Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik
Instytut Podstaw Informatyki PAN, Warszawa
Malgorzata.Marciniak@ipipan.waw.pl, Agnieszka.Mykowiecka@ipipan.waw.pl,
Piotr.Rychlik@ipipan.waw.pl

# Automatyczne wydobywanie terminologii dziedzinowej z korpusów tekstowych

słowa kluczowe: terminologia dziedzinowa, korpusy tekstowe

## 1 Wprowadzenie

Wydobywanie terminologii z korpusów dziedzinowych jest zadaniem polegającym na identyfikacji fraz, zwanych terminami, typowych dla dziedziny, której dotyczą zgromadzone teksty. Pojęcie terminologii dziedzinowej jest powszechnie stosowane, opracowywane są słowniki terminów np.: ekonomicznych (Gęsicki i Gęsicki 1996), literackich (Głowiński, Kostkiewiczowa i Okopień-Słowińska 2010), sztuk pięknych (Kozankiewicz 1976). Słowniki te składają się zwykle z haseł będących terminami dziedzinowymi wraz z opisem ich znaczenia. Problemy związane z tworzeniem, opracowaniem, normalizacją terminologii dotyczącej jakiejś dziedziny stanowią zakres zainteresowania naukowców wielu specjalności. Temat ten przewija się w pismach bibliotekoznawczych np.: „Przeglądzie bibliotecznym" (2003 1/2), „Zagadnieniach informacji naukowej" czy „Praktyce i Teorii Informacji Naukowej i Technicznej". Jest on interesujący dla językoznawców i tłumaczy, a także dla naukowców zajmujących się daną dziedziną.

Pomimo wielu publikacji dotyczących terminologii, brak jest precyzyjnej definicji, która mogłaby posłużyć do automatyzacji procesu wyboru haseł, które powinny znaleźć się w takich słownikach oraz do oceny uzyskanych wyników. Przyjrzyjmy się jak hasła *termin* oraz *terminologia* opisane są w słownikach języka polskiego.

Słownik pod redakcją Doroszewskiego (Doroszewski 1958-1969) określa termin jako „wyraz albo połączenie wyrazowe o specjalnym, konwencjonalnie ustalonym znaczeniu naukowym lub technicznym", natomiast terminologię jako „ogół terminów, którymi posługuje się dana dziedzina wiedzy, techniki; mianownictwo". Słownik PWN (Drabik, Kubiak-Sokół i Sobol 2012) te same hasła definiuje następująco: termin to „wyraz lub wyrażenie o specjalnym

znaczeniu w jakiejś dziedzinie", natomiast terminologię stanowi „ogół terminów i słownictwa używanego w danej dziedzinie wiedzy, zwłaszcza nauki lub techniki".

Powyższe definicje słownikowe nie wskazują jednak drogi postępowania w przypadku realizacji zadania przez program komputerowy. Dlatego też na potrzeby automatycznej ekstrakcji terminologii dziedzinowej najpierw zdefiniujemy co będziemy rozumieli pod pojęciem terminu.

Większość słowników zawierających terminologię dziedzinową składa się z fraz rzeczownikowych. Wprawdzie niektóre podejścia np. (Savova i inni 2003) biorą pod uwagę frazy czasownikowe, jednak zwykle terminami są formy nominalne, czyli *recytacja wiersza*, a frazy czasownikowe *recytowanie wiersza* czy *recytować wiersz* nie są uwzględniane. Na potrzeby niniejszego zadania ograniczymy więc pojęcie wyrażenia dziedzinowego do frazy rzeczownikowej. Jednak nie każda napotkana w tekście fraza rzeczownikowa jest terminem dziedzinowym. Przez termin dziedzinowy będziemy rozumieli frazę rzeczownikową, która w tekstach dziedzinowych występuje dostatecznie często by przypuszczać, że opisuje pojęcie istotne dla dziedziny i równocześnie częstość występowania tej frazy w tekstach spoza dziedziny jest niższa.

Zadanie ekstrakcji terminologii z tekstów dziedzinowych może być realizowane z myślą o wielu zastosowaniach, począwszy od tworzenia słowników i tezaurusów terminologii dziedzinowej, czy tworzenia słowników wielojęzycznych wykorzystywanych do automatycznego tłumaczenia, przez indeksowanie tekstów lub powiązaną z nim ekstrakcję informacji, aż po tworzenie ontologii dziedziny, której dotyczą zgromadzone teksty. W zależności od planowanego zastosowania konstrukcja identyfikowanych fraz może się nieco zmieniać, tak by rezultaty były najlepiej dopasowane do planowanego celu.

## 2 Automatyczna ekstrakcja terminologii dziedzinowej

W literaturze opisanych jest wiele rozwiązań służących realizacji zadania ekstrakcji terminologii z tekstów dziedzinowych, patrz (Korkontzelos, Klapafti i Manandhar 2008) i (Pazienza, Pennacchiotti i Zanzotto 2004). Wszystkie te metody wymagają zgromadzenia korpusu tekstów reprezentatywnych dla danej dziedziny. Zawarte w nim teksty poddawane są wstępnej analizie językowej przypisującej poszczególnym wyrazom nazwy części mowy (POS), a w językach fleksyjnych również ich pełną interpretację morfologiczną. Niezależnie

od przyjętych szczegółowych rozwiązań, proces ekstrakcji terminologii składa się z kilku, omówionych w dalszej części niniejszego rozdziału, etapów.

## 2.1 Identyfikacja fraz

Do identyfikacji fraz będących kandydatami na terminy dziedzinowe wykorzystywana jest wiedza lingwistyczna dotycząca ich konstrukcji gramatycznej w rozważanym języku. Takie podejście opisane jest w pracach: (Justeson i Katz 1995) i (Frantzi, Ananiadou i Mima 2000). W tym celu wykorzystywane jest zwykle płytkie parsowanie (ang. *shallow parsing*). Istnieją wprawdzie rozwiązania, w których stosowana jest pełna, głęboka analiza syntaktyczna (ang. *deep parsing*) całych zdań, patrz Savova i inni (2003), ale takie podejście wydaje się nie być efektywne. Należy też wspomnieć, że są także rozwiązania, w których pomijana jest całkowicie analiza syntaktyczna. Na przykład, w pracy (Wermter i Hahn 2005) kandydatami na terminy jest n kolejnych słów (n-gramy). Wybór gramatycznie poprawnych fraz dokonywany jest wówczas na późniejszym etapie.

W języku polskim terminami mogą być następujące frazy:

- rzeczowniki, akronimy lub skróty pochodzące od rzeczownika;
- rzeczowniki z sekwencją przymiotników (występujących przed lub po rzeczowniku);
- powyższe sekwencje rzeczownikowe modyfikowane tego samego typu frazą w dopełniaczu;
- frazy rzeczownikowe modyfikowane frazami przyimkowymi;
- koordynacja wymienionych powyżej fraz.

Jednak nie każda fraza zbudowana zgodnie z powyższymi regułami jest terminem dziedzinowym. Juan C. Sager w książce *A Practical Course in Terminology Processing* (1990) wskazuje kryteria, które powinny spełniać frazy terminologiczne. Jednym z nich jest możliwość ustalenia znaczenia frazy bez rozpatrywania kontekstu, w którym ona wystąpiła. Dlatego też frazy będące terminami dziedzinowym nie powinny zawierać pewnych słów np. *inny, niektóry, jakiś, pewien, poprzedni.* Fraza *inna faktura materiału* nie tworzy terminu, podczas gdy tak samo zbudowana fraza *bogata faktura materiału*, jest dobrą kandydatką na termin. Inną grupą słów, które nie powinny występować w wybranych frazach, są określenia czasu, które są uniwersalne dla każdej dziedziny. Są to między innymi słowa typu *dzień, miesiąc, godzina* oraz konkretne nazwy *poniedziałek* czy *grudzień*. Powszechnie stosowanym rozwiązaniem w przypadku automatycznej ekstrakcji terminologii jest opracowanie listy słów, które nie powinny występować w terminach.

W języku polskim, grupą wyrażeń prowadzących do tworzenia fraz niebędących właściwymi kandydatami na terminy są przyimki złożone. Jeśli rozważymy frazę *na gruncie muzyki współczesnej,* to bez wykluczenia przyimka złożonego *na gruncie* z możliwości tworzenia fraz terminologicznych, przy próbie identyfikowania wszystkich fraz rzeczownikowych otrzymamy frazę *grunt muzyki współczesnej*. Proponujemy więc wykluczenie przyimków złożonych (i ich fragmentów) z fraz kandydujących do miana terminu.

## 2.2 Szeregowanie fraz

Wyselekcjonowane wstępnie frazy rzeczownikowe różnią się między sobą potencjalną przydatnością przy tworzeniu terminologii dziedzinowej. Przykładowo, teksty dotyczące muzyki zawierać mogą frazy *orkiestra symfoniczna, drugi utwór,* czy *kardiolog.* Ta pierwsza jest niewątpliwie terminem związanym z wykonywaniem muzyki, druga raczej nie zostanie zaklasyfikowana jako termin, natomiast trzecia pochodzi z innej dziedziny i mogła pojawić się na przykład w kontekście opisywania pozamuzycznego wykształcenia muzyka czy kompozytora. Aby w tysiącach fraz rzeczownikowych zidentyfikować te najściślej związane z konkretną dziedziną, potrzebna jest metoda szeregowania ich tak, by frazy znajdujące się na początku wyznaczonej listy były tymi najbardziej charakterystycznymi. Porządek ten może wyznaczać miara liczbowa uwzględniająca cechy, które uznamy za ważne dla wyróżnienia terminów dziedzinowych. Najważniejszą informacją, jaką możemy wykorzystać przy konstrukcji takiej miary, jest liczba wystąpień frazy w tekście. Im częściej fraza występuje, tym większa jest szansa, że jest to fraza związana z daną dziedziną. Biorąc pod uwagę jedynie częstości występowania fraz nie uwzględnilibyśmy jednak co najmniej dwóch ważnych obserwacji. Po pierwsze, frazy wielowyrazowe występują rzadziej niż pojedyncze wyrazy, tak więc kryterium częstości powinno być różne dla fraz o różnej długości. Po drugie, część terminów dziedzinowych pojawia się bardzo często wewnątrz innych fraz, na przykład fraza *orkiestra symfoniczna* pojawia się w takich dłuższych określeniach jak: *nowojorska orkiestra symfoniczna, filadelfijska orkiestra symfoniczna* czy *narodowa orkiestra symfoniczna,* i wtedy częstość występowania terminu, który stanowi część dłuższej frazy (w tym przypadku *orkiestra symfoniczna*), byłaby zaniżona. Rozwiązaniem tego problemu jest uwzględnianie wystąpień fraz wewnętrznych. Rodzi ono jednak kolejne problemy, gdyż nie każda fraza wewnętrzna jest terminem dziedzinowym. Jeśli jakaś fraza występuje zawsze w otoczeniu tych samych słów, najbardziej prawdopodobne jest, że to cała, dłuższa sekwencja, a nie jej fragment tworzy termin dziedzinowy. Natomiast duża zmienność kontekstów często wskazuje na to, że rozważany fragment jest terminem. Przykładowo, wyszukując frazy wewnętrzne w

terminie *Uniwersytet Kardynała Stefana Wyszyńskiego*, zidentyfikujemy frazę *Uniwersytet Kardynała Stefana*. Jeśli występuje ona tylko w kontekście podanej frazy dłuższej, to powinna zostać odrzucona jako kandydat na termin. Z kolei duża liczba kontekstów dla frazy *kościół parafialny* (m.in. *stary, dawny, rzymskokatolicki, zakrystia, wieża*) jest argumentem przemawiającym za potraktowaniem tej frazy jako termin dziedzinowy. Propozycją miary istotności frazy jako kandydata na termin dziedzinowy, która uwzględnia poruszone wyżej problemy, jest współczynnik *C-value* (Frantzi, Ananiadou i Mima 2000):

$$C\text{-}value(f) = \begin{cases} l(f) * freq(f), jeżeli\ fraza\ f\ nigdzie\ nie\ występuje\ jako\ zagnieżdżona \\ l(f) * \left( freq(f) - \frac{\sum_{b \in T_f} freq(b)}{P(T_f)} \right), w\ przeciwnym\ przypadku \end{cases}$$

Przez $T_f$ oznaczmy zbiór fraz zawierających badaną frazę *f*. W zacytowanym wzorze liczba wystąpień danej frazy (*freq(f)*) mnożona jest przez wartość *l(f)* zwiększającą końcową wartość dla fraz dłuższych (typowym rozwiązaniem jest przyjęcie, że *l(f)* równe jest logarytmowi z długości frazy $f$ zwiększonemu o 1 lub jest to logarytm z długości frazy dla fraz wielowyrazowych i wybrana stała dla pojedynczych wyrazów). Składnik odejmowany obniża wagę tych fraz, które występują wewnątrz innych fraz, o wartość stanowiącą wynik podzielenia liczby wszystkich wystąpień badanej frazy (zarówno tych samodzielnych jak i wewnętrznych) przez liczbę różnych typów kontekstów ($P(T_f)$), w których te wystąpienia mają miejsce. Wartość ta jest tym większa, im mniej różnych typów kontekstów ma fraza dla tej samej ogólnej ich liczby. Jeśli fraza nigdy nie występuje samodzielnie, a tylko w jednym kontekście innej frazy rzeczownikowej, to końcowa wartość współczynnika równa jest zeru.

Przykładowe wyniki uzyskane dla zbioru około jednego miliona słów tekstów dotyczących muzyki (głównie jazzu), wykazują, że współczynnik *C-value* ma zbliżoną wartość (181.9) dla pojedynczego słowa (*kompozycja*), które wystąpiło ponad 1800 razy w prawie 500 różnych kontekstach, jak i dla frazy dwuwyrazowej (*muzyka jazzowa*), która wystąpiła około 180 razy w prawie 60 różnych kontekstach (179.4). Natomiast pojedyncze słowo *charakter*, które wystąpiło 185 razy w 96 różnych kontekstach ma współczynnik *C-value* równy 18,4.

Współczynnik *C-value* wykorzystywany jest do porządkowania fraz zidentyfikowanych w tym samym zbiorze tekstów. Porównywanie współczynników uzyskanych dla różnych zbiorów wymaga ich normalizacji.

## 2.3 Uproszczona forma podstawowa frazy

Dla języka polskiego (podobnie jak dla innych języków fleksyjnych) zliczanie częstości fraz, zarówno tych, co wystąpiły samodzielnie w tekście, jak i tych zagnieżdżonych w innych frazach, nie jest czynnością polegającą na prostym porównywaniu napisów. Następującą frazę w narzędniku: *renowacją nawy głównej* należy zidentyfikować jako wystąpienie frazy podstawowej: *renowacja nawy głównej*. W jej wnętrzu wystąpiły cztery zagnieżdżone frazy o następujących formach podstawowych: *renowacja nawy*, *nawa główna*, *renowacja* oraz *nawa*. Żadna z tych form nie może być rozpoznana, a tym samym nie może być zarejestrowane jej wystąpienie, w rozważanej frazie w narzędniku, i tylko dwie z nich dają się zidentyfikować przez proste porównywanie napisów w formie podstawowej. Jeżeli chcemy więc ustalić częstość wystąpienia frazy o podanej formie podstawowej, to musimy znaleźć w tekście wystąpienia form tej frazy we wszystkich przypadkach. Problem zliczania częstości fraz może być rozwiązany na kilka sposobów. Pierwszą nasuwającą się metodą jest sprowadzenie wszystkich wystąpień fraz rzeczownikowych w tekście do form podstawowych. Zadanie to wymaga jednak jednoznacznej i poprawnej lematyzacji fraz, co dla języka polskiego nie zawsze jest zadaniem prostym do realizacji przez program komputerowy. Rozważmy przykładową frazę w dopełniaczu *ołtarza głównego kościoła parafialnego,* dla której decyzja o tym czy przymiotnik *główny* opisuje *ołtarz* czy *kościół* wymaga wiedzy o świecie. Powyższej frazie możemy przypisać następujące lematy: *ołtarz główny kościoła parafialnego* jak i *ołtarz głównego kościoła parafialnego.* Decyzja jakiego stopnia przymiotnika należy użyć lematyzując np. frazę *wyższej szkoły* nastręcza trudności. Nie wiemy czy powinna to być *wyższa szkoła* czy też *wysoka szkoła.*

Innym rozwiązaniem jest wykorzystanie uproszczonej formy podstawowej, którą uzyskujemy lematyzując poszczególne elementy frazy. Dla wyżej rozważanej frazy jest to forma *renowacja nawa główny*, a dla zagnieżdżonych w niej fraz: *renowacja nawa*, *nawa główny*, *renowacja* oraz *nawa*. Ich wystąpienia w oznaczonym morfologicznie tekście, z lematami przypisanymi do poszczególnych słów, dają się zidentyfikować za pomocą prostego porównywania napisów. Rozwiązanie to również nie jest doskonałe, gdyż należy pamiętać, że istnieją frazy, które mają różne właściwe formy podstawowe, a ich uproszczone formy podstawowe są takie same. Przewagą tego rozwiązania jest jednak prostota zastosowania go do omawianego zadania. Poniżej omówimy kiedy takie rozwiązanie prowadzi do błędnego połączenia różnych znaczeniowo fraz.

- Frazy, w których element główny jest modyfikowany frazami rzeczownikowymi różniącymi się liczbą, np. *remont nawy bocznej* oraz *remont naw bocznych*. Jeśli zadaniem jest opracowanie słownika terminologicznego, to połączenie tych fraz uproszczoną frazą podstawową jest korzystne. Należy podkreślić, że w przypadku opracowywania słownika terminologicznego niezbędne jest ustalenie zasad, według których dokonywany jest wybór odpowiedniej formy podstawowej terminu umieszczanego w słowniku. Jednak w przypadku, gdy terminologia ma posłużyć do ekstrakcji informacji z tekstów może być istotne rozróżnienie obu fraz i ustalenie czy remont dotyczył obu naw bocznych czy tylko jednej.
- Frazy zawierają przymiotniki, które wystąpiły w różnych stopniach, np. *lewa nawa wysoka* oraz *lewa nawa wyższa*. Obie frazy mają taką samą formę podstawową *lewy nawa wysoki*. Pierwsza fraza jest stwierdzeniem faktu, że nawa jest wysoka, natomiast druga fraza mówi o względnej wielkości lewej nawy, zapewne w porównaniu z nawą prawą.
- Frazy zawierające imiesłowy przymiotnikowe w formie zanegowanej i niezanegowanej mają te same uproszczone formy podstawowe np. *niewydany tomik poezji* oraz *wydany tomik poezji* mają identyczne formy podstawowe *wydać tomik poezja* gdyż dla imiesłowów przymiotnikowych informacja o zanegowaniu jest jedną z kategorii gramatycznych.
- Frazy zawierające słowa należące do różnych klas gramatycznych (części mowy), a posiadających takie same uproszczone formy podstawowe. Sytuacja taka zachodzi dla fraz zawierających rzeczowniki odsłowne (gerundia) oraz imiesłowy przymiotnikowe, które mają takie same formy podstawowe będące bezokolicznikami. Istnieją więc frazy, które pomimo różnych konstrukcji syntaktycznych są sprowadzone do takiej samej uproszczonej formy podstawowej. Przykładem takich fraz są *uzgodnienie*$_{ger}$ *terminu renowacji* oraz *uzgodniony*$_{ppas}$ *termin renowacji*, obie mają uproszczoną formę podstawową *uzgodnić termin renowacja*. W pierwszej frazie elementem głównym jest rzeczownik odsłowny *uzgodnienie*, który jest modyfikowany frazą rzeczownikową w dopełniaczu *terminu renowacji*. Elementem głównym drugiej frazy jest natomiast rzeczownik *termin*, który modyfikowany jest imiesłowem przymiotnikowym *uzgodniony*.

Opisane powyżej sytuacje występują w danych sporadycznie. Niektórych z nich można uniknąć dodając do formy podstawowej informację o klasie gramatycznej słowa lub kategorii gramatycznej prowadzącej do konfliktu.

## 2.4 Wyodrębnianie zagnieżdżonych fraz

Jedną z istotnych cech wykorzystywanej przez nas metody *C-value* jest zwrócenie uwagi na istnienie fraz terminologicznych, które jako maksymalne frazy występują w tekstach albo bardzo rzadko albo wcale. Cecha ta jest o tyle ważna, że korpusy dziedzinowe zwykle nie są bardzo duże, wobec tego uwzględnienie w procesie ekstrakcji terminologii jedynie fraz maksymalnych może prowadzić do pominięcia istotnych terminów. W metodzie tej zaproponowano więc, by we frazach maksymalnych rozpoznawać wszystkie poprawne gramatycznie frazy wewnętrzne. Podejście takie prowadzi jednak do tworzenia fraz semantycznie niepoprawnych, które powstają przez obcięcie istotnego elementu frazy. Przykładowo, we frazie *druga wojna światowa* możemy wyróżnić trzy poprawne gramatycznie terminy. Są to: *druga wojna*, *wojna* oraz *wojna światowa*. O ile dwie ostatnie frazy są poprawnymi terminami o tyle pierwsza z nich nie powinna być rozważana jako termin. W pracy (Marciniak i Mykowiecka 2015) zaproponowaliśmy metodę pozwalającą na wyeliminowanie lub znaczne obniżenie rangi takich „urwanych" fraz. Metoda ta dzieli każdą maksymalną frazę na co najwyżej dwie frazy zagnieżdżone. Jako miejsce podziału wybiera najsłabsze ogniwo w danej frazie. Wybór ten dokonywany jest na podstawie siły powiązań poszczególnych par słów (bigramów) liczonych dla całego korpusu na podstawie lematów słów. Siła wiązania jest ustalana przy pomocy współczynnika NPMI (ang. Normalised Pointwise Mutual Information, znormalizowana punktowa informacja wzajemna) zaproponowanego w (Bouma 2009). Dla bigramu „x y", współczynnik ten jest liczony według poniższego wzoru, gdzie x i y są lematami kolejnych słów w tekście; p(x) i p(y) to częstość wystąpienia lematów w rozważanym korpusie a p(x,y) to częstość bigramu „x y".

$$NPMI(x,y) = \ln\frac{p(x,y)}{p(x)*p(y)} / -ln(p(x,y))$$

Współczynnik NPMI jest tym większy im częściej słowa x i y występują obok siebie. Jeśli słowa te zawsze ze sobą sąsiadują, cały współczynnik równy jest 1. Zaproponowana metoda pozwala na wyeliminowanie fraz *druga wojna* oraz *pierwsza wojna* z ogólnodostępnego (http://zil.ipipan.waw.pl/Korpus%20plWikiEcono) korpusu tekstów ekonomicznych, gdyż wartość NPMI dla powyższych biogramów wynosi odpowiednio 0.36 oraz 0.25 podczas gdy dla bigramu *wojna światowa* jest to 0.79.

## 2.5 Analiza kontrastywna

Opisana powyżej metoda szeregowania fraz pozwala na selekcję z tekstu częstych fraz o strukturze odpowiadającej terminom dziedzinowym. Aby jednak otrzymany zbiór terminów był wystarczająco obszerny, trzeba przetworzyć stosunkowo dużo danych, a im większy jest zbiór, tym większa jest szansa, że składające się na niego teksty zawierają istotną liczbę wystąpień terminów z różnych innych dziedzin. W analizowanym zbiorze tekstów dotyczących muzyki, takimi terminami są przykładowo *silna osobowość*, *stary kontynent* czy *wielkie szczęście*. Metodą pozwalającą na odrzucenie terminów należących do innych dziedzin tematycznych jest porównanie list terminów otrzymanych dla różnych kolekcji tekstów. Dobór tych kolekcji ma istotny wpływ na wynik porównania. Jeżeli jako korpus kontrastywny wybierzemy zbiór tekstów ogólnych, mamy szanse na wyeliminowanie terminów typu *własny sposób* czy *łatwe zadanie*. Natomiast jeżeli chcemy wyeliminować terminy specjalistyczne z innej, pokrewnej dziedziny, powinniśmy analizować zbiory tekstów z tego właśnie zakresu. Samo porównanie częstości występowania terminu w różnych zbiorach danych nie daje pożądanych rezultatów, gdyż zbiory te mają z reguły różną wielkość. W literaturze opisanych jest kilka strategii porównywania terminów pochodzących z analizy różnych zbiorów tekstów. Każda z tych metod opiera się na innej mierze różnic w rozkładzie terminów w porównywanych tekstach. W przygotowanym programie zaimplementowano trzy z nich. Metoda pierwsza (Rayson i Garside 2000) wykorzystuje współczynnik LL (logarytm wiarygodności, ang. Log-Likelihood) mówiący na ile różni się częstość konkretnego terminu w dwóch porównywanych korpusach. Druga metoda (Bonin i inni 2010) łączy częstość występowania w korpusie dziedzinowym z odwrotną częstością występowania w korpusie ogólnym (liczoną jako stosunek wielkości korpusu do częstości badanego terminu).

Trzeci sposób porównywania (CSmw) (Basili i inni 2001) przeznaczony dla terminów wielowyrazowych, uwzględnia nie tylko częstość występowania pełnych terminów, ale też częstość występowania słów stanowiących element główny badanej frazy. W tym współczynniku częstość terminu jest wygładzona tak, by zmniejszyć wariancje wyników dla terminów rzadkich. Niezależnie od wielości podejmowanych prób, brak jest jednak uniwersalnej, skutecznej metody rozróżniania terminów występujących rzadko, od terminów pochodzących z innej dziedziny.

## 3 Program TermoPL

TermoPL [1] jest programem do automatycznego wyszukiwania terminów ze zbioru dokumentów dotyczących wybranej dziedziny wiedzy. Przed uruchomieniem programu należy przetworzyć teksty dokumentów przez tager języka polskiego [2], który podzieli wejściowy tekst na tzw. tokeny, czyli słowa, znaki interpunkcyjne oraz inne ciągi znaków, przypisując im ujednoznacznioną interpretację morfologiczną. Wyniki działania tagera powinny być zapisane w pliku tekstowym, w którym zastosowano system kodowania znaków UTF-8. Program akceptuje trzy formaty plików wejściowych: format przyjęty w Narodowym Korpusie Języka Polskiego (NKJP) (Przepiórkowski i inni 2012), XCES oraz następujący prosty format:

```
forma #lemat #tag#,
```

w którym każdy token opisany jest przez jedną linię tekstu składającą się z jego formy ortograficznej, lematu oraz interptetacji morfologicznej opisanej przy pomocy znaczników przyjętych w NKJP. W formacie tym poszczególne zdania oddzielone są specjalną linią tekstu:

```
& #& #interp#.
```

TermoPL przetwarza pliki wejściowe zdanie po zdaniu wyszukując w nich najdłuższe ciągi tokenów zgodne z zadaną gramatyką. Użytkownik może wybrać standardową gramatykę wbudowaną w program lub zdefiniować swoją własną. Zbiór reguł określających gramatykę standardową jest następujący:

```
NPP : $NAP NAP_GEN*;
NAP[agreement] : AP* N AP*;
NAP_GEN[case = gen] : NAP;
AP : ADJ | PPAS | ADJA DASH ADJ;
N[pos = subst, ger];
ADJ[pos = adj];
ADJA[pos = adja];
PPAS[pos = ppas];
DASH[form = "-"];
```

[1] Program TermoPL można pobrać ze strony http://zil.ipipan.waw.pl/TermoPL.
[2] Przykładowo, można skorzystać z serwisu demo.clarin-pl.eu/demo/tager.html#

Symbole NAP i NAP_GEN oznaczają frazy rzeczownikowe, z tą jednak różnicą, że NAP_GEN jest frazą rzeczownikową w dopełniaczu ([case = gen]). Zakłada się również, że tokeny dopasowane do wzorca NAP (a także do NAP_GEN) zgadzają się między sobą co do liczby, rodzaju i przypadka ([agreement]). Standardowa gramatyka rozpoznaje frazy składające się z rzeczownika (N[pos = subst, ger]), przed którym i po którym może wystąpić dowolna liczba modyfikatorów przymiotnikowych (AP), np. *miejska bibioteka publiczna*. Frazy te z kolei mogą być modyfikowane przez dowolną liczbę tego samego typu fraz w dopełniaczu, np. *Sekretariat Naczelnika Wydziału Działalności Gospodarczej Komendy Powiatowej Policji*.

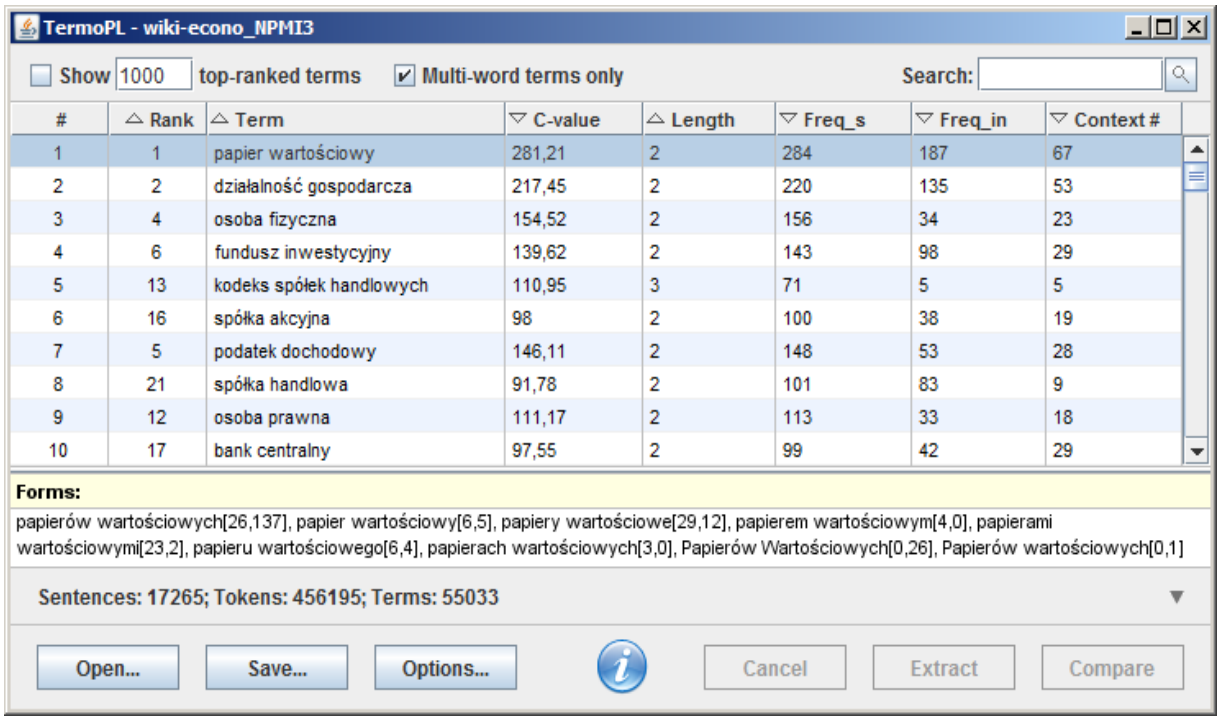
Znalezione w tekście frazy pasujące do tego opisu, po przekształceniu na uproszczoną formę podstawową, stają się początkową listą kandydatów na terminy. W następnym kroku, program odnajduje w rozpoznanych wcześniej frazach inne podfrazy zgodne z przyjętą gramatyką stosując rekurencyjnie jedną z dostępnych metod selekcji fraz zagnieżdżonych, o których mowa była w rozdziale 2.4. Wygenerowane w ten sposób frazy dołączane są do listy kandydatów.

Użytkownik może samodzielnie zdefiniować listy słów i fraz, które nie są uwzględniane przez program. Mogą być to frazy zawierające określone słowa (np. *ten, taki, jaki*), zawierające przyimki złożone (np. *na modłę, pod płaszczykiem, u podstaw*), lub będące terminami ogólnymi (np. *aktualny stan, druga połowa, mała zmiana*).

Zidentyfikowane w tekście terminy mogą zostać przekształcone z formy uproszczonej na formę podstawową. Przykładowo sekwencja *prawo karny* zamieniona zostaje na prawidłową formę *prawo karne*. Użytkownik ma wpływ na to, które tokeny zostaną finalnie przekształcone na formę podstawową. W standardowej gramatyce są to tylko tokeny dopasowane do symbolu NAP, o czym informuje program znak $ przed tym symbolem. Na ogół forma podstawowa terminu generowana przez program jest frazą w liczbie pojedynczej, chyba że wszystkie formy tego terminu w badanym korpusie są w liczbie mnogiej (np. *drogi oddechowe*).

Po skompletowaniu pełnej listy potencjalnych terminów, program szereguje je obliczając dla każdego z nich współczynnik *C-value*. Posortowana lista teminów wyświetlana jest na ekranie monitora w postaci tabeli, która oprócz formy uproszczonej lub podstawowej (Term), zawiera: kolejny numer terminu na liście (#), jego ocenę (Rank), *C-value* (C-value), długość (Length), całkowitą liczbę wystąpień (Freq_s), liczbę wystąpień w kontekście innych fraz (Freq_in) oraz liczbę tych kontekstów (Context #). Przykładowe rezultaty

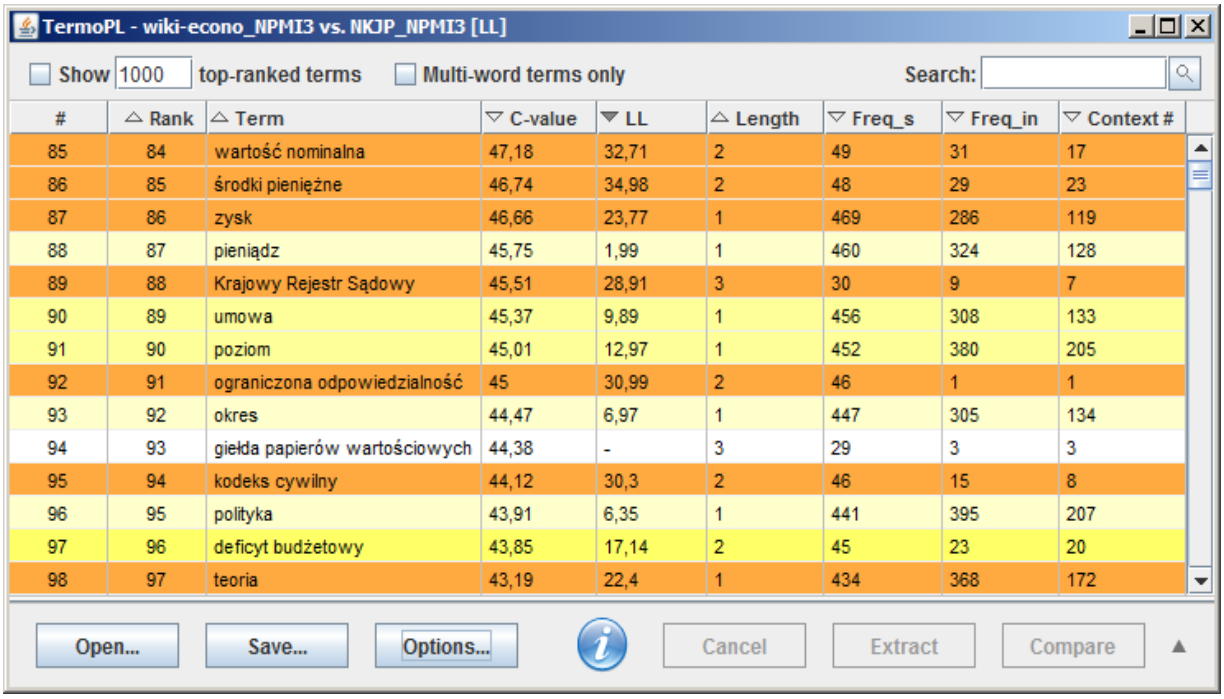
przedstawione są na rysunku 1. Zawierają one tylko formy wielowyrazowe, gdyż wybrana została opcja *Multi-word terms only*.




| # | Rank | Term | C-value | Length | Freq_s | Freq_in | Context # |
|---|---|---|---|---|---|---|---|
| 1 | 1 | papier wartościowy | 281,21 | 2 | 284 | 187 | 67 |
| 2 | 2 | działalność gospodarcza | 217,45 | 2 | 220 | 135 | 53 |
| 3 | 4 | osoba fizyczna | 154,52 | 2 | 156 | 34 | 23 |
| 4 | 6 | fundusz inwestycyjny | 139,62 | 2 | 143 | 98 | 29 |
| 5 | 13 | kodeks spółek handlowych | 110,95 | 3 | 71 | 5 | 5 |
| 6 | 16 | spółka akcyjna | 98 | 2 | 100 | 38 | 19 |
| 7 | 5 | podatek dochodowy | 146,11 | 2 | 148 | 53 | 28 |
| 8 | 21 | spółka handlowa | 91,78 | 2 | 101 | 83 | 9 |
| 9 | 12 | osoba prawna | 111,17 | 2 | 113 | 33 | 18 |
| 10 | 17 | bank centralny | 97,55 | 2 | 99 | 42 | 29 |

Rysunek 1. Fragment listy terminów dla korpusu wiki-ekono.

Wyodrębniony zbiór terminów można zapamiętać w pliku tekstowym i użyć do porównania z innym zbiorem terminów, o czy była mowa w podrozdziale 2.5. W tabeli rezultatów pojawi się wówczas nowa kolumna zawierająca obliczony współczynnik istotności terminu w korpusie. Poszczególne wiersze tabeli zostaną także oznaczone różnymi odcieniami kolorów w zależności od tego, na ile reprezentatywny jest dany termin w badanych korpusach. Im ciemniejszy (bardziej nasycony) jest kolor, tym bardziej dany termin jest reprezentatywny dla jednego z dwóch porównywanych korpusów. Rożne odcienie koloru żółtego wskazują na to, że termin jest relatywnie częściej używany w aktualnie badanym korpusie. Zieleń oznacza sytuację odwrotną, tzn. że dany termin jest bardziej charakterystyczny dla korpusu porównawczego. Przykładowe wyniki porównania za pomocą metody LL pokazane są na rysunku 2.




| # | Rank | Term | C-value | LL | Length | Freq_s | Freq_in | Context # |
|---|---|---|---|---|---|---|---|---|
| 85 | 84 | wartość nominalna | 47,18 | 32,71 | 2 | 49 | 31 | 17 |
| 86 | 85 | środki pieniężne | 46,74 | 34,98 | 2 | 48 | 29 | 23 |
| 87 | 86 | zysk | 46,66 | 23,77 | 1 | 469 | 286 | 119 |
| 88 | 87 | pieniądz | 45,75 | 1,99 | 1 | 460 | 324 | 128 |
| 89 | 88 | Krajowy Rejestr Sądowy | 45,51 | 28,91 | 3 | 30 | 9 | 7 |
| 90 | 89 | umowa | 45,37 | 9,89 | 1 | 456 | 308 | 133 |
| 91 | 90 | poziom | 45,01 | 12,97 | 1 | 452 | 380 | 205 |
| 92 | 91 | ograniczona odpowiedzialność | 45 | 30,99 | 2 | 46 | 1 | 1 |
| 93 | 92 | okres | 44,47 | 6,97 | 1 | 447 | 305 | 134 |
| 94 | 93 | giełda papierów wartościowych | 44,38 | - | 3 | 29 | 3 | 3 |
| 95 | 94 | kodeks cywilny | 44,12 | 30,3 | 2 | 46 | 15 | 8 |
| 96 | 95 | polityka | 43,91 | 6,35 | 1 | 441 | 395 | 207 |
| 97 | 96 | deficyt budżetowy | 43,85 | 17,14 | 2 | 45 | 23 | 20 |
| 98 | 97 | teoria | 43,19 | 22,4 | 1 | 434 | 368 | 172 |

Rysunek 2. Fragment listy terminów z korpusu wiki-ekono porównanego z korpusem 1M NKJP.

## 4 Podsumowanie

Każda dziedzina wiedzy czy forma komunikacji posiada własne, charakterystyczne dla niej, słownictwo. W artykule omówiliśmy problemy związane z ekstrakcją terminologii z tekstów dziedzinowych w języku polskim oraz przedstawiliśmy program TermoPL służący do realizacji tego zadania[3]. Wykorzystaliśmy w nim ogólnie uznaną metodę *C-value*, która wymagała dostosowania do fleksyjnego charakteru języka polskiego. Wprowadziliśmy również modyfikację w procesie rozpoznawania fraz zagnieżdżonych, która ogranicza rozpoznawanie fraz „urwanych". Omówiony program powstał w ramach realizacji projektu ClarinPL[4].

[3] Program zaprezentowany w niniejszym artykule zawiera implementację metod opracowywanych w ciągu paru ostatnich lat wspólnie przez Małgorzatę Marciniak i Agnieszkę Mykowiecką. Małgorzata Marciniak jest główną autorką rozdziałów 1, 2.1, 2.3 i 2.4. Agnieszka Mykowiecka jest główną autorką rozdziałów 2.2 i 2.5. Piotr Rychlik zaimplementował prezentowaną wersję programu i jest głównym autorem rozdziału 3. Wszyscy współautorzy mają swój wkład w ostateczną redakcję całości tekstu.

[4] Praca finansowana w ramach wkładu krajowego na rzecz udziału w międzynarodowym przedsięwzięciu „*CLARIN ERIC: Wspólne zasoby językowe i infrastruktura technologiczna*".

## 5 Bibliografia


Basili R., Moschitti A., Pazienza M. T., Zanzotto F. M. 2001: „A contrastive approach to term extraction.” *Terminologie et intelligence artificielle. Rencontres*, s. 119-128.

Bonin F., Dell'Orletta F., Venturi G., Montemagni S. 2010: „A contrastive approach to multi-word term extraction from domain corpora.” W *Proceedings of the 7th International Conference on Language Resources and Evaluation*, s. 19–21.

Bouma G. 2009: „Normalized (Pointwise) Mutual Information in Collocation Extraction.” W *Proceedings of the Biennial GSCL Conference*, Tübingen, s. 31–40.

Doroszewski W. (red.) 1958-1969: *Słownik języka polskiego,* PWN, Warszawa.

Drabik L., Kubiak-Sokół A., Sobol E. 2012: *Słownik języka polskiego PWN*, PWN, Warszawa.

Frantzi K., Ananiadou S., Mima H. 2000: „Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method.” W *Int. Journal on Digital*, s. 115–130.

Gęsicki Ł., Gęsicki M. 1996: *Słownik terminów ekonomiczno-prawnych.* Interfart.

Głowiński M., Kostkiewiczowa T., Okopień-Słowińska A. 2010: *Słownik terminów literackich,* Ossolineum.

Justeson J. S., Katz S. M. 1995: „Technical terminology: some linguistic properties and an algorithm for identification in text.”, *Natural Language Engineering*, s. 9–27.

Korkontzelos I., Klapafti I. P., Manandhar, S. 2008: „Reviewing and Evaluating Automatic Term Recognition Techniques.” *Advances in Natural Language Processing, Lecture Notes in Artificial Intelligence, Volume 5221*, s. 248–259.

Kozankiewicz S. (red.) 1976: *Słownik terminologiczny sztuk pięknych.* Państwowe Wydawnictwo Naukowe, Warszawa.

Marciniak M., Mykowiecka A. 2015: „Nested Term Recognition Driven by Word Connection Strength.” *Terminology* 21, nr 2, s. 180–204.

Pazienza, M. T., Pennacchiotti, M., Zanzotto, F. M. 2004: „Terminology Extraction: An Analysis of Linguistic and Statistical Approaches.”, *Knowledge Mining, Proceedings of the NEMIS 2004 Final Conference, Studies in Fuzziness and Soft Computing, Volume 185,* s. 255–279.

Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. 2012: *Narodowy Korpus Języka Polskiego*, PWN, Warszawa.

Rayson P., Garside R. 2000: „Comparing Corpora Using Frequency Profiling.” W *Proceedings of the Workshop on Comparing Corpora*, *Volume 9*, Association for Computational Linguistics, s. 1–6.

Sager J. C. 1990: *A Practical Course in Terminology Processing*, John Benjamins, Amsterdam, Philadelphia.

Savova G. K., Harris M., Johnson T., Pakhomov S. V., Chute C. G. 2003: „A Data-Driven Approach for Extracting <<the Most Specific Term>> for Ontology Development.” W *Proceedings of AMIA, Annual Symposium*, s. 579–583.

Wermter J., Hahn U. 2005: „Massive Biomedical Term Discovery.” W *Discovery Science, volume 3735 of Lecture Notes in Computer Science*, s. 281–293.

Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik

Streszczenie

Każda dziedzina wiedzy czy forma komunikacji posiada własne, charakterystyczne dla niej, słownictwo. Słowniki zawierające słowa i wielowyrazowe terminy dziedzinowe służące identyfikacji istotnych pojęć i ich leksykalnych odpowiedników w tradycyjnym podejściu tworzone były przez zajmujących się tą dziedziną specjalistów. Metoda ta jest jednak bardzo czasochłonna a zatem, zwłaszcza dla szybko zmieniających się dziedzin, niewystarczająca. W niniejszej pracy prezentujemy program pozwalający na automatyczną identyfikację fraz rzeczownikowych będących potencjalnymi terminami dziedzinowymi oraz ich uporządkowanie według miary przybliżającej stopień ich istotności.

Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik
Automatic extraction of domain terminology from text corpora
key words: domain terminology, text corpora

Abstract

Every knowledge domain or a form of communication has its own characteristic vocabulary. In traditional approach, dictionaries containing words and multi-word terms identifying important concepts and their lexical equivalents were created by specialists in a subject area. This method, however, is very time-consuming and therefore inadequate, especially for rapidly changing domains. In this paper we present a computer program allowing for automatic identification of noun phrases being a potential domain terms and their ranking according to some measure of significance.