 milionów segmentów. Aby uzyskać pełnowartościowy zbiór porównawczy, należałoby istniejący rozszerzyć o inne typy składniowe (zwłaszcza obejmujące czasowniki).

Problem wielosegmentowych jednostek leksykalnych poruszany w książce jest bardzo rozległy i – jak widać na podstawie powyższej sekcji – ma wiele możliwości rozwoju.

2. Wskazówki praktyczne

Przygotowany na potrzeby pracy zbiór porównawczy może stać się wygodnym narzędziem do rozbudowywania różnorakich baz leksykalnych, a także do projektowania, trenowania i ewaluacji algorytmów wykorzystujących jednostki wielosegmentowe. Dodatek A prezentuje 300 jednostek ze zbioru porównawczego, które uzyskały najwyższą punktację. W dodatku B przedstawione są szczegółowe wyniki działania algorytmu ekstrakcji w postaci tabel i wykresów. Dodatek C zawiera praktyczne wskazówki dotyczące liczenia miar asocjacji dla danych językowych.

Bibliografia

Słowniki

Słownik języka polskiego. (2007). Drabik, L., Sobol, E. (oprac.). Warszawa, PWN
Słownik języka polskiego PAN. (1969). Doroszewski, W. (red.). t. 1-11. Warszawa, PWN
Słownik języka polskiego PWN. (1988). Szymczak, M. (red.). t. 1-3. Warszawa, PWN
Słownik poprawnej polszczyzny PWN. (1997). Doroszewski, W. (red.). Warszawa, PWN
Słownik współczesnego języka polskiego. (1996). Dunaj, B. (red.). Warszawa, Wilga

Literatura cytowana

Attia, M., Toral, A., Tounsi, L., Pecina, P., van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions. W: *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Pekin, 18-26

Baldwin, T. (2006). Compositionality and Multiword Expressions: Six of One, Half a Dozen of the Other? W: *Invited Talk at theCOLING/ACL 2006 Workshop on Multiword Expressions*, Sydney

Baldwin, T., Bannard, C., Tanaka T., Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. W: *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.* Sapporo, 89-96

Baldwin, T., Kim, S.N. (2010). Multiword Expressions. W: Indurkhya, N., Damerau, F. (red.). *Handbook of Natural Language Processing.* Boca Raton, CRC Press. 267-292

Baldwin, T., Villavicencio, A. (2002). Extracting the Unextractable: A Case Study on Verb--particles. W: *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*. Taipei, 98-104

Bannard, C., Baldwin, T., Lascarides, A. (2003). A Statistical Approach to the Semantics of Verb-Particles. W: *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.* Sapporo, 65-72

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

Bański, P., Moszczyński, R. (2008). Enhancing an English-Polish electronic dictionary for multiword expression research. W: *Studia kognitywne*, t. 8, 288-302

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Cooccurrence*. Tybinga, Gunter Narr Verlag

Basaj, M. (1982). Ekwiwalencja tłumaczeń frazeologizmów. W: Basaj, M., Rytel, D. (red.). *Z problemów frazeologii polskiej i słowiańskiej*, t. 1. Wrocław, 157-166

Bauer, L. (1983). *English Word-formation.* Cambridge, Cambridge University Press

Bąba, S., Liberek, J. (2001). *Słownik frazeologiczny współczesnej polszczyzny.* Warszawa

Berry-Rogghe, G. (1973). The Computation of Collocations and their Relevance in Lexical Studies. W: Aitken, A., Bailey, R., Hamilton-Smith, N. (red.). *The Computer and Literary Studies.* Edynburg, 103-112

Berry-Rogghe, G. (1974). Automatic Identification of Phrasal Verbs. W: Mitchell, J. (red.). *Computers in the Humanities*, Edinburgh University Press, Edynburg, 16-26

Bień, J. (1991). *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji.* Rozprawy Uniwersytetu Warszawskiego, t. 383. Warszawa, Wydawnictwa Uniwersytetu Warszawskiego

Bień, J., Woliński, M. (2003). Wzbogacony korpus „Słownika frekwencyjnego polszczyzny współczesnej". W: Linde-Usiekniewicz, J., Huszcza, R. (red.). *Prace językoznawcze dedykowane Profesor Jadwidze Sambor*. Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa, 6-10.

Blaheta, D., Johnson, M. (2001). Unsupervised learning of multi-word verbs. W: *Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations.* Tuluza, 54-60

Blancafort, H., Daille, B., Gornostay, T., Heid, U., Mechoulam, C., Sharoff, S. (2010). TTC. Terminology Extraction, Translation Tools and Comparable Corpora. W: *Proceedings, 14th EURALEX International Congress.* Leeuwarden, 263-268

Bogusławski, A. (1976). O zasadach rejestracji jednostek języka. *Poradnik Językowy*, *z. 8*, 356-364

Bogusławski, A. (1987). Obiekty leksykograficzne a jednostki języka. W: Saloni, Z. (red.). *Studia z polskiej leksykografii współczesnej*, t. 2, Białystok, 13-34

Bogusławski, A. (1989). Uwagi o pracy nad frazeologią. W: Saloni, Z. (red.). *Studia z polskiej leksykografii współczesnej*, t. 3, Białystok, 13-30

Bogusławski, A., Wawrzyńczyk, J. (1993). *Polszczyzna jaką znamy. Nowa sąda słownikowa.* Warszawa

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. W: *Proceedings of the Biennial GSCL Conference,* 31-40

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. W: *Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, 977-981

Breidt, E. (1993). Extraction of V-N-collocations from text corpora: A feasibility study for German. W: *Proceedings of theWorkshop on Very Large Corpora: Academic and Industrial Perspectives.* Columbus, OH, 74-83

Buczyński, A. (2004). *Pozyskiwanie z Internetu tekstów do badań lingwistycznych*. Praca magisterska, Uniwersytet Warszawski

Bungum, L., Gambäck, B., Lynum, A., Marsi, E. (2013). Improving Word Translation Disambiguation by Capturing Multiword Expressions with Dictionaries. W: *The 9th Workshop on Multiword Expressions (MWE)*. Atlanta, GA, Association for Computational Linguistics, 21-30

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. W: *Proceedings of LREC*. Las Palmas, 934-1940

Caseli de, H., Ramisch, C., Nunes, M., Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language resources and evaluation, 44(1-2)*, 59-77

Chlebda, W. (1991). *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy.* Opole, WSP

Chlebda, W. (2001). Frazematyka. W: Bartmiński, J. (red.). *Współczesny język polski*, Wydawnictwo UMCS, Lublin, 335-342

Choueka, Y., Klein, S., Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, *4(1)*, 34-38

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. W: *Proceedings of the RIAO Conference on User--Oriented Content-Based Text and Image Handling.* Cambridge, 609-624

Church, K. (1980). *On memory limitations in natural language processing*. Praca magisterska. Uniwersytet MIT.

Church, K., Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics, 16(1)*, 22-29.

Church, K., Gale, W., Hanks, P., Hindle, D. (1991). Using Statistics in Lexical Analysis. W: Zernik U. (red.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Hillsdale, NJ, Lawrence Erlbaum, 115-164

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., Flickinger, D. (2010). Muliword Expressions: linguistic precission and reusability. W: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, 1941-1947

Czerepowicka, M. (2011). „Toposław" jako narzędzie znakowania jednostek wieloczłonowych. W: Matusiak-Kempa, I., Przybyszewski, S. (red.). *Nowe zjawiska w języku, tekście, komunikacji. Kontekst a komunikacja.* Olsztyn, 28-35.

Dagan, I., Church, K. (1994). Termight: Identifying and translating technical terminology. W: *Proceedings of the 4th ANLP Conf. (ANLP 1994).* Stuttgart, ACL, 34-40

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. W: Klavans, I., Resnik, P. (red.), *The Balancing Act. Combining symbolic and statistical approaches to language*, t. 1. Cambridge, MIT Press, 49-66

Dias, G. (2003). Multiword Unit Hybrid Extraction. W: *Workshop on Multiword Expressions of the 41st ACL meeting*, 41-48

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19(1)*, 61-74

Edmonds, Ph. (1997). Choosing the word most typical in context using a lexical cooccurrence network. W: *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1997).* Madryt, 507-509

Evans, D., Ginther-Webster, K., Hart, M., Lefferts, R., Monarch, I.(1991). Automatic indexing using selective nlp and first-order thesauri. W: *Proceedings of the RIAO*, t. 2, 624-643

Evert, S. (2005). *The Statistics of Word Cooccurrences. Word Pairs and Collocations.* Praca doktorska. Uniwersytet w Stuttgarcie

Evert, S. (2008). Corpora and collocations. W: Lüdeling, A., Kytö, M. (red.). *Corpus Linguistics. An International Handbook.* Mouton de Gruyter. Berlin

Evert, S., Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, *19(4)*,450-466

Fazly, A., Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. W: Grégoire, N., Evert, S., Kim, S.N. (red.). *Proceedings of the ACL Workshop on A Broader Perspective on MWEs (MWE 2007).* Praga, 9-16

Ferret, O. (2002). Using collocations for topic segmentation and link detection. W: *Proceedings of COLING 2002*. Taipei, 1-7

Finlayson, M., Kulkarni, N. (2011). Detectiong multi-word expressions improves word sense disambiguation. W: Kordoni i in. (red.). *Proceedings of the ACL Workshop on MWEs: form Parsing and Generation to the Real World.* Portland, 20-24

Firth, J. (1957). A synopsis of linguistic theory 1930-1957. W: *Studies in Linguistic Analysis*. Oxford, 1-32

Francis, W., Kucera, H. (1964). *Manual of information to accompany a standard sample of present- day edited American english, for use with digital computers.* Providence, RI, Department of Linguistics, Brown University

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multiword terms: the C- value/NC-value method. *International Journal on Digital Libraries*, *3(2)*, 115-130

Gajęcki, M. (2009). Słownik fleksyjny jako biblioteka języka C. W: Lubaszewski, W. (red.). *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* Kraków, Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH

Gayen, V., Sarkar, K. (2013). Automatic Identification of Bengali Noun-Noun Compounds Using Random Forest. W: *The 9th Workshop on Multiword Expressions (MWE)*. Atlanta, GA, Association for Computational Linguistics, 64-72

Geffroy, A., Lafon, P., Seidel, G., Tournier, M. (1973). Lexicometric Analysis of Co-occurrences. W: Aitken, A., Bailey, R., Hamilton-Smith, N. (red.). *The Computer and Literary Studies.* Edynburg, Edinburgh University Press

Goldman, J., Nerima, L., Wehrli, E. (2001). Collocation extraction using a syntactic parser. W: *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*. Tuluza, 61-66

Graliński, F., Savary, A., Czerepowicka, M., Makowiecki, F. (2010). Computational lexicography of multi-word units: How efficient can it be? W: *Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010).* Pekin, 2-10

Grégoire, N. (2009). *Untangling Multiword Expressions. A study on the representation and variation of Dutch multiword expressions.* Praca doktorska. Uniwersytet w Utrechcie.

Green, S., de Marneffe, M., Bauer, J., Manning, C. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with french. W: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 725-735

Grochowski, M. (1982). *Zarys leksykologii i leksykografii. Zagadnienia synchroniczne.* Toruń

Halliday, M.A.K. (1966). Lexis as a linguistic level. W: Bazell, C. E., Catford, J. C., Halliday, M.A.K., Robins, R. (red.). *In memory of J. R. Firth*. Londyn, 148-162

Hardcastle, D. (2005). Using the Distributional Hypothesis to Derive Cooccurrence Scores from the British National Corpus. W: *Proceedings of Corpus Linguistics.* Birmingham

Harris, Z. S. (1962). *String Analysis of Sentence Structure.* Mouton, The Hague

Heid, U., Gojun, A. (2012). Term candidate extraction for terminography and CAT. An overview of TTC. W: *Proceedings of the 15th EURALEX International Congress*. Oslo, 585-594

Heid, U., Weller, M. (2010). Corpus-derived data on German multiword expressions for lexicography. W: *Proceedings of the Euralex International Congress.* Leeuwarden, 331-340

Hockett, C. F. (1958). A *Course in Modern Linguistics*. New York, Macmillan

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, MA, USA. MIT Press

Janus, D., Przepiórkowski, A. (2007). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. W: Walinski, J., Kredens, K., Gozdz-Roszkowski, S. (red.). *The proceedings of Practical Applications in Language and Computers PALC 2005.* Frankfurt nad Menem, Peter Lang

Jurafsky, D., Martin, J. (2008). *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, Prentice Hall

Justeson, J., Katz, S. (1995a). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics, 21*, 1-27

Justeson, J., Katz, S. (1995b). Technical terminology: some linguistic properties and an algorithm for identification in text. W: *Natural Language Engineering, 1*, 9-27

Kaplan, R., Kay, M. (1981). Phonological rules and finite-state transducers. W: *Linguistic Society of America Meeting Handbook, Fifty-Sixth Annual Meeting*. Nowy Jork, 27-30

Karlsson, F. (2008). Early generative linguistics and empirical methodology. W: Lüdeling, A., Kytö, M. (red.). *Corpus Linguistics. An International Handbook*. Berlin, 15-30

Katz, G., Giesbrecht, E. (2006). Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis. W: *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12-19

Khokhlova, M. (2008). Extracting Collocations in Russian: Statistics vs. Dictionary. W: *JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles*, 613-624.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. *Computational Linguistics, 29 (3)*, 333-347

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The sketch engine. W: *Proceedings of the 11th EURALEX International Congress*. Lorient, 105-116

Kiss, T., Strunk, J. (2002). Viewing sentence boundary detection as collocation identification. W: Busemann, S. (red.). *Proceedings of KONVENS 2002.* Saarbrücken, 75-82

Kita, K., Kato, Y., Omoto, T., Yano, Y. (1994). Automatically Extracting Collocations from Corpora for Language Learning. W: Wilson, A., McEnera, A. (red.). *UCREL Technical Papers*, t. 4, *Corpora in Language Education and Research, A Selection of Papers from Talc94 1 (1)*, 21-33

Kjellmer, G. (1994). A *Dictionary of English Collocations.* Claredon Press, Oxford

Kopotev, M., Pivovarova, L., Kochetkova, N., Yangarber, R. (2013). Automatic Detection of Stable Grammatical Features in N-Grams. W: *The 9th Workshop on Multiword Expressions (MWE)*. Atlanta, GA, Association for Computational Linguistics, 73-81

Korzycki, M. (2008). *Transducer skończenie stanowy jako narzędzie rozpoznawania form tekstowych wyrazów polskich.* Praca doktorska. Akademia Górniczo-Hutnicza, Kraków.

Kosek, I. (2001). Z zagadnień leksykalizacji i derywacji: wyrażenia z segmentem na. W: Biolik, M. (red.). *Prace językoznawcze Uniwersytetu Warmińsko-Mazurskiego w Olsztynie, z. III*, 77-86

Kosek, I. (2008). *Fleksja i składnia nieciągłych imiennych jednostek leksykalnych.* Olsztyn, UWM

Krčmář, L., Ježek, K., Pecina, P. (2013). Determining Compositionality of Word Expressions Using Word Space Models. W: *The 9th Workshop on Multiword Expressions (MWE)*. Atlanta, GA, Association for Computational Linguistics, 42-50

Krčmář, L., Ježek, K., Poesio, M. (2012). Detection of semantic compositionality using semantic spaces. W: *Lecture Notes in Computer Science 7499*, LNAI, 353-361

Krenn, B. (2000). Collocation mining: Exploiting corpora for collocation identification and representation. W: *Proceedings of KONVENS 2000*. Ilmenau, 209-214

Krenn, B., Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. W: *Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations.* Tuluza, 39-46

Krstev, C., Stanković, R., Obradović, I., Vitas, D., & Utvić, M. (2010). Automatic construction of a morphological dictionary of multi-word units. W: *Advances in Natural Language Processing*. Springer. Berlin Heidelberg, 226-237

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. W: *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 17-22

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków

Landauer, T., Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2)*, 211-240

Lareau, F., Dras, M., Börschinger, B., Dale, R. (2011). Collocations in Multilingual Natural Language Generation: Lexical Functions meet Lexical Functional Grammar. W: *Proceedings of the 2011 Australasian Language Technology Workshop*. Canberra, 95-104

Lewicki, A.M., (1982). Problemy opracowania słownika frazeologicznego. *Biuletyn Slawistyczny*, t. 7, 5-25

Lewicki, A.M. (1986). Składnia związków frazeologicznych. *Biuletyn Polskiego Towarzystwa Językoznawczego*, t. 40, 75-83

Lewicki, A.M., Pajdzińska, A. (2001). Frazeologia, W: Bartmiński, J. (red.). *Encyklopedia kultury polskiej XX wieku*, t. II. Lublin

Lewicki, A.M., Pajdzińska, A., Rejakowa, B. (1987). Z *zagadnień frazeologii: problemy leksykograficzne*. Warszawa, PWN

Lin, D. (1998). Extracting Collocations from Text Corpora. W: *Proceedings of the First Workshop on Computational Terminology.* Montreal, 57-63

Lin, D. (1999). Automatic identification of non-compositional phrases. W: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.* Morristown, NJ, 317-324

Lubaszewski, W. (2009). Wyraz. W: Lubaszewski, W. (red.). *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* Kraków, Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 15-36

Lubaszewski, W., Wróbel, H., Gajęcki, M., Moskal, B., Orzechowska, A., Pietras, P., Pisarek, P., Rokicka T. (2001). *Słownik fleksyjny języka polskiego.* Kraków

Lü, Y., Zhou, M. (2004). Collocation Translation Acquisition Using Monolingual Corpora. W: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*. Barcelona, 167-174

Lyons, J. (1976). *Wstęp do językoznawstwa.* Warszawa, PWN

Miller, G.A. *WordNet: A Lexical Database for English.* Communications of the ACM, t. 38, nr 11, 39-41

Łaziński, M. (2000). *Korpus PWN*. Warszawa, ISJP, LVII-LVIII

Manning, C., Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press. Cambridge

Maynard, D., Ananiadou, S. (1999). Identifying Contextual Information for Multi-Word Term Extraction. W: *Proceedings of 5th International Congress on Terminology and Knowledge Engineering (TKE)*, 212-221

Maziarz, M., Piasecki, M., Szpakowicz, S. (2012). Approaching plWordNet 2.0. W: *Proceedings of the 6th Global Wordnet Conference*. Matsue, Japonia

McCarthy, D., Keller, B., Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. W: Bond, F., Korhonen, A., McCarthy, D., Villavicencio, A. (red.). *Proceedings of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*. Sapporo, ACL, 73-80

McKeown, K., Radev, D. (2000). Collocations. W: Dale, R., Moisl, H., Somers, H. (red.). *A Handbook of Natural Language Processing*. Marcel Dekker, 507-522

Michelbacher, L., Evert, S., Schütze, H. (2007). Asymmetric association measures. W: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*. Borovetz, 1-6

Mielczuk, I., Clas, A., Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Editions Duculot, Louvain la Neuve

Mitra, M., Buckley, Ch., Singhal, A., Caride C. (1997). An analysis of statistical and syntactic phrases. W: *Proceedings of RIAO*, t. 97, 200-214

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Claredon Press. Oxford

Moreno-Ortiz, A., Pérez-Hernández, C., Del-Olmo, M. Á. (2013). Managing Multiword Expressions in a Lexicon-Based Sentiment Analysis System for Spanish. W: *The 9th Workshop on Multiword Expressions (MWE)*. Atlanta, GA, Association for Computational Linguistics, 42-50

Mosiołek-Kłosińska, K. (2002). Innowacje frazeologiczne jako źródło powstawania nowych jednostek leksykalnych. W: Lewicki, A.M. *Problemy frazeologii europejskiej*, t. 5. Lublin, 21-33

Moszczyński, R. (2007). A Practical Classification of Multiword Expressions. W: *Proceedings of the ACL 2007 Student Research Workshop*. Praga, Association for Computational Linguistics, 19-24

Moszczyński, R. (2010). Towards a bilingual lexicon of information technology multiword units. W: *Proceedings of the XIV Euralex International Congress*, 1-4

Nivre, J., Nilsson, J. (2004). Multiword Units in Syntactic parsing. W: *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edynburg, Edinburgh University Press

Ogrodniczuk, M. (2003). Nowa edycja wzbogaconego korpusu słownika frekwencyjnego. W: *Językoznawstwo w Polsce. Stan i perspektywy.* Polska Akademia Nauk – Komitet Językoznawstwa, Uniwersytet Opolski, Opole, 181-190

Orliac, B., Dillinger, M. (2003). Collocation extraction for machine translation. W: *Proceedings of Machine Translation Summit IX*. New Orleans, LA, 292-298

Pajdzińska, A. (1988). *Związki frazeologiczne nazywające akt mowy. Semantyka i składnia.* Lublin

Pajdzińska, A. (1991). Wartościowanie we frazeologii. W: Puzynina, J., Anusiewicz, J. (red.). *Język a kultura*, t. 3, Wrocław, 15-28

Pantel, P., Lin, D. (2001). A Statistical Corpus-Based Term Extractor. W: Stroulia, E., Matwin, S. (red.). *AI 2001, Lecture Notes in Artificial Intelligence*. Springer-Verlag, 36-46

Paulo, J., Correia, M., Mamede, N., Hagege, C. (2002). Using morphological, syntactical, and statistical information for automatic term acquisition. W: *Advances in Natural Language Processing.* Springer, 219-227

Pearce, D. (2001). Synonymy in Collocation Extraction. W: *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations.* Pittsburgh, PA, 41-46

Pecina, P. (2008a). A Machine Learning Approach to Multiword Expression Extraction. W: *Proceedings of the 6th International Conference on Language Resources and Evaluation Workshop: Towards a Shared Task for Multiword Expressions*. Marrakesz, 54-57

Pecina, P. (2008b). Reference data for Czech collocation extraction. W: *Proceedings of the 6th International Conference on Language Resources and Evaluation Workshop: Towards a Shared Task for Multiword Expressions*. Marrakesz, 11-14

Pecina, P. (2009). *Lexical association measures: Collocation extraction*. Praga

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation, 44(1-2)*, 137-158

Pecina, P., Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. W: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, Sidney, 651-658

Pęzik, P. (2009). Extraction of multiword expressions for corpus-based discourse analysis. W: *Studies in cognitive corpus linguistics. Lodz Studies in Language – Volume 18*

Pęzik, P. (2012).Wyszukiwarka PELCRA dla danych NKJP. W: Przepiórkowski, A., Bańko, M., Górski, R., Lewandowska-Tomaszczyk, B. (red.). *Narodowy Korpus Języka Polskiego*. Warszawa, Wydawnictwo Naukowe PWN, 253-274

Pęzik, P. (2013). Paradygmat Dystrybucyjny w Badaniach Frazeologicznych. Powtarzalność, Reprodukcja i Idiomatyzacja. W: Stalmaszczyk, P. (red.). *Metodologie Językoznawstwa. Ewolucja Języka, Ewolucja Teorii Językoznawczych*. Wydawnictwo Uniwersytetu Łódzkiego

Pęzik, P. (2014). Graph-Based Analysis of Collocational Profiles. W: Jesenšek, V., Grzybek, P. (red.). *Phraseologie Im Wörterbuch Und Korpus (Phraseology in Dictionaries and Corpora)*. ZORA 97. Maribor, Bielsko-Biała, Budapest, Kansas, Praha: Filozofska fakulteta, 227-43

Piasecki, M., Szpakowicz, S., Broda, B. (2012). *A Wordnet from the Ground Up*. Wroclaw

Pisarek, P. (2009). Słownik fleksyjny. W: Lubaszewski, W. (red.). *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* Kraków, Uczelniane Wydawnictwa Naukowo- Dydaktyczne AGH, 37-68

Piskorski, J. (2004). Extraction of Polish Named-Entities. W: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lizbona, 313-316

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14(3)*, 130-137.

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

Przybylska, R. (2002). *Polisemia przyimków polskich w świetle semantyki kognitywnej*. Kraków, Universitas

Ramisch, C. (2012). A *generic and open framework for multiword expressions treatment: from acquisition to applications*. Praca doktorska, Uniwersytet w Grenoble, Grenoble, Francja

Ramisch, C., Schreiner, P., Idiart, M., Villavicencio, A. (2008). An evaluation of methods for the extraction of multiword expressions. W: Grégoire, N., Evert, S., Krenn, B. (red.). *Proceedings of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008).* Marrakesz, 50-53

Ramisch, C., Villavicencio, A., Boitet, C. (2010). mwetoolkit: a framework for multiword expression identification. W: *Proceedings of the Seventh LREC (LREC 2010).* Malta, ELRA, 662-669

Reddy, S., McCarthy, D., Manandhar, S. (2011). An empirical study on compositionality in compound nouns. W: *Proceedings of 5th International Joint Conference on Natural Language Processing.* Chiang Mai, 210-218

Rejakowa, B. (1982). Ekwiwalencja tłumaczenia związku frazeologicznego o identycznej strukturze formalnej i znaczeniowej w przekładach z języka słowackiego na język polski. W: Basaj, M., Rytel, D. (red.). Z *problemów frazeologii polskiej i słowiańskiej*, t. I. Wrocław, 173-182

Resnik, Ph. (1997). Selectional preferences and sense disambiguation. W: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*. Waszyngton, 52-57

Ritz, J. (2006). Collocation extraction: Needs, feeds and results of an extraction system for German. W: *Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trydent, 41-48

Rokitiański, M. (2009). Słownik wyrazów wielosegmentowych. W: Lubaszewski, W. (red.). *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* Kraków, Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 69-78

Rudolf, M. (2004). *Metody automatycznej analizy korpusu tekstów polskich. Pozyskiwanie, wzbogacanie, przetwarzanie informacji lingwistycznych.* Warszawa

Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. W: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*. Mexico City, 1-15

Saussure de, F. (1961). *Kurs językoznawstwa ogólnego*. PWN, Warszawa

Savary, A. (2009). Multiflex: a Multilingual Finite-State Tool for Multi-Word Units. W: *Proceedings of CIAA 2009.* Springer Verlag, 237-240

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., Makowiecki, F. (2012). SEJFEK – a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. W: *Proceedings of Cognitive Aspects of the Lexicon (COGALEX-III), a Workshop at COLING 2012*. Mumbai, 195-214

Schone, P., Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? W: Lee, L., Harman, D. (red.). *Proceedings of the 2001 EMNLP (EMNLP 2001).* Pittsburgh, PA, ACL, 100-108

Seretan, V. (2011a). A Collocation-Driven Approach to Text Summarization. W: *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*. Montpellier, 9-14

Seretan, V. (2011b). *Syntax-Based Collocation Extraction.* Springer (Text, Speech and Language Technology, t. 44)

Seretan, V., Nerima, L., Wehrli, E. (2003). Extracion of Multi-Word Collocations Using Syntactic Bigram Composition. W: *International Conference on Recent Advances in NLP.* Borovets, 424-431

Seretan, V., Nerima, L., Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. W: *Proceedings of the 11th EURALEX International Congress, EURALEX 2004.* Lorient, 755-766

Seretan, V., Wehrli, E. (2009). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, *43(1)*, 71-85

Silva, J., Dias, G., Guilloré, S., Lopes, J. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. W: *Progress in Artificial Intelligence.* Springer, 113-132

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford, Oxford University Press

Sinclair, J. (red.). (1995). *Collins COBUILD English Dictionary.* Harper Collins. Londyn

Skorupka, S. (1982). Klasyfikacja jednostek frazeologicznych i jej zastosowanie w leksykografii. W: Basaj, M., Rytel, D. (red.). Z *problemów frazeologii polskiej i słowiańskiej*, t. I, Wrocław, 7-15

Skubalanka, T. (1972). O ekspresywności języka. W: *Annales Universitatis Mariae Curie-Skłodowska. Sectio F, Nauki Filozoficzne i Humanistyczne. Vol. 27*, 123-135

Smadja, F. (1991). *Retrieving collocational knowledge from textual corpora. An application: Language generation.* Praca doktorska. Uniwersytet Kolumbia

Smadja, F. (1992). How to compile a bilingual collocational lexicon automatically. W: *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques.* San Jose, CA

Smadja, F. (1993). Retrieving collocations from text: XTRACT. *Computational Linguistics, 19*, 143-177

Smadja, F., McKeown, K. (1990). Automatically extracting and representing collocations for language generation. W: *Proceedings of the 28th annual meeting on Association for Computational Linguistics.* Stroudsburg, 252-259

Spärck Jones, K. (2001). Natural language processing: a historical review. W: Zampolli, A., Calzolari, N., Palmer, M. (red.). *Current Issues in Computational Linguistics: in Honour of Don Walker.* Amsterdam, Kluwer, 3-16

Straś, E. (2008). *Kategoria intensywności we frazeologii języka polskiego i rosyjskiego*. Katowice, Wydawnictwo Uniwersytetu Śląskiego

Terra, E., Clarke, Ch. (2003). Frequency estimates for statistical word similarity measures. W: *Proceedings of HLT-NAACL 2003.* Edmonton, Alberta, 244-251

Todirascu, M., Gledhill, Ch. (2008). Extracting Collocation in Context: The case of Verb--Noun Constructions in English and Romanian. W: *Recherches Anglaises et Nord--Américaines (RANAM).* Strasburg

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. W: De Raedt, L., Flach, P. (red.), *Proceedings of the 12th European Conference on Machine Learning (ECML- 2001).* Freiburg, 491-502

Van de Cruys, T., Villada Moirón, B. (2007). Semantics-based multiword expression extraction. W: *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions.* Association for Computational Linguistics, 25-32

Vetulani, G., Vetulani, Z., Obrębski, T. (2008). Verb-noun collocation SyntLex dictionary: Corpus-based approach. W: *Proceedings of the 6th International Language Resources and Evaluation (LREC'08).* Marrakesz, 1561-1564

Villada Moirón, B. (2005). *Data-driven identification of fixed expressions and their modifiability.* Praca doktorska. Uniwersytet w Groningen

Villada Moirón, B., Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. W: *Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trydent, 33-40

Washtell, J., Markert, K. (2009). A Comparison of Windowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations. W: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, t. 2. Association for Computational Linguistics Stroudsburg, 628-637

Wolinski, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, *XXII–XXIII*, 39-55.

Wolinski, M. (2006). Morfeusz – a practical tool for the morphological analysis of Polish. W: Kłopotek, M. A., Wierzchon, S. T., Trojanowski, K. (red.). *Intelligent Information Processing and Web Mining, Advances in Soft Computing*, s. 503-512. Springer-Verlag, Berlin.

Woźniak, M. (2011). Automatic extraction of multiword lexical units from Polish text. W: *Proceedings of the 5th Language & Technology Conference.* Poznań, 187-193

Wróbel, H. (1995). Problemy dyskusyjne syntaktycznej klasyfikacji polskich leksemów. *Studia gramatyczne*, t. 11, 7-18

Wróbel, H. (1996). Nowa propozycja klasyfikacji syntaktycznej polskich leksemów. W: Wróbel, H. (red.). *Studia z leksykologii i gramatyki języków słowiańskich.* Kraków, 53-60

Yagunova, E.V., Pivovarova, L.M. (2010). The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts. *Automatic Documentation and Mathematical Linguistics, 44(3)*, 164-175

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. W: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 189-196

Zhang, Y., Kordoni, V., Villavicencio, A., Idiart, M. (2006). Automated Multiword Expression Prediction for Grammar Engineering. W: Moirón, B.V., Villavicencio, A., McCarthy, D., Evert, S., Stevenson, S. (red.). *Proceedings of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006).* Sidney, 36-44

Żmigrodzki, P. (2009). *Wprowadzenie do leksykografii polskiej*. Katowice, Wydawnictwo Uniwersytetu Śląskiego

Dodatek A. 300 najwyżej ocenionych jednostek leksykalnych w zbiorze porównawczym

Legenda

Nrg – nieregularność
Prg – pragmatyczność
Knw – konwencjonalizacja
SLks – stałość leksykalna
SSkł – stałość składniowa
Npk – nieprzekładalność
ZL – zamkniętość lewostronna
ZP – zamkniętość prawostronna
O/E – obrazowość / ekspresywność

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
martwa natura	2	0	2	2	2	2	2	0	1	32
dobre imię	2	0	2	2	2	2	0	2	1	32
wolna ręka	2	0	2	2	2	2	2	2	1	32
panna młoda	2	0	2	2	2	2	2	2	1	32
gołym okiem	2	0	2	1	2	2	2	0	1	31
błędne koło	2	0	2	2	2	2	2	0	0	30
młoda para	2	0	2	2	2	2	2	2	0	30
siła robocza	2	0	2	2	2	2	2	2	0	30
wojna domowa	2	0	2	2	1	2	0	2	1	30
pan młody	2	0	2	2	2	2	2	2	0	30
sztuki wyzwolone	2	0	2	2	2	2	2	2	0	30
kawałek chleba	2	0	2	2	2	1	2	2	1	30
święty spokój	2	0	1	2	2	2	2	0	1	29
zdrowy rozsądek	1	0	2	2	2	2	2	0	1	28
czarna dziura	2	0	2	2	2	0	0	2	1	28
świętej pamięci	2	2	0	2	2	2	2	0	0	28
żelazna kurtyna	2	0	2	2	2	0	2	2	1	28
królestwo niebieskie	2	0	2	2	2	1	0	2	0	28
kość słoniowa	1	0	2	2	2	1	2	2	2	28
chleb żywy	2	0	2	2	1	1	2	2	1	28
naród wybrany	2	0	2	2	2	1	0	2	0	28
pewnego razu	2	0	1	2	2	2	2	0	0	27
siwy włos	2	0	2	1	1	1	2	0	1	27
przyszywana ciotka	2	0	2	2	1	2	1	0	1	27
gwiazda morska	2	0	1	2	2	1	2	2	1	27
rynek pracy	1	0	2	2	2	1	2	0	1	26
wymiar sprawiedliwości	1	0	2	2	2	1	2	0	1	26
święta wojna	1	0	2	2	2	1	2	0	1	26

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
dobra wola	1	0	2	2	2	2	0	2	0	26
zimna wojna	1	0	2	2	2	1	2	0	1	26
dobra nowina	2	0	2	2	2	0	0	2	0	26
dobre wychowanie	1	0	2	2	2	2	0	2	0	26
zielone pojęcie	2	0	0	2	2	2	2	0	1	26
fałszywy prorok	2	0	2	2	1	1	2	0	0	26
drogi oddechowe	1	0	2	2	2	2	2	0	0	26
szata roślinna	2	0	2	2	1	1	2	0	0	26
ojciec duchowny	2	0	2	2	1	1	2	2	0	26
rachunek sumienia	2	0	2	2	1	1	2	0	0	26
siła rzeczy	2	0	0	2	2	2	2	2	1	26
duża litera	1	0	2	1	2	2	2	0	0	25
cywilizacja łacińska	1	0	2	1	2	2	0	2	0	25
słowo wcielone	2	0	1	2	2	1	2	0	0	25
obóz zagłady	2	0	2	1	1	1	2	0	0	25
światłość świata	2	0	1	2	2	1	2	0	0	25
działalność gospodarcza	1	0	2	2	2	1	2	0	0	24
papiery wartościowe	1	0	2	2	2	1	2	1	0	24
dzieło sztuki	1	0	2	2	2	1	2	0	0	24
prawo jazdy	1	0	2	2	2	1	0	2	0	24
obcy język	1	0	2	2	1	2	2	0	0	24
tajna policja	1	0	2	2	2	1	2	0	0	24
dobry uczynek	1	0	2	2	2	1	0	2	0	24
zielona herbata	1	0	2	2	1	1	2	0	1	24
rajski ptak	1	0	2	2	1	2	2	0	0	24
wyższa uczelnia	1	0	2	2	2	1	2	0	0	24
intensywna terapia	1	0	2	2	2	1	2	2	0	24
jama ustna	0	0	2	2	2	2	2	0	1	24
wieki średnie	1	0	2	2	2	1	0	2	0	24
obóz koncentracyjny	1	0	2	2	1	2	2	2	0	24
liczba mnoga	1	0	2	2	1	2	2	2	0	24
literatura piękna	1	0	2	2	2	1	0	2	0	24
łódź podwodna	1	0	2	2	1	2	2	0	0	24
wiek produkcyjny	1	0	2	2	2	1	0	2	0	24
liczba pojedyncza	1	0	2	2	2	1	2	0	0	24
pytanie retoryczne	2	0	2	2	0	1	0	2	0	24
ogród zoologiczny	1	0	2	2	2	1	2	0	0	24
wóz pancerny	1	0	2	2	2	1	2	0	0	24
sąd ostateczny	1	0	2	2	2	1	0	2	0	24
wyrzuty sumienia	1	0	2	2	2	1	2	0	0	24
zawrót głowy	2	0	0	2	2	1	2	0	1	24

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
obóz pracy	1	0	2	2	1	2	2	0	0	24
wrogowie ludu	1	0	2	2	1	1	0	2	1	24
rzut serca	1	0	2	2	1	2	2	0	0	24
rzut oka	2	0	2	2	2	1	0	0	2	23
dobre strony	1	0	2	1	1	2	0	2	0	23
najwyższy czas	2	0	0	1	2	2	2	2	0	23
kamera papieska	1	0	2	1	2	1	2	0	0	23
lata międzywojenne	1	0	2	1	2	1	0	2	0	23
ułamek sekundy	1	0	1	2	2	1	2	0	1	23
droga życia	2	0	1	2	0	1	2	0	1	23
zbawienie świata	2	0	2	2	1	1	1	0	0	23
towarzysz broni	2	0	2	2	1	1	0	1	0	23
pan stworzenia	2	0	2	2	1	0	1	0	1	23
klatka schodowa	0	0	2	2	2	2	2	2	0	22
szkoła średnia	1	0	2	2	1	1	0	2	0	22
szkoły wyższych	1	0	2	2	1	1	2	2	0	22
parki narodowe	1	0	2	2	1	1	2	0	0	22
ropa naftowa	0	0	2	2	2	2	2	1	0	22
wolne rodniki	0	0	2	2	2	2	0	2	0	22
stare dzieje	1	0	2	1	2	2	1	0	0	22
święty wojownik	1	0	2	2	2	0	2	0	0	22
wysoki obcas	1	0	2	2	1	1	2	2	0	22
ludzka rodzina	1	0	2	2	1	0	0	2	1	22
testy punktowe	1	0	2	2	0	2	0	2	0	22
klatka piersiowa	0	0	2	2	2	2	2	0	0	22
grzech pierworodny	1	0	2	2	0	2	0	2	0	22
nazwa własna	0	0	2	2	2	2	2	2	0	22
skóra właściwa	0	0	2	2	2	2	0	2	0	22
pęcherzyki płucne	0	0	2	2	2	2	2	0	0	22
naczynia krwionośne	0	0	2	2	2	2	2	2	0	22
metale ciężkie	1	0	2	2	1	1	0	2	0	22
osoba boska	1	0	2	2	2	0	2	2	0	22
pole minowe	0	0	2	2	2	2	2	2	0	22
przyrost naturalny	1	0	2	2	1	1	0	2	0	22
próba kliniczna	1	0	2	2	2	0	2	0	0	22
jama brzuszna	1	0	2	2	1	1	2	0	0	22
język urzędowy	1	0	2	2	1	1	2	0	0	22
droga krzyżowa	1	0	2	2	1	1	0	2	0	22
sieć osadnicza	1	0	2	2	1	1	0	2	0	22
materiał źródłowy	1	0	2	2	1	1	2	2	0	22
zwierzęta domowe	1	0	2	2	1	1	0	2	0	22

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
nowotwory złośliwe	1	0	2	2	0	1	0	2	1	22
punkt wyjścia	1	0	2	2	1	1	0	2	0	22
wyznanie wiary	1	0	2	2	1	1	2	0	0	22
system wartości	1	0	2	2	1	1	2	0	0	22
środek ciężkości	1	0	2	2	1	1	0	2	0	22
rzecz jasna	1	1	0	1	1	2	0	2	1	21
ochrona przyrody	1	0	2	2	2	1	1	0	0	21
służba zdrowia	1	0	2	2	2	1	1	1	0	21
złe samopoczucie	0	0	2	1	2	2	2	0	0	21
naoczny świadek	1	0	1	2	1	2	2	0	0	21
wychowanie seksualne	1	0	2	2	2	1	1	0	0	21
życie duchowe	1	0	2	2	2	1	1	0	0	21
wiek szkolny	1	0	2	1	1	1	2	0	0	21
świat zewnętrzny	1	0	1	2	2	1	0	2	0	21
zmiany skórne	1	0	2	1	1	1	2	0	0	21
woda żywa	1	0	1	2	1	1	0	2	1	21
materiał własny	1	0	1	2	2	1	2	0	0	21
pomoce naukowe	1	0	1	2	2	1	2	0	0	21
wybuch wojny	2	0	0	1	1	1	2	0	1	21
poczucie bezpieczeństwa	1	0	2	2	2	1	1	0	0	21
poczucie wyższości	1	0	2	2	2	1	1	0	0	21
radość życia	1	0	1	2	2	1	2	0	0	21
słowo honoru	1	0	2	2	2	1	1	0	0	21
gra słów	2	0	2	2	2	1	0	0	1	21
piłka nożna	0	0	2	2	1	2	0	2	0	20
koniec świata	2	0	1	2	2	1	0	0	2	20
lewa strona	0	0	2	2	1	2	0	2	0	20
prawa strona	0	0	2	2	1	2	0	2	0	20
dawne czasy	1	0	2	2	1	0	0	2	0	20
święty obraz	1	0	2	1	2	1	1	0	0	20
swobodne zwierciadło	0	0	2	2	1	2	0	2	0	20
śmierć krzyżowa	1	0	2	2	1	0	2	0	0	20
ciepła woda	1	0	1	1	2	1	2	0	0	20
otwarte drzwi	2	0	0	2	1	0	2	2	1	20
nocna koszula	0	0	2	2	1	2	2	0	0	20
zły sen	1	0	1	1	2	0	2	0	1	20
język naturalny	1	0	2	2	1	0	0	2	0	20
kierownik duchowy	0	0	2	2	2	1	2	0	0	20
stan zapalny	0	0	2	2	1	2	0	2	0	20
praca magisterska	0	0	2	2	1	2	1	2	0	20
przekrój czynny	0	0	2	2	1	2	0	2	0	20

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
rodzaj ludzki	0	0	2	2	1	2	2	0	0	20
odma opłucnowa	0	0	2	2	1	2	2	2	0	20
rok akademicki	0	0	2	2	2	1	2	0	0	20
współżycie seksualne	0	0	2	2	2	1	2	0	0	20
błona podstawna	0	0	2	2	2	1	1	2	0	20
broń jądrowa	1	0	2	1	2	1	0	1	0	20
płeć męska	0	0	2	2	2	1	2	0	0	20
święcenia kapłańskie	0	0	2	2	2	1	2	0	0	20
szkoła ludowa	1	0	2	2	0	1	1	2	0	20
naczynia włosowate	0	0	2	2	2	1	0	2	0	20
węzeł chłonny	0	0	2	2	2	1	2	2	0	20
klasa średnia	1	0	2	2	0	1	1	2	0	20
wody płodowe	0	0	2	2	2	1	2	0	0	20
nauki ścisłe	0	0	2	2	2	1	1	2	0	20
łożysko naczyniowe	0	0	2	2	2	1	2	2	0	20
śródbłonek naczyniowy	0	0	2	2	2	1	2	0	0	20
ruch jednostajny	0	0	2	2	2	1	0	2	0	20
mięśnie gładkie	0	0	2	2	2	1	0	2	0	20
dystych elegijny	0	0	2	2	2	1	2	0	0	20
tkanka ziarninowa	0	0	2	2	2	1	0	2	0	20
aktywność zawodowa	0	0	2	2	2	1	2	0	0	20
plamka katodowa	0	0	2	2	2	1	2	2	0	20
wyraz twarzy	1	0	2	2	0	1	2	0	0	20
zapalenie oskrzeli	0	0	2	2	1	2	1	2	0	20
środek nocy	1	0	0	2	2	1	2	0	1	20
zbieg okoliczności	2	0	0	2	1	1	2	0	0	20
literatura przedmiotu	0	0	2	2	2	1	0	2	0	20
język ciała	1	0	2	1	1	1	1	0	1	20
zachód słońca	0	0	2	2	1	2	2	0	0	20
kątem oka	1	0	0	2	2	1	2	0	1	20
prawda wiary	1	0	2	2	1	0	2	0	0	20
podaż płynów	0	0	2	2	2	1	2	0	0	20
szkoła podstawowa	1	0	2	2	2	0	1	0	0	19
opinia publiczna	0	0	2	1	2	1	2	0	0	19
podstawa prawna	1	0	2	2	1	1	1	0	0	19
dom dziecka	1	0	2	2	1	1	1	0	0	19
wysoki poziom	2	0	1	1	1	0	0	1	1	19
stare stworzenie	1	0	2	2	2	0	1	0	0	19
przyszłe życie	1	0	2	1	1	0	2	0	0	19
święta księga	1	0	2	2	2	0	1	0	0	19
nauki przyrodnicze	0	0	1	2	2	2	1	2	0	19

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
język literacki	1	0	1	2	1	1	2	0	0	19
ciśnienie parcjalne	0	0	2	1	2	1	0	2	0	19
produkt kartezjański	1	0	2	1	1	0	2	2	0	19
szok termiczny	1	0	2	2	1	1	1	0	0	19
stacja benzynowa	0	0	2	1	1	2	2	0	0	19
aparat fotograficzny	0	0	2	2	2	2	1	0	0	19
świat materialny	1	0	1	2	1	1	2	0	0	19
pomoce dydaktyczne	0	0	2	1	2	1	2	0	0	19
dobra materialne	1	0	1	2	1	1	2	0	0	19
więzień polityczny	1	0	2	2	1	1	1	0	0	19
przerzuty odległe	0	0	1	2	1	1	2	0	2	19
efekt uboczny	1	0	2	1	0	1	0	2	0	19
osoba prywatna	1	0	1	2	1	1	2	0	0	19
narządy wewnętrzne	1	0	2	1	0	1	1	2	0	19
prawo wewnętrzne	1	0	1	2	1	1	0	2	0	19
bezpieczeństwo wewnętrzne	1	0	2	1	1	0	0	2	0	19
poczucie humoru	1	0	2	2	1	1	1	0	0	19
punkt odniesienia	1	0	1	2	1	1	0	2	0	19
dom modlitwy	1	0	2	2	1	1	1	0	0	19
dom starców	1	0	2	2	1	1	1	0	0	19
świadectwo wiary	1	0	2	1	0	1	2	0	0	19
istota rzeczy	2	0	0	1	1	1	2	0	0	19
znak czasu	2	0	2	2	1	1	0	0	1	19
język obcy	1	0	1	1	1	1	0	2	0	18
tworzywo sztuczne	0	0	2	2	1	1	2	0	0	18
telefon komórkowy	0	0	2	2	1	1	1	2	0	18
ochrona środowiska	0	0	2	2	1	1	1	2	0	18
dłuższa chwila	1	0	1	1	1	1	2	0	0	18
osobowy byt	0	0	2	2	1	1	2	0	0	18
wynikowy zbiór	0	0	2	2	1	1	2	0	0	18
młode pokolenie	1	0	1	1	2	0	2	0	0	18
ogólna liczba	1	0	1	1	1	1	2	0	0	18
dzikie zwierzę	0	0	2	2	2	0	2	0	0	18
otwarta przestrzeń	1	0	1	1	1	1	2	0	0	18
dobre słowo	2	0	0	2	1	0	0	2	0	18
pełnej krwi	0	0	2	2	1	1	2	0	0	18
wolna chwila	1	0	0	2	1	2	0	2	0	18
państwa bałtyckie	1	0	2	1	1	1	0	1	0	18
komórka tuczna	0	0	2	2	1	1	0	2	0	18
unia hipostatyczna	0	0	2	2	2	0	0	2	0	18

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
zdanie względne	0	0	2	2	1	1	1	2	0	18
język ojczysty	1	0	1	1	0	2	0	2	0	18
zabór rosyjski	0	0	2	2	1	1	2	0	0	18
jelito grube	0	0	2	2	1	1	2	0	0	18
przekrój poprzeczny	0	0	2	2	0	2	2	0	0	18
rośliny naczyniowe	0	0	2	2	1	1	0	2	0	18
podręczniki szkolne	0	0	2	2	0	2	2	0	0	18
komórka jajowa	0	0	2	2	1	1	0	2	0	18
zaimek dzierżawczy	0	0	2	2	1	1	1	2	0	18
tętnica płucna	0	0	2	2	1	1	2	0	0	18
bulla papieska	0	0	2	2	2	0	2	0	0	18
krajobraz kulturowy	1	0	1	1	1	1	2	0	0	18
rok liturgiczny	0	0	2	2	1	1	2	0	0	18
linia spektralna	0	0	2	2	1	1	0	2	0	18
blizny przerostowe	0	0	2	2	1	1	1	2	0	18
człowiek wierzący	1	0	1	1	1	1	0	2	0	18
pasma górskie	0	0	2	2	1	1	2	0	0	18
seminarium duchowne	0	0	2	2	1	1	2	0	0	18
temperatura elektronowa	0	0	2	2	1	1	0	2	0	18
endotoksyna bakteryjna	0	0	2	2	1	1	2	0	0	18
dziennik podawczy	0	0	2	2	1	1	0	2	0	18
sytuacja rodzinna	1	0	1	2	2	1	1	0	0	18
śluby zakonne	0	0	2	2	1	1	2	0	0	18
jelito cienkie	0	0	2	2	1	1	2	0	0	18
legat papieski	0	0	2	2	1	1	2	0	0	18
kariera zawodowa	0	0	2	2	1	1	2	0	0	18
środki materialne	1	0	1	1	1	1	1	2	0	18
flora bakteryjna	0	0	2	2	1	1	2	0	0	18
sklep spożywczy	0	0	2	2	1	1	0	2	0	18
płyta nagrobna	0	0	2	2	1	1	0	2	0	18
liczebnik złożony	0	0	2	2	1	1	2	2	0	18
bezrobocie strukturalne	0	0	2	2	1	1	2	2	0	18
wyrażenie argumentowe	0	0	2	2	1	1	0	2	0	18
praca doktorska	0	0	2	2	1	1	2	2	0	18
płytka krwi	0	0	2	2	1	1	2	0	0	18
pełnia życia	1	0	0	2	2	1	2	0	0	18
znak zapytania	0	0	2	2	1	1	0	2	0	18
sens życia	1	0	2	1	1	1	1	0	0	18
pora dnia	0	0	2	2	1	1	2	0	0	18
mieszkaniec miasta	0	0	2	2	1	1	2	0	0	18

Hasło	Nrg	Prg	Knw	SLks	SSkł	Npk	ZL	ZP	O/E	Suma
układ równań	0	0	2	2	1	1	2	0	0	18
pole bitwy	0	0	2	2	1	1	2	0	0	18
klasa referencji	0	0	2	2	1	1	0	2	0	18
kierunek studiów	0	0	2	2	1	1	2	0	0	18
interfejs użytkownika	0	0	2	2	1	1	2	0	0	18
ruch oporu	0	0	2	2	1	1	0	2	0	18
prawo karne	0	0	2	2	2	1	1	1	0	17
rok szkolny	0	0	2	2	2	1	1	0	0	17
siły zbrojne	0	0	2	2	2	1	1	0	0	17
stan rzeczy	1	0	1	1	2	1	0	1	0	17
pora roku	1	0	2	2	2	2	0	0	0	17
konserwator zabytków	0	0	2	2	2	1	1	0	0	17
rodzinne miasto	0	0	1	2	1	2	2	0	0	17
ciężka choroba	0	0	2	1	0	1	2	0	1	17
czerwone wino	1	0	1	1	1	1	0	1	1	17
ciemne oczy	0	0	1	2	2	0	2	0	1	17
prawdziwa przyjaźń	1	0	1	2	1	0	2	0	0	17
starsza pani	1	0	1	2	1	0	2	0	0	17
spektroskopia emisyjna	0	0	2	1	1	1	2	0	0	17
gospodarstwo rolne	0	0	2	2	1	2	1	0	0	17
artykuł spożywczy	0	0	2	1	0	2	2	0	0	17
związek małżeński	0	0	2	2	1	2	1	0	0	17
rozmnażanie płciowe	0	0	2	2	2	1	0	1	0	17
karabin maszynowy	1	0	2	2	0	1	1	0	0	17
płeć żeńska	0	0	1	2	2	1	2	0	0	17
władza poznawcza	0	0	2	2	2	1	1	1	0	17
gatunek ludzki	0	0	2	1	1	1	2	0	0	17
kult maryjny	0	0	2	1	1	1	0	2	0	17

Dodatek B. Rezultaty działania algorytmu

100 najwyżej ocenionych wyrażeń w niezawierającym anotacji korpusie notatek prasowych PAP

konferencja prasowa	przejście graniczne	turniej tenisowy
wybory prezydenckie	państwo członkowskie	rekord świata
wybory parlamentarne	kościół katolicki	minister gospodarki
sekretarz generalny	robotnik przymusowy	sztab wyborczy
kampania wyborcza	rynek pracy	ordynacja wyborcza
rzecznik prasowy	szef sztabu	system obrony
projekt ustawy	okres przejściowy	bank centralny
minister obrony	kraj kandydujący	klasyfikacja generalna
wojna światowa	finanse publiczne	wypłata odszkodowania
nowelizacja ustawy	grupa przestępcza	rok więzienia
mistrz świata	olej napędowy	przedstawiciel władzy
minister sprawiedliwości	wzrost gospodarczy	wymiar sprawiedliwości
strona internetowa	ochrona środowiska	list gończy
kurs akcji	indeks rynku	prezes zarządu
komitet wyborczy	polityka zagraniczna	prokurator generalny
partia polityczna	zbrodnia wojenna	minister zdrowia
stopa procentowa	telewizja publiczna	szef klubu
związek zawodowy	zachmurzenie umiarkowane	badanie przeprowadzone
opinia publiczna	strona polska	rok życia
sekretarz stanu	ropa naftowa	notowanie ciągłe
cena akcji	minister kultury	poprawa koniunktury
igrzyska olimpijskie	klub parlamentarny	praca przymusowa
rynek podstawowy	punkt widzenia	wzrost zachmurzenia
szef rządu	mistrz olimpijski	minister rolnictwa
szef dyplomacji	zysk netto	program wyborczy
minister skarbu	atak terrorystyczny	akcja protestacyjna
obroty bieżące	runda turnieju	akcja serii
biuro prasowe	rzecznik rządu	dyrektor generalny
pozbawienie wolności	zachmurzenie duże	wynik badania
projekt nowelizacji	akcja spółki	cena hurtowa
koniec roku	etap wyścigu	integracja europejska
służba zdrowia	telefon komórkowy	klub piłkarski
wyścig kolarski	tytuł mistrza	
służba specjalna	piłka nożna	

Tabela 1: 100 najwyżej ocenionych wyrażeń w niezawierającym anotacji korpusie notatek prasowych PAP

500 najlepiej ocenionych wyrażeń w klasyfikacji SVM

łacina starożytna
naczynia krwionośne
mięśnie gładkie
pokój gościnny
ewangelie synoptyczne
nawozy sztuczne
jama ustna
praca badawcza
zaimek dzierżawczy
przekrój czynny
redaktor naczelny
powiązanie anaforyczne
alternatywy wybrane
przyrost naturalny
szkoła średnia
praca botaniczna
kampania wyborcza
ogród zoologiczny
język obcy
cząstki ciężkie
martwica oparzeniowa
plazma argonowa
głębokość krytyczna
infekcja wirusowa
rzecz jasna
oddech własny
test śródskórny
strofa trójdzielna
grzech pierworodny
coś podobne
hymnografia łacińska
gospodarka rynkowa
mediatorzy lipidowi
diagnostyka alergologiczna
infekcja bakteryjna
prawo wewnętrzne
reprezentacja semantyczna
nazwy zmienne
ściganie karne
semantyka formalna
nauki ścisłe
układ immunologiczny
zdanie względne
fraza nominalna
grzyby pleśniowe
życie publiczne
pole minowe
łuk kaskadowy
przekrój poprzeczny
testy punktowe
osoba prywatna
czas wolny
opinia publiczna
choroba zakaźna
rok akademicki
energia kinetyczna
rozkład radialny
literatura piękna
ciśnienie parcjalne
osoba dorosła
praca zbiorowa
strategia szczera
głosowanie szczere
śródbłonek naczyniowy
prace budowlane
wartość logiczna
droga powrotna
sąd ostateczny
elektrony swobodne
ćwiczenia duchowne
rzecz oczywista
produkt kartezjański
interpretacja kumulatywna
sobór ekumeniczny
istota żywa
społeczność lokalna
krążenie płucne
bezpieczeństwo wewnętrzne
stan wojenny
metale ciężkie
leki przeciwalergiczne
poradnia alergologiczna
obóz przyfabryczny
siły zbrojne
woda żywa
dieta eliminacyjna
powstanie styczniowe
sprawy beneficjalne
alternatywy dostępne
związek religijny
choroba wrzodowa
struktura składniowa
błona komórkowa
polityka zagraniczna
głosowanie sondażowe
składnik odżywczy
notacja graficzna
unia hipostatyczna
seminarium duchowne
nawyki żywieniowe
rzecz ważna
poezja religijna
edukacja ekologiczna
doświadczenie religijne
produkt spożywczy
aparat fotograficzny
dziennik podawczy
wiązka laserowa
prawo moralne
związek wyznaniowy
kościół parafialny
środki finansowe
drogi oddechowe
osoba bezrobotna

historia powszechna
cecha charakterystyczna
test naskórkowy
czas przeszły
tętnica płucna
mechanizm obronny
rodzajnik nieokreślony
ochrona środowiska
łożysko naczyniowe
wyprysk kontaktowy
służba wojskowa
narządy wewnętrzne
system obronny
ocena moralna
nazwa własna
formułka wstępna
kościół rzymskokatolicki
okupacja hitlerowska
odmiana dystrybutywna
artykuły spożywcze
umowa międzynarodowa
płeć żeńska
nietolerancja pokarmowa
ubój rytualny
osoba duchowna
stan prawny
kościół katolicki
skłonności homoseksualne
prawa rządzące
szkoły wyższymi
język angielski
państwa bałtyccy
zasady moralne
urząd pracy
szpital psychiatryczny
toksyna oparzeniowa
linia kolejowa
wyrażenie argumentowe
wolność wewnętrzna
zespół nerczycowy
wstawka poetycka
partia polityczna
choroba nowotworowa
fraza przymiotnikowa
alergen pokarmowy
wentylacja zastępcza
bóg żywy
środowisko naturalne
materiały budowlane
łuk swobodny
tkanka skórna
fraza czasownikowa
przepuklina przeponowa
stan jonizacyjny
badania naukowe
chleb żywy
ciśnienie tętnicze
grupa wiekowa
dom rekolekcyjny
przyszywana ciotka
symbol predykatywny
stan psychiczny
dokument fundacyjny
tożsamość płciowa
natura ludzka
dojrzałość uczuciowa
rośliny naczyniowe
kompleks leśny
sprawy wewnętrzne
funkcja ekspresywna
związek małżeński
wóz pancerny
życie płodowe
szkoła katedralna
kierownictwo duchowe
kula ziemska
bulla papieska
skóra właściwa
siła robocza
państwo piastowskie
administracja państwowa
głosowanie strategiczne
ktoś bliski
woda mineralna
osoba świecka
organ administracji
pochodzenie naturalne
alergia pokarmowa
syn marnotrawny
układ nerwowy
płyta nagrobna
preferencje indywidualne
warunek konieczny
substancja chemiczna
środek farmakologiczny
grupa zawodowa
zmiany skórne
kult maryjny
środowisko przyrodnicze
dusza ludzka
krew tętnicza
konstrukcja dzierżawcza
rozum ludzki
kompleksy immunologiczne
linia spektralna
psychologia rozwojowa
odpowiedzialność moralna
wartość odżywcza
życie wieczne
naoczny świadek
wydział teologiczny
wentylacja mechaniczna
księgozbiór domowy
użycie generyczne
osoba boska
filozofia moralna
istota ludzka
słownik staropolski
organizm ludzki
płeć męska

relacje międzyludzkie
prawo kanoniczne
szok termiczny
sytuacja materialna
powstanie listopadowe
profil spektralny
funkcja ilustracyjna
środek spożywczy
układ odpornościowy
normy moralne
sekretarz generalny
nawa boczna
choroba alergiczna
park narodowy
materiały archiwalne
broń jądrowa
lek przeciwzapalny
rośliny uprawne
opatrunek biologiczny
warunki pracy
zabór rosyjski
funkcja ornamentacyjna
cywilizacja bizantyńska
zaimek osobowy
traktat pokojowy
język logiczny
dom mieszkalny
rodzina ludzka
życie codzienne

krajobraz kulturowy
procesy naprawcze
państwa europejscy
naczynia płucne
stan wyjściowy
czas prac
władza państwowa
kraje europejskie
szkoła mariacka
życie duchowe

chwila obecna
barwniki spożywcze
nawozy mineralne
choroba psychiczna
odczyn alergiczny
flora bakteryjna
drzewa owocowe
edukacja formalna
składnik pokarmowy
gatunek ludzki
członek czynny
wymiar sprawiedliwości
łódź podwodna
objawy uboczne
plazma termiczna
reguła leksykalna
postępowanie jurysdykcyjne
podstawa prawna
kwantyfikator egzystencjalny
szata roślinna
temperatura elektronowa
kwantyfikacja równoległa
sztuki wyzwolone
umysł ludzki
przeszkody ekstradycyjne
baza wojskowa
rodzaj ludzki
królestwo niebieskie
piśmiennictwo średniowieczne
kategoria składniowa
liczebnik złożony
objętość oddechowa
idea słowiańska
uraz termiczny
komórka docelowa
fraza jądrowa
botanika farmaceutyczna
proste zdanie
państwo niemieckie

legat papieski
objawy chorobowe
sklep spożywczy
grupa etniczna
czynnik wzrostowy
język naturalny
młodzież szkolna
pochodzenie roślinne
odmiana kolektywna
substancje obce
strona tytułowa
produkty żywnościowe
środek geometryczny
prawo rzymskie
podstawa programowa
liczba mnoga
integracja europejska
przewodnictwo cieplne
prawo międzynarodowe
państwo obce
przewód tętniczy
jednostka osadnicza
grupa społeczna
wody płodowe
kancelaria papieska
zakład przemysłowy
naród słowiański

stan wejściowy
władze miejskie
tendencje homoseksualne
sieć osadnicza
ludność żydowska
wdech mechaniczny
program telewizyjny
prawo miejskie
człowiek wierzący
więź emocjonalna
program autorski
partia komunistyczna
materiał źródłowy
pomoc materialna
państwa zachodni
katedra gnieźnieńska
reakcja alergiczna
prosty człowiek
płeć przeciwna
sytuacja życiowa
wymiana gazowa
dom rodzinny
katedra romańska
cywilizacja łacińska
obraz radiologiczny
pieśń pasyjna
umowa dwustronna
edukacja środowiskowa
związki chemiczne
nakłady finansowe
miłość ludzka
droga krzyżowa
urzędnicy kancelaryjni
komórka bakteryjna
ruch turystyczny
przywilej uniwersytecki
sfera erotyczna
sytuacja polityczna
życie wspólnotowe
dojrzałość emocjonalna
badanie diagnostyczne
układ oddechowy
schorzenia alergiczne
reguła konstrukcyjna
pieśń maryjna
ktoś obcy
klasa polityczna
polityka społeczna
stan początkowy
przepisy prawne
wojna domowa
zwierzęta doświadczalne
rośliny kwiatowe
życie rodzinne
pracownik naukowy
wychowanie seksualne
mniejszość narodowa
postępowanie sądowe
odmiana neutralna
pytanie retoryczne

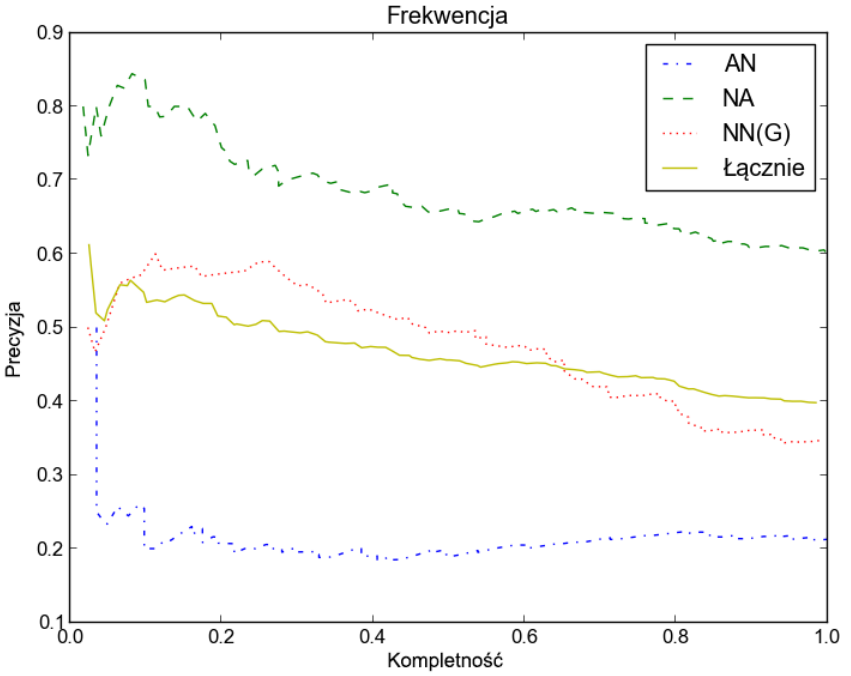
200 najniżej ocenionych wyrażeń w klasyfikacji SVM

godność człowieka
własne prawo
pan drogi
własna praca
początek roku
dziesiątki lat
profil linii
różnica temperatur
krótka przerwa
wzrost ciśnienia
szerszym zakres
wzajemna miłość
szkodliwe działanie
stałe ciśnienie
wysokość ciśnienia
przepływ krwi
teoria humoru
liczba alternatyw
możliwa interpretacja
różnorodne formy
lewa dłoń
odpowiedni moment
Starszy brat
koniec wieku
ogromne znaczenie
poszczególne elementy
podstawowe znaczenie
kolagen typu
najpoważniejszy problem
cel wychowania
specjalna komisja
skład drewna
brak środków
różne przedmioty
osobowy byt
wszelkie stworzenie
duża litera
relacja zależności
ilość alternatyw
znaczenie zdania
poszczególne przedmioty
duża skala
problematyka bezrobocia
połowa uczniów
dany teren
pełna świadomość
choroba skóry
prawidłowy rozwój
ogromna większość
własna wartość
istotna różnica
własne słowo
różne grupy
odpowiednie przygotowanie
zbiór alternatyw
zastosowanie diety
dzień tygodnia
wielkie znaczenie
własne ciało
zdrowie człowieka
współczynniki transportu
płuca szczura
najbliższa okolica
tajna policja
ogólna liczba
krótki okres
aktywna polityka
porządny człowiek
radykalna zmiana
kontekst użycia
nadmierna ilość
wartość ciśnienia
duża waga
przyszły uniwersytet
królewski dokument
przepływ gazu
wielka miłość
zbiór referentów
wróg ludu
osobowy podmiot
temperatura elektronów
jedyna kobieta
własna skóra
górna granica
własna rodzina
duża liczba
najmniejsza wątpliwość
zapalenie stawów
definicja operatora
różne źródła
jedyna droga
własny kraj
prawdziwy człowiek
pozytywna ocena
liczba ludzi
komórki gospodarza
najmłodsze pokolenie
własne doświadczenie
wiedza odbiorcy
różne metody
droga życia
minimalna ilość
różne instytucje
jedyne źródło
poszczególne grupy
jedyna szansa
ludzkie życie
mnóstwo ludzi
afgańska wojna
operator łączenia
zbiór znaczników

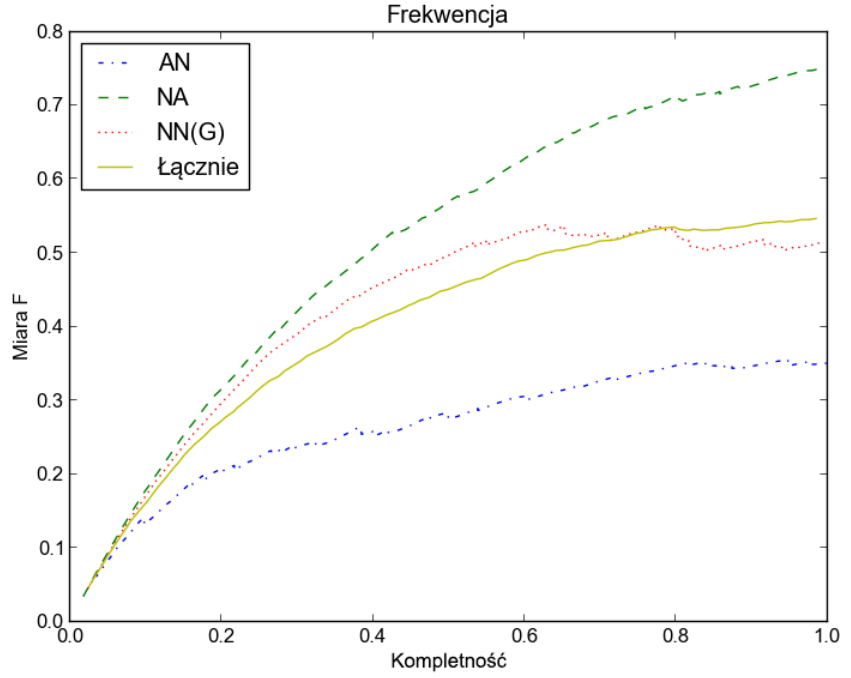
poczucie wolności
minister wojny
konkretny przypadek
wszelkie próby
szeroki zasięg
większa ilość
różne sytuacje
szeroki krąg
wartość współczynnika
kobiece ciało
afgańska ziemia
najprostszy sposób
duża odległość
katalog biblioteki
najtrudniejszy okres
współczesna kultura
najmocniejsza strona
wyrażenie typu
większej liczba
zasadnicze pytanie
obecny stan
własna droga
sposób życia
pojedynczy człowiek
miłość rodziców
najtrudniejsza sytuacja
miejsce popełnienia
denotacja typu
język ciała
rozwój osobowości
aktywny udział
ciepła woda
znaczenie wypowiedzi
rola nauczyciela
wzajemna pomoc
wielkie dzieło
różne dziedziny
pewien raz
kąt oka
najbliższa przyszłość
znacznik odmiany
możliwe znaczenie
ogólna zasada
chora ciekawość
analiza znaczenia
ciemne oczy
ograniczony zakres
konkretny człowiek
źródło emisji
niniejsza książka
spis członków
jedyny sposób
francuska polityka
życie dorosłe
różne sprawy
pora dnia
pan stworzenia
najważniejszy element
dokument króla
poszczególne jednostki
wysoka temperatura
najważniejszy czynnik
najtrudniejsze zadanie
dzień wojny
gwałtowny ruch
ścisły związek
właściwy organ
zbiór preferencji
ilość narkotyków
użycie operatora
najważniejszy element

czas popełnienia
otwarci drzwiach
najliczniejsza grupa
pan domu
zadanie typu
najwcześniejsze dzieciństwo
połowa roku
świat sztuki
własny świat
pewien dzień
poszczególny człowiek
pora roku
aktualny stan
pewna sprawa
pewien człowiek
działanie operatorów
codzienne życie
ziemskie życie
zdanie typu

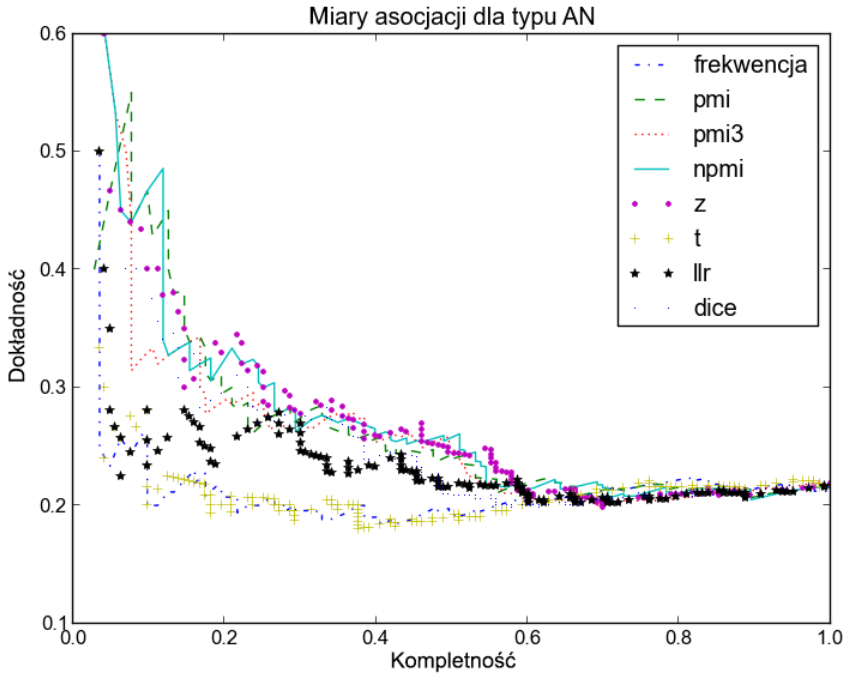
Uzupełniające wyniki oceny skuteczności algorytmu



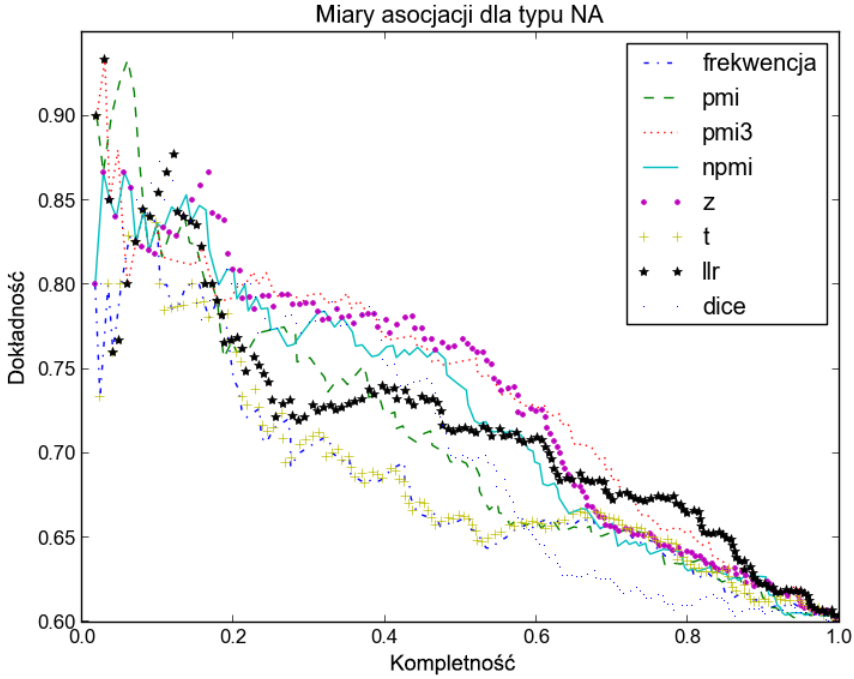
Wykres 3: Krzywa dokładności-kompletności dla klasyfikatora SVM



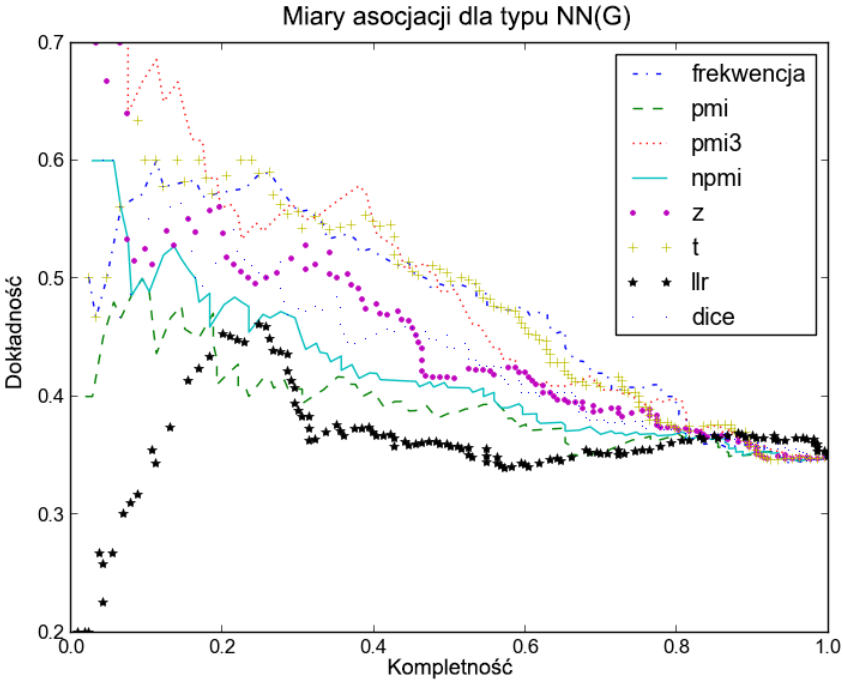
Wykres 4: Miara F – Frekwencja



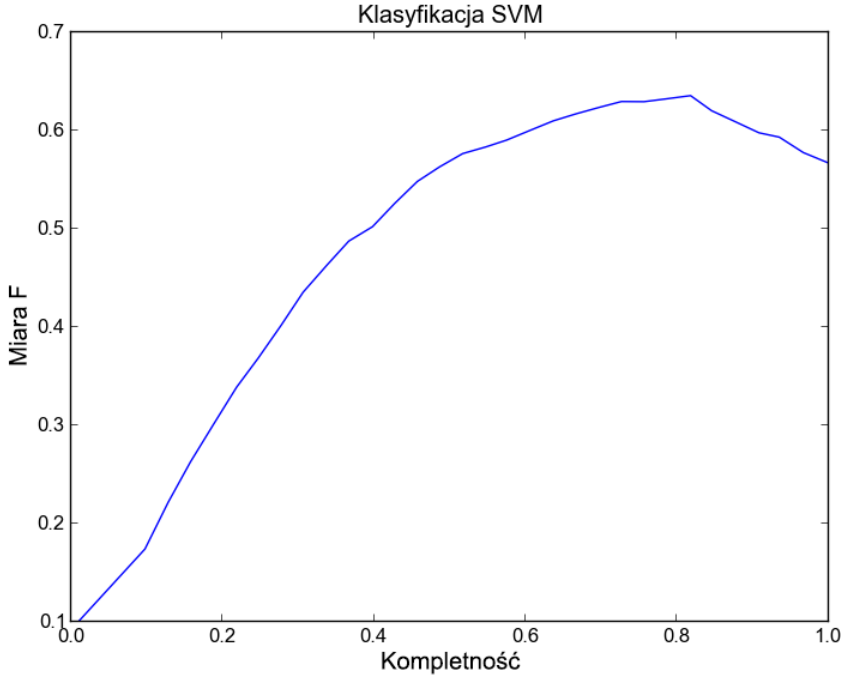
Wykres 5: Miary asocjacji – typ AN



Wykres 6: Miary asocjacji – typ NA



Wykres 7: Miary asocjacji – typ NN(G)



Wykres 8: Miara F – klasyfikacja SVM

Uzupełniające wyniki oceny skuteczności algorytmu

	AN	NA	NN(G)	Łącznie
PMI				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.28 / 0.22 / 0.21	0.75 / 0.66 / 0.60	0.39 / 0.37 / 0.35	0.49 / 0.43 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.31 / 0.33 / 0.35	0.44 / 0.63 / 0.75	0.34 / 0.46 / 0.52	0.37 / 0.48 / 0.55
AP	0.28	0.71	0.40	0.48
PMI^3				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.27 / 0.2 / 0.21	0.79 / 0.73 / 0.6	0.55 / 0.42 / 0.35	0.55 / 0.46 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.29 / 0.3 / 0.35	0.45 / 0.66 / 0.75	0.40 / 0.49 / 0.52	0.38 / 0.5 / 0.55
AP	0.28	0.74	0.49	0.51
NPMI				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.28 / 0.22 / 0.21	0.78 / 0.69 / 0.6	0.45 / 0.39 / 0.35	0.55 / 0.44 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.3 / 0.3 / 0.35	0.45 / 0.64 / 0.75	0.36 / 0.47 / 0.52	0.37 / 0.49 / 0.55
AP	0.29	0.73	0.43	0.49
z-score				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.29 / 0.21 / 0.21	0.78 / 0.72 / 0.6	0.51 / 0.42 / 0.35	0.54 / 0.46 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.3 / 0.31 / 0.35	0.45 / 0.66 / 0.75	0.38 / 0.49 / 0.52	0.38 / 0.5 / 0.55
AP	0.29	0.74	0.48	0.51
t-score				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.2 / 0.2 / 0.21	0.71 / 0.66 / 0.6	0.54 / 0.45 / 0.35	0.49 / 0.45 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.25 / 0.31	0.44 / 0.63 / 0.75	0.39 / 0.52 / 0.52	0.36 / 0.49 / 0.55
AP	0.23	0.69	0.49	0.48
LLR				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.26 / 0.21	0.73 / 0.71 / 0.6	0.38 / 0.34 / 0.35	0.47 / 0.43 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.28 / 0.31	0.44 / 0.65 / 0.75	0.34 / 0.44 / 0.52	0.36 / 0.48 / 0.55
AP	0.25	0.72	0.36	0.46
Dice				
P (R = 0.3) / P (R = 0.6) / (R = 1)	0.28 / 0.2 / 0.21	0.78 / 0.65 / 0.6	0.47 / 0.4 / 0.35	0.52 / 0.43 / 0.4
F (R = 0.3) / F (R = 0.6) / (R = 1)	0.29 / 03	0.44 / 0.63 / 0.75	0.37 / 0.49 / 0.52	0.37 / 0.48 / 0.55
AP	0.27	0.71	0.45	0.49

Tabela 2: Wyniki algorytmu dla miar asocjacji. Wartości precyzji (P) i miary F (F) podawane są dla dwóch progów kompletności (R): 0.8 i 1

Dodatek C. Obliczanie miar asocjacji

Obliczanie wartości miar asocjacji może sprawiać trudności osobom badającym język, niemającym jednak ugruntowanej wiedzy matematycznej. W tym miejscu omawiam krok po kroku procedurę liczenia takich miar, przy wykorzystaniu zrównoważonego podkorpusu NKJP. Za przykład niech posłuży wyrażenie *pies ogrodnika*.

1. Przygotowanie danych frekwencyjnych

Do przeprowadzenia obliczeń potrzebne są cztery wartości: frekwencja wyrazu pierwszego (*pies*), drugiego (*ogrodnik*), frekwencja bigramu (*pies*, *ogrodnik*), wreszcie ilość wszystkich bigramów w badanym korpusie. W celu obliczenia tej ostatniej wartości najlepiej posłużyć się liczbą wyrazów (tokenów) w całym korpusie pomniejszoną o 1 (np. jeśli korpus składa się z miliona wrazów, potrzebna wartość to 999 999). Frekwencję danego wyrażenia oznaczam jako *f (wyrażenie)*, ilość bigramów jako *N*. Należy tu zwrócić uwagę, że mówiąc o frekwencji wyrazu, mamy na myśli frekwencję leksemu, a nie jego konkretnej formy, zatem do frekwencji wyrazu *pies* wliczamy formy *pies*, *psa*, *psami* itd.

- f (*pies*) = 26 715
- f (*ogrodnik*) = 1780
- f (*pies*, *ogrodnik*) = 50
- N = 240 192 460

2. Przygotowanie tablic dwudzielczych

Miary asocjacji porównują frekwencję obserwowaną z frekwencją oczekiwaną. Dane z punktu pierwszego stanowią właśnie *wartości obserwowane*, które dobrze jest przedstawić za pomocą tablicy dwudzielczej. Tablica taka składa się podstawowo z czterech komórek, oznaczanych tradycyjnie jako O_{11}, O_{12}, O_{21}, O_{22}. Ich zawartość jest następująca:

- O_{11} – frekwencja poszukiwanego bigramu, tj. ilość wszystkich wystąpień w korpusie sekwencji leksemów *pies*, *ogrodnik* (np. *pies ogrodnika*, *psem ogrodnika* itd.). O_{11} w naszym przykładzie wynosi 50.
- O_{12} – frekwencja wszystkich bigramów, w których pierwszym leksemem **jest** *pies*, a drugim – dowolny leksem **poza** *ogrodnik*. Przykładowo mogą to być sekwencje *pies szczekał*, *pies się*, *pies sąsiada*. Wartość tę najłatwiej policzyć, odejmując od całościowej liczby wystąpień leksemu *pies* liczbę wystąpień sekwencji leksemów *pies*, *ogrodnik*. W przykładzie O_{12} jest równe 26 665 (26 715 – 50).
- O_{21} – frekwencja wszystkich sekwencji, w których pierwszym leksemem **nie jest** *pies*, a drugim **jest** *ogrodnik*. Np. *przyszedł ogrodnik*, *tamtego ogrodnika*, *Zimni ogrodnicy*. Wartość tę można uzyskać, odejmując od liczby wszystkich wystąpień leksemu *ogrodnik* liczbę wystąpień sekwencji *pies*, *ogrodnik*. W przykładzie mamy O_{21} = 1730 (1780 – 50).
- O_{22} to ilość wszystkich bigramów, których pierwszym leksemem **nie jest** *pies*, a drugim **nie jest** *ogrodnik*. O_{22} należy obliczyć za pomocą wzoru $O_{22} = N - f(pies) - f(ogrodnik) + f(pies, ogrodnik)$. Obliczenie dla przykładu wygląda następująco: O_{22} = 240 192 460 – 26 715 – 1780 + 50 = 240 164 015.

Sumy wartości z powyższych komórek (tzw. wartości brzegowe) dla kolumn i wierszy oznaczamy odpowiednio jako *C* i *R*.

W naszym przykładzie zawartość tablicy wygląda następująco (dane z poprzedniego punktu są wytłuszczone):

	wyraz 2 = ogrodnik	*wyraz 2 ≠ ogrodnik*	*R*
wyraz 1 = pies	O_{11}: **50**	O_{12}: 26 665	R_1: **26 715**
wyraz 1 ≠ pies	O_{21}: 1730	O_{22}: 240 164 015	R_2: 240 165 745
C	C_1: **1780**	C_2: 240 190 680	N: **240 192 460**

Warto zauważyć, że na potrzeby większości miar asocjacji wystarczy znać wartości O_{11}, R_1 i C_1 – a więc dane, które dostępne są już w punkcie 1, w takich przypadkach nie ma potrzeby dodatkowego obliczania wartości komórek O_{12}, O_{21} i O_{22}.

Na podstawie wartości brzegowych *R* i *C* możemy obliczyć *wartości oczekiwane* (oznaczane odpowiednio jako E_{11}, E_{12}, E_{21} i E_{22}), według schematu:

	wyraz 2 = ogrodnik	*wyraz 2 ≠ ogrodnik*
wyraz 1 = pies	$E_{11} = \frac{R_1 C_1}{N}$	$E_{22} = \frac{R_1 C_2}{N}$
wyraz 1 ≠ pies	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Zatem wartości oczekiwane dla wyrażenia *pies ogrodnika* kształtują się następująco:

	wyraz 2 = ogrodnik	*wyraz 2 ≠ ogrodnik*
wyraz 1 = pies	$E_{11} = 0.198$	$E_{22} = 26\,714.8$
wyraz 1 ≠ pies	$E_{21} = 1779.8$	$E_{22} = 240\,163\,965.2$

3. Obliczenie miary

Mając do dyspozycji dane z punktu 2, możemy obliczyć wartość każdej z opisanych w książce miar asocjacji. Dla przykładu sposób obliczenia PMI (*Pointwise Mutual Information*) i współczynika Dice'a:

PMI

Miara ta wyrażona jest wzorem: $PMI = \log_2 \frac{O_{11}}{E_{11}}$.

Odczytujemy zatem wartości $O_{11} = 50$ i $E_{11} = 0.198$, a następnie podstawiamy do wzoru:

$$PMI_{(pies,ogrodnik)} = \log_2 \frac{50}{0.198} = \log_2 252.53 = 7.98,$$

gdzie $\log_2$ oznacza logarytm o podstawie 2. Można go wyliczyć za pomocą kalkulatora, komputera (np. Excel) lub tablic matematycznych.

Współczynnik Dice'a

Korzystamy z wzoru: $Dice = \frac{2O_{11}}{R_1 + C_1}$.

Potrzebne wartości: $O_{11} = 50$, $R_1 = 26\,715$, $C_1 = 1780$.

Zatem: $Dice = \frac{2 * 50}{26715 + 1780} = 0.0035$.