

POLITECHNIKA POZNAŃSKA
Wydział Informatyki i Zarządzania
Instytut Informatyki

Praca magisterska

Zastosowanie ontologii do odkrywania wiedzy

Dariusz Daćko

Promotor:

dr hab. inż. Joanna Józefowska, prof. PP

Poznań 2007

(karta pracy dyplomowej)

Spis treści

Spis treści	5
1 Wstęp	9
2 Biblioteki i katalogi przedmiotowe	15
2.1 Opis przedmiotowy publikacji	16
2.2 Niedostatki katalogu przedmiotowego	17
3 Transformacja słownika haseł do tezaurusa	19
3.1 Kartoteka haseł przedmiotowych i formalnych	19
3.2 Słownik haseł przedmiotowych	21
3.2.1 Typy haseł	22
3.2.2 Zawartość słownika	23
3.3 Systemy organizacji wiedzy	24
3.4 Koncepcja rozwiązania	25
3.5 Inne możliwe zastosowania tezaurusa	25
4 Projekt	29
4.1 Język haseł przedmiotowych KABA	29
4.2 Budowanie niejawnych hiperonimii	32
4.2.1 Wiązanie haseł z ich leksemami	33
4.2.2 Wiązanie haseł z ich dopowiedzeniami	34
4.2.3 Wiązanie nazw zależnościowych	34
4.2.4 Wiązanie hasła z hasłami zaczynającymi się nim	36
4.2.5 Relacje równoważności	36
4.2.6 Odsyłacze orientacyjne uzupełniające	37
4.2.7 Inne reguły	38
4.3 Zastosowania hierarchii tezaurusa	39
4.3.1 Schemat bazy danych	39
4.3.2 Przeglądarka tezaurusa	41
4.3.3 Wyszukiwarka książek	41
4.3.4 Badania statystyczne tematyki książek	42

5	Implementacja katalogu przedmiotowego	45
5.1	Zastosowane technologie	45
5.2	Architektura systemu	46
5.3	Odmiana rzeczowników przez liczbę	48
5.4	Kartoteka haseł wzorcowych	51
5.4.1	Opis funkcji kartoteki i jej klas	52
5.4.2	Błędy w słowniku KABA	62
5.4.3	Nazwy geograficzne	63
5.4.4	Liczba gramatyczna haseł	66
5.4.5	Indeks form haseł	68
5.4.6	Reprezentacja tekstowa	72
5.4.7	Wywołanie programu	77
6	Implementacja budowy niejawnych hiperonimii	83
6.1	Opis klas i ich metod	84
6.2	Wiązanie haseł z ich leksemami	87
6.2.1	Opis działania	87
6.2.2	Możliwe udoskonalenia	89
6.2.3	Analiza wyników	90
6.2.4	Przykłady powiązań	91
6.3	Wiązanie haseł z ich dopowiedzeniami	92
6.3.1	Opis działania	92
6.3.2	Możliwe udoskonalenia	94
6.3.3	Analiza wyników	95
6.3.4	Przykłady powiązań	97
6.4	Wiązanie nazw zależnościowych	98
6.4.1	Opis działania	98
6.4.2	Możliwe udoskonalenia	99
6.4.3	Analiza wyników	99
6.5	Wiązanie hasła z hasłami zaczynającymi się nim	100
6.5.1	Odmiana przez liczbę hasła podstawowego w hasle złożonym .	101
6.5.2	Niepewność relacji	102
6.5.3	Opis działania	103
6.5.4	Możliwe udoskonalenia	104
6.5.5	Analiza wyników	105
6.5.6	Przykłady powiązań	105
6.6	Propozycje pozostałych reguł wiązania haseł	106
6.6.1	Wiązanie określników z tematami	106
6.6.2	Wiązanie części obiektu geograficznego z jego całością	106
6.6.3	Zastępowanie leksemu przez leksem szerszy	107

<i>Spis treści</i>	7
6.7 Pozostałe czynności do wykonania	107
6.8 Wnioski	109
7 Podsumowanie	111
Literatura	113
A Zawartość płyty CD	115

Rozdział 1

Wstęp

Rozwój sztucznej inteligencji sprawił, że coraz częściej buduje się systemy oparte na wiedzy oraz ją odkrywające. Jedną z możliwych definicji określa wiedzę jako: „ogół wiarygodnych informacji o rzeczywistości wraz z umiejętnością ich wykorzystywania” [9]. Inaczej mówiąc wiedza to pełna, pewna i zrozumiała informacja. Te trzy dodatkowe cechy sprawiają, że systemy wykorzystujące wiedzę powinny działać lepiej od systemów wykorzystujących jedynie informację. Podobnie odkryta wiedza powinna być bardziej użyteczna od informacji.

Jednym ze sposobów reprezentacji wiedzy jest ontologia. Niestety jej wykorzystanie w systemach informatycznych jest trudne. Obecnie analiza wypowiedzi w języku naturalnym przy użyciu ontologii daje wyniki niejednoznaczne i niepewne, ze względu na wieloznaczność wyrażen języka naturalnego. Wydaje się, że powodem trudności w wykorzystaniu ontologii do analizy języka naturalnego jest traktowanie wypowiedzi w sposób za mało znaczeniowy, ograniczając proces zrozumienia wyrazu jedynie do wiedzy czy składa się on z tych samych liter co zapamiętany wzorzec.

Natomiast wydaje się, że nie powinno być trudności podczas analizy wyrażen zapisanych przy użyciu słownictwa kontrolowanego, gdyż w przeciwieństwie do języka naturalnego takie słownictwo jest jednoznaczne. Słownictwo kontrolowane jest zbiorem słów, które posiadają jednoznaczną i niepowtarzalną definicję [15]. Dzięki temu ich wystąpienie w wypowiedzi może być bezbłędnie i jednoznacznie rozpoznane. Jeśli pewien wyraz języka naturalnego posiada kilka znaczeń to, dla zachowania jednoznaczności, do słownika kontrolowanego dodaje się go z pewną informacją jednoznacznie określającą znaczenie. Może to być na przykład dziedzina pojęcia podana w nawiasach po właściwej nazwie pojęcia. Jeśli dwa wyrazy języka naturalnego mają taką samą definicję (takie samo znaczenie), to do słownika dodawany jest tylko jeden wyraz (tak zwany termin przyjęty), a pozostałe wyrazy (terminy odrzucone) uznawane są za jego alias. W wyrażeniach można używać jedynie terminów przyjętych.

Obecnie słownictwa kontrolowanego używa się najczęściej w bibliotekach do opi-

su zawartości publikacji [15]. Publikacja opisywana jest przez bibliotekarzy jednym lub kilkoma słowami kluczowymi słownictwa kontrolowanego zbudowanego specjalnie z myślą o tym zastosowaniu. Przeglądając słowa kluczowe publikacji można w prosty sposób dowiedzieć się na jaki temat jest publikacja. Zaletą słownictwa kontrolowanego jest łatwość zbudowania systemu zwracającego bezbłędne odpowiedzi. Można na przykład skonstruować system wyszukujący publikacje na podany temat. System taki będzie posiadał bardzo wysoką precyzję, to znaczy wszystkie lub prawie wszystkie wyszukane publikacje będą na interesujący czytelnika temat – podany za pomocą słowa kluczowego.

Niestety zastosowanie słownictwa kontrolowanego zamiast przeszukiwania pełnotekstowego publikacji zmniejsza tak zwany współczynnik recall. Oznacza to, że część publikacji na interesujący czytelnika temat nie zostanie zwrócona. Wynika to z tego, że w opisie publikacji nie podaje się wszystkich słów kluczowych dotyczących publikacji, a jedynie te które najlepiej określają jej temat. Na przykład publikacja o aloesie zwyczajnym, czyli jednym z bardziej popularnych gatunków rodziny aloesów, będzie opisana jednym hasłem „Aloes zwyczajny”. Jeśli czytelnik chciałby odnaleźć publikacje o aloesie, to możliwe że będzie szukał hasła „Aloesy”, opisującego całą rodzinę. Niestety w takim przypadku znajdzie wszystkie publikacje o aloesach z wyjątkiem publikacji o aloesach zwyczajnych. Co więcej, prawdopodobnie nie będzie nawet świadomy niekompletności wyniku. Można temu zaradzić poszukując oprócz podanego słowa kluczowego także wszystkich jego hiponimów to jest terminów węższych. Wiedzę o hiponimach można zdobyć z ontologii zewnętrznej lub zbudowanej bezpośrednio na słowach słownictwa kontrolowanego. To drugie podejście ma tą zaletę, że unika się błędów i niejednoznaczności podczas przechodzenia między słownictwem kontrolowanym a słownictwem ontologii.

Podsumowując można stwierdzić, że stosując słownictwo kontrolowane uzyskuje się informację pewną. Stosując dodatkowo ontologię wydaje się, że można uzyskać system dający pełne lub przynajmniej o wiele pełniejsze odpowiedzi, dzięki czemu uzyskiwana informacja byłaby bardziej użyteczna.

Obecnie biblioteczne słownictwa kontrolowane posiadają przeważnie jedynie część hiperonimii to jest relacji termin szerszy – termin węższy. Najczęściej dla haseł nie podaje się hiperonimii, jeśli wynikają one z nazwy hasła. Na przykład nie znajdziemy w słowniku innej informacji o tym, że hasło „Siuksowie (Indianie)” jest hiponimem hasła „Indianie” niż informacja zawarta w tekście pierwszego hasła. Tekst haseł konstruuje się według określonych zasad nazywanych językiem haseł przedmiotowych. Wykorzystując te zasady oraz hiperonimie jawnie zapisane w słowniku jesteśmy w stanie utworzyć większość hiperonimii, które rzeczywiście istnieją między pojęciami reprezentowanymi przez hasła. Pozostałe, brakujące relacje można by w przyszłości dodać ręcznie, tak aby otrzymać pełną ontologię.

Budując system oparty na bibliotecznym słownictwie kontrolowanym oraz na ontologii, nie trzeba się ograniczać do wyszukiwania dokumentów. Wydaje się, że elementy te będzie można wykorzystać także w takich zastosowaniach eksploracji danych jak: hierarchiczne poszukiwanie zbiorów częstych, czy klasyfikacja, a także do obliczania innych statystyk.

Łatwo zauważyć analogię między eksploracją zasobów bibliotecznych a eksploracją zasobów Internetu. Obie dziedziny mają ze sobą wiele wspólnego, jednak jednocześnie zasoby biblioteczne posiadają kilka cech ułatwiających ich eksplorację. Tematy publikacji bibliotecznych, w przeciwieństwie do dokumentów w Internecie, opisane są słownictwem kontrolowanym. Wynika z tego kilka faktów. Po pierwsze, jak już wspomniano, możliwe jest użycie podczas analizy ontologii. Dzięki istnieniu zasad języka haseł przedmiotowych możliwa jest analiza morfologiczna oraz składowa hasła. Dzięki utworzeniu ontologii prawdopodobnie w przyszłości będzie możliwa także analiza semantyczna haseł. W przeciwieństwie do Internetu języki haseł przedmiotowych istniejące w poszczególnych krajach mają bardzo podobną budowę. Często mówi się o nich, że są kompatybilne z językiem Biblioteki Kongresu, tak zwanym „Library of Congress Subject Headings” [4]. Język używany w Polsce także jest kompatybilny z LCSH i nazywa się językiem KABA („Katalogi Automatyczne Bibliotek Akademickich”). Dzięki kompatybilności języków wdrożenie systemu w bibliotekach innych krajów powinno być stosunkowo łatwe.

Biblioteki cechują się mniejszym niż Internet rozmiarem danych w nich przechowywanych oraz ich mniejszą ulotnością. Dzięki temu analiza oraz poprawianie działania systemu powinno być łatwiejsze. W bibliotekach można zastosować prostszy algorytm obliczania popularności dokumentu niż w Internecie. Wystarczy jako popularność publikacji przyjąć miarę w pewien sposób zależną od częstości jej wypożyczeń przez czytelników. W wyszukiwarkach bibliotecznych, podobnie jak w internetowych, do określania pozycji dokumentu można by stosować personalizację użytkowników. Wszystkie wymienione wspólne cechy oraz różnice mogą sprawić, że „Semantic Library Catalogue” powstanie szybciej niż Semantic Web.

Celem pracy magisterskiej było zaprojektowanie i zaimplementowanie zastosowań ontologii przede wszystkim do semantycznych analiz statystycznych tematyki publikacji w bibliotekach. Efekty pracy powinny być możliwe do wdrożenia w Wielkopolskiej Bibliotece Cyfrowej [14] opartej na platformie dLibra. Pierwszym zadaniem na drodze do realizacji tego celu było zapoznanie się z semantycznymi funkcjami obecnie udostępnianymi przez biblioteki. Drugim zadaniem było zaprojektowanie semantycznych statystyk tematyki publikacji w bibliotece, takich jak znajdowanie tematów, na które występuje najwięcej książek w bibliotece. Trzecim zadaniem było wybranie najbardziej interesujących i możliwych do realizacji algorytmów oraz ich implementacja z wykorzystaniem istniejącej ontologii. Na końcu

należało ocenić wyniki działań tych algorytmów.

Po przeanalizowaniu semantycznych funkcji oferowanych obecnie przez biblioteki oraz ustaleniu brakujących funkcji, przyjrzano się możliwości ich realizacji z wykorzystaniem istniejących obecnie ontologii. Okazało się, że nie istnieją polskie ontologie¹. Jednak okazało się, że prawdopodobnie będzie możliwe przekształcenie bibliotecznego słownictwa kontrolowanego w ontologię. Dlatego też pojawiło się kolejne zadanie wynikające z trzeciego zadania z pierwotnego zbioru zadań. Było nim utworzenie między hasłami bibliotecznego słownictwa kontrolowanego KABA hiperonimii wynikających z języka haseł przedmiotowych i tym samym utworzenie ontologii. Zadanie to można podzielić na cztery podzadania. Po pierwsze należało przeanalizować słownik KABA oraz język haseł przedmiotowych KABA pod kątem możliwości przekształcenia słownika w ontologię. Następnie należało zaprojektować algorytmy budowy nie podanych jawnie hiperonimii. Trzecim podzadaniem było zaimplementowanie struktury programistycznej reprezentującej słownik KABA i umożliwiającej wykonywanie na nim operacji analitycznych. Natomiast czwartym podzadaniem było wykorzystanie tej struktury do implementacji algorytmów budowy niejawnych hiperonimii.

W trakcie realizacji trzeciego podzadania okazało się, że słownik KABA zawiera bardzo dużo błędów. Wynikało to z tego, że zawartość słownika KABA nigdy nie była komputerowo sprawdzana pod względem poprawności. Dlatego dodatkowo należało wykryć błędy znajdujące się w słowniku KABA.

Podsumowując praca magisterska składała się z następujących zadań:

- a) Analiza funkcjonalności systemu dLibra oraz innych bibliotek,
- b) Zaprojektowanie przykładowych semantycznych ulepszeń (oryginalnie tylko obliczania statystyk),
- c) Konwersja słownika KABA do ontologii, a w tym:
 - i) Analiza języka haseł przedmiotowych KABA,
 - ii) Zaprojektowanie algorytmów budowania niejawnych hiperonimii,
 - iii) Implementacja reprezentacji programistycznej słownika KABA,
 - iv) Wykrycie błędów w słowniku KABA,
 - v) Implementacja algorytmów budowania niejawnych hiperonimii,
- d) Implementacja wyżej wymienionych ulepszeń z wykorzystaniem ontologii,
- e) Ocena uzyskanych wyników.

Cztery główne zadania z powyższej listy występowały w pierwotnym zbiorze zadań do zrealizowania. Natomiast konwersja słownika do ontologii i jej podzadania wynikły z pierwotnego zbioru zadań.

Budowanie ontologii okazało się tak dużym przedsięwzięciem, że nie wystarczyło czasu na implementację zaprojektowanych zastosowań zbudowanej ontologii.

¹Pierwsze polskie ontologie zaczęły być tworzone dopiero w 2007 roku bazując na projekcie WordNet.

W rozdziale drugim opisano funkcje semantyczne oferowane obecnie w polskich bibliotekach dzięki polskiemu centralnemu katalogowi przedmiotowemu NUKAT. Przedstawiono także niedogodności związane z jego używaniem. W rozdziale trzecim opisano dokładniej katalog przedmiotowy, używane przez niego słownictwo kontrolowane, a także koncepcję rozwiązania problemów ukazanych w poprzednim rozdziale. W rozdziale czwartym przedstawiono projekt budowania ontologii na podstawie słownictwa kontrolowanego oraz projekty jej zastosowań w bibliotekach. Na początku rozdziału został omówiony język haseł przedmiotowych KABA potrzebny do zbudowania ontologii. W rozdziale piątym opisano architekturę systemu oraz stworzoną w ramach pracy magisterskiej bibliotekę programistyczną do pracy ze słownikiem haseł przedmiotowych. W ramach biblioteki zaimplementowano między innymi następujące funkcje:

- import rekordów słownika KABA zapisanych w formacie MARC,
- dzielenie nagłówek haseł przedmiotowych na jednostki języka haseł przedmiotowych,
- wykrywanie błędów w rekordach słownika KABA,
- obsługa specjalnych formatów złożonych haseł zawierających nazwy geograficzne,
- odmiana przez liczbę gramatyczną haseł będących rzeczownikami,
- budowanie indeksu wszystkich form, jakie mogą przyjmować hasła przedmiotowe w innych hasłach z nich złożonych.

Biblioteka ta umożliwi w miarę prosty sposób implementację algorytmów tworzenia powiązań hierarchicznych między hasłami. Implementacja tych algorytmów oraz analiza ich działania opisana jest w rozdziale siódmym. Pod koniec tego rozdziału opisano także propozycje udoskonalenia zbudowanej ontologii oraz jej analizę.

Ze względu na obszerność pracy magisterskiej w tekście pominięto następujące jej elementy:

- dokładny opis języka haseł przedmiotowych KABA (jest jedynie pobieżny),
- opis formatu MARC oraz opis biblioteki MARC4J do odczytu rekordów MARC,
- opis przebiegu importu słownika KABA, w tym szczegółowy opis dzielenia nagłówek na jednostki oraz opis metod wykrywania błędów w rekordach KABA i wykrytych błędów,
- przeprowadzone analizy statystyczne zawartości rekordów KABA,
- opis platformy dLibra, Wielkopolskiej Biblioteki Cyfrowej oraz pakietu reprezentującego i importującego dane z tej biblioteki,
- opis analizatora morfologicznego SAM-95 oraz dokładny opis metody jego wykorzystania do odmiany rzeczowników przez liczbę gramatyczną,
- projekt relacji hierarchicznych szczególnego typu dla nazw geograficznych,
- dokładny opis złożoności pamięciowej i obliczeniowej.

W trakcie pracy magisterskiej okazało się, że zastosowaną strukturę organizacji wiedzy powinno się raczej nazywać tezaurem a nie ontologią, choć ich dokładne definicje nie zostały jeszcze ustalone. W dalszej części pracy będzie używana nazwa tezaurus.

Rozdział 2

Biblioteki i katalogi przedmiotowe

W bibliotekach zasadniczo stosuje się dwa podejścia do opisu treści książek w celu ułatwienia ich wyszukiwania: klasyfikacje oraz katalogi przedmiotowe.

Klasyfikacja to struktura hierarchiczna dziedzin wiedzy. Struktura taka organizuje dziedziny o różnym poziomie szczegółowości. Na samej górze klasyfikacji są dziedziny najbardziej ogólne na przykład „Nauka”, poniżej znajdują się dziedziny bardziej szczegółowe takie jak: „Biologia”, „II Wojna Światowa”, a na samym dole są dziedziny najbardziej szczegółowe takie jak: „Układ pokarmowy” albo „Bitwy II Wojny Światowej”. Cechą wyróżniającą klasyfikacje jest ustrukturalizowanie pozwalające na zlokalizowanie interesującej nas dziedziny poprzez przejście struktury od góry do dołu. Klasyfikacja jest przeważnie mało szczegółowa, gdyż nie grupuje tematów czy pojęć a jedynie ich klasy. Najczęściej stosowanymi klasyfikacjami bibliotecznymi są: Uniwersalna Klasyfikacja Dziesiętna (UKD, ang. Universal Decimal Classification, UDC) [16], Klasyfikacja Dziesiętna Deweya (ang. Dewey Decimal Classification, DDC) oraz Klasyfikacja Biblioteki Kongresu Stanów Zjednoczonych (ang. Library of Congress Classification, LCC).

Opisanie książki przy pomocy klasyfikacji polega na przyporządkowaniu jej do klasy najlepiej opisującej jej tematykę. Celem klasyfikacji jest umożliwienie odnalezienia książek z interesującej nas dziedziny. Dzięki istnieniu hierarchii dziedzin możemy odnaleźć interesującą nas dziedzinę przeglądając hierarchię pojęć.

Z kolei w katalogu przedmiotowym znajdują się nie ogólne klasy, ale ogólne oraz także bardziej szczegółowe tematy takie jak: „Biologia”, „Zapalenie żołądka i jelita”, „Jelito ślepe”, „Wojna światowa (1939-1945)”, „Bitwa o Warszawę (1945)”. Wyszukiwanie interesującego nas tematu polega na znalezieniu jego hasła w indeksie alfabetycznym, a nie jak w klasyfikacji poprzez przeglądanie hierarchii. Katalog przedmiotowy dzięki swojej szczegółowości nadaje się bardzo dobrze do dokładnego opisu tematu książek i takie też jest jedno z jego głównych zadań. Drugim zadaniem jest umożliwienie odnalezienia książek na wybrany temat. Ze względu na swoją szczegółowość w poszczególnych krajach przeważnie istnieją tylko jeden

lub dwa katalogi przedmiotowe. W Stanach Zjednoczonych używa się katalogu Biblioteki Kongresu (ang. Library of Congress Subject Headings, LCSH), we Francji – katalogu RAMEAU. W Polsce najczęściej używa się katalogu KABA (Katalogi Automatyczne Bibliotek Akademickich). W Bibliotece Narodowej używa się odrębnego katalogu – katalogu Biblioteki Narodowej. Niektóre biblioteki specjalistyczne posiadają odrębne katalogi przedmiotowe zawierające słownictwo dostosowane do ich potrzeb. Na przykład biblioteki medyczne na całym świecie, w tym w Polsce, korzystają z katalogu nazywanego MeSH (Medical Subject Headings).

Opisanie książki w katalogu przedmiotowym polega na przyporządkowaniu jej hasła opisującego jej temat. W przypadku gdy nie ma takiego hasła w katalogu, należy je utworzyć. Hasła katalogu przedmiotowego często są powiązane relacjami hierarchicznymi i skojarzeniowymi. Ich celem jest przekierowanie czytelnika do terminów szerszych, węższych lub podobnych w przypadku, gdy hasło wybrane w indeksie alfabetycznym nie określa tematu jaki miał na myśli czytelnik. W przeciwieństwie do klasyfikacji relacje hierarchiczne nie są wykorzystywane do lokalizacji haseł poprzez przechodzenie od góry do dołu. Jednym z powodów takiego stanu musi być to, że takie przejście składałoby się ze zbyt dużej liczby kroków. Szczegółowość katalogu przedmiotowego sprawia, że relacje hierarchiczne i skojarzeniowe między jego hasłami wykorzystuje się jedynie w celach pomocniczych.

Klasyfikacją stosowaną w Polsce jest UKD, jednak posiada ona mało poziomów szczegółowości i jest stosowana jedynie w mniejszych, nienaukowych bibliotekach. Wobec tego praktycznie jedynymi strukturami organizacji tematyki książek w Polsce są wspomniane dwa katalogi przedmiotowe. Jak już napisano jeden z nich stosuje się dla zbiorów Biblioteki Narodowej. Drugi stosowany jest w pozostałych większych bibliotekach naukowych i akademickich. Biblioteki cyfrowe takie jak Wielkopolska Biblioteka Cyfrowa [14] oparta na platformie dLibra [2] które były początkowym obszarem zainteresowania w ramach pracy magisterskiej nie stosują żadnych systemów organizacji wiedzy. Zapewne wynika to z faktu, że ze względu na małą liczbę publikacji w nich przechowywanych potrzeba organizacji tematyki książek nie była w nich do tej pory zauważana. Wydaje się, że w najbliższym czasie powinno się to zmienić.

2.1 Opis przedmiotowy publikacji

Każdą z książek znajdujących się w bibliotece opisuje się jednym lub kilkoma hasłami katalogu przedmiotowego [1]. Hasła te określają temat książki. Na przykład anglojęzyczna książka „No lights, no sirens”, będąca biografią policjanta patrolującego ulice Nowego Jorku, opisana jest trzema hasłami:

Cea, Robert (1961-) – pamiętnik

Korupcja w policji – Stany Zjednoczone

Przemoc policyjna – Stany Zjednoczone

gdyż przede wszystkim opisuje korupcję i przemoc w policji. Pierwsze hasło występuje dlatego, że Robert Cea jest dość znaną osobą. W opisie książek powinny pojawić się hasła dla wszystkich głównych i pobocznych tematów pojawiających się w książce. Wszystkie hasła którymi opisane są książki zindeksowane są w katalogu przedmiotowym książek. Dzięki temu możliwe jest odnalezienie wszystkich książek na dany temat – na przykład o korupcji w policji. Wystarczy odnaleźć w katalogu przedmiotowym kartę szukanego hasła i przejrzeć karty wszystkich książek, które są umieszczone za kartą hasła. Katalogi komputerowe (OPAC) nie różnią się sposobem działania od katalogów tradycyjnych. Zmienił się w nich jedynie nośnik informacji z kart na rekordy.

Ze względu na dużą liczbę haseł w katalogu konieczne było wprowadzenie reguł określających sposób budowy haseł. Bez nich forma szukanego hasła byłaby trudna do przewidzenia przez szukającego go czytelnika i tym samym odnalezienie hasła w alfabetycznym indeksie byłoby utrudnione. Reguły te nazywają się językiem haseł przedmiotowych i zostaną omówione w rozdziale trzecim.

2.2 Niedostatki katalogu przedmiotowego

Używanie katalogów przedmiotowych w ich obecnym stanie często jest uciążliwe z dwóch powodów opisanych poniżej.

Co się stanie jeśli spróbujemy odnaleźć wszystkie książki, które poruszają tematykę policjantów? Postępując tak jak poprzednio, czyli znajdując hasło dotyczące policjantów oraz przeglądając wszystkie książki nim opisane, uzyskamy jedynie listę książek traktujących ogólnie o policjantach. Natomiast nie znajdziemy książek na takie tematy jak: kształcenie policjantów, policjantki, etyka zawodowa policjantów a w tym: przemoc i korupcja w policji; gdyż tematy te mają osobne karty. Na liście książek na temat policji nie znajdziemy więc bardzo znanej książki biograficznej Roberta Cea. Aby znaleźć wszystkie książki o policjantach, musimy odnaleźć w katalogu wszystkie hasła będące terminami węższymi hasła „Policjanci”, a następnie przejrzeć listę książek o tych hasłach, co może zająć bardzo dużo czasu. Pierwszym problemem może być określenie wszystkich haseł będących terminami węższymi interesującego nas hasła. Aby to zrobić, powinniśmy przejrzeć w katalogu wszystkie hasła sąsiadujące alfabetycznie z interesującym nas hasłem. Niektóre z nich mogą być terminami węższymi, na przykład hasła: „Policjanci – kształcenie” lub „Policjanci – deontologia”. Jeśli hasło ma terminy węższe „oddalone alfabetycznie” od niego, to wpisuje się je na karcie hasła, podając tym samym bezpośrednio jego hiponimy. Drugim problemem jest to, że musimy zlokalizować te hasła (co zajmuje dużo

czasu zarówno w kartkowym jak i komputerowym katalogu) oraz przejrzeć listy książek, często duplikujące się wzajemnie. Dodatkowo procedurę tą należy powtórzyć, jeśli jeden z odnalezionych terminów węższych sam posiada inne terminy węższe. Opisana procedura szukania zajmuje przeważnie dość dużo czasu.

Jeśli już nawet znajdziemy wszystkie książki na interesujący nas temat, to którą z nich powinniśmy wybrać? Jedną z możliwości jest złożenie zamówienia na wszystkie książki, przejrzanie ich zawartości oraz wybranie kilku z nich, które wypożyczymy i spróbujemy przeczytać. Prawdopodobnie ze względu na ograniczenie liczby zamawianych książek nie będziemy mogli zamówić wszystkich książek. Koniecznym może być kilkukrotne składanie zamówień na mniejszą liczbę książek. Bylibyśmy w o wiele prostszej sytuacji, gdybyśmy wiedzieli, które z książek na interesujący nas temat są warte wypożyczenia. Taką informację można by uzyskać z historii wypożyczeń książek – książki częściej wypożyczane w przeszłości są prawdopodobnie ciekawsze od pozostałych.

Opisane wady nie mogły być naprawione w katalogach kartkowych. Natomiast katalogi komputerowe pozwalają na automatyczne przetwarzanie danych przez komputer, mimo tego że nie jest to obecnie stosowane w bibliotekach. Komputer, wykorzystując umiejętności semantyczne oraz przetwarzania języka naturalnego, mógłby wykonywać uciążliwe operacje, które do tej pory musiał wykonać czytelnik. Wymienione metody są często stosowane w Internecie do organizacji wiedzy. Koncepcja ta zostanie szczerzej przedstawiona w następnym rozdziale.

Rozdział 3

Transformacja słownika haseł do tezaurusa

W aktualnym rozdziale zostanie omówiona koncepcja rozwiązania opisanych w poprzednim rozdziale niedostatków bibliotecznego katalogu przedmiotowego poprzez zbudowanie dla niego tezaurusa. Tezaurus w porównaniu do słownictwa kontrolowanego jest krokiem naprzód w kierunku organizacji wiedzy dzięki zawieraniu powiązań semantycznych (hierarchicznych i skojarzeniowych) między hasłami. Dzięki niemu komputer odciąży czytelnika od wykonywania czasochłonnych operacji logicznych podczas posługiwania się bibliotecznym katalogiem przedmiotowym.

Na początku zostanie szczegółowo omówiony katalog przedmiotowy oraz jego podstawowa część – słownik. Następnie zostaną omówione relacje hierarchiczne obecnie istniejące w katalogu przedmiotowym i różne systemy organizacji wiedzy. Po podaniu wymienionych informacji zostanie przedstawiona koncepcja rozwiązania. Na końcu zostaną podane inne możliwe problemy, jakie można rozwiązać w opisany sposób.

3.1 Kartoteka haseł przedmiotowych i formalnych

Katalog przedmiotowy składa się z kartoteki haseł przedmiotowych (khp) oraz list książek opisanych poszczególnymi hasłami [3]. W kartotece haseł przedmiotowych zapisane są wszystkie hasła, których można używać do opisu książek. Oprócz katalogu przedmiotowego w bibliotece istnieją także katalogi umożliwiające odnalezienie książki po jej tytule lub autorze. Takie katalogi wykorzystują kartotekę haseł formalnych (khf). Formalność oznacza, że hasła nie dotyczą przedmiotu książki a jej opisu formalnego (autor, tytuł, instytucje będące autorem zbiorowym). Cecha formalna jednej z książek może stać się cechą przedmiotową innej. Na przykład

istnieje wiele książek napisanych przez Adama Mickiewicza. „Adam Mickiewicz” będzie więc hasłem w kartotece haseł formalnych, umożliwiającym odnalezienie tych książek po ich autorze. Jednak w bibliotece znajduje się też trochę książek na temat osoby Adama Mickiewicza (na przykład o jego życiu lub twórczości). A więc „Adam Mickiewicz” będzie też hasłem w kartotece haseł przedmiotowych. Podobnie tytuł „Biblia” będzie zarówno hasłem kartoteki haseł formalnych jak i przedmiotowych. Jeśli pewne hasło występuje w obu kartotekach, to jego postać jest taka sama – mówi się, że hasła obu kartotek są ujednolicone a same kartoteki spójne. Obie kartoteki tworzą wspólnie jedną kartotekę, która w systemie KABA nazywa się Centralną Kartoteką Haseł Wzorcowych (CKHW). W dalszej części pracy będzie nas interesowała jedynie kartoteka haseł przedmiotowych, jednak będziemy uwzględniać występujące w niej ujednolicone tytuły dzieł i hasła osobowe.

W czerwcu 2006 roku kartoteka haseł przedmiotowych systemu KABA zawierała około 1,3 miliona rekordów, podczas gdy w grudniu 1999 roku jeszcze tylko 320 tysięcy rekordów. Interfejs WWW kartoteki haseł przedmiotowych systemu KABA umożliwia także dostęp do kartoteki haseł przedmiotowych Biblioteki Narodowej oraz polskiego katalogu MeSH, choć w niniejszej pracy nie będziemy się nimi zajmować.

Kartoteka haseł przedmiotowych ma za zadanie zindeksowanie wszystkich haseł użytych lub przygotowanych do użycia w opisie książek. Jej rolą jest umożliwienie zbudowania indeksu tych haseł i ich wyszukiwania przez czytelnika. Jednak jak już wcześniej wspomniano, ze względu na bardzo duży rozmiar khp wyszukiwanie haseł nie byłoby możliwe, gdyby nie tworzone ich według pewnych reguł. Jedną z podstawowych zasad języka haseł przedmiotowych jest tworzenie haseł o następującej budowie:

Hasło podstawowe (temat hasła) – pierwsze wyszczególnienie znaczenia tematu
[– drugie wyszczególnienie znaczenia tematu, ...].

Na przykład książka omawiająca korupcję w policji niezależnie od kraju będzie opisana hasłem „Korupcja w policji”, natomiast książka ograniczająca omówienie korupcji w policji do Polski będzie opisana hasłem podstawowym z tak zwanym określnikiem: „Korupcja w policji – Polska”. Określniki mogą ograniczać podstawowy temat między innymi do określonego miejsca, czasu lub dowolnego pojęcia jak w hasle: „Aeronautyka – przyrządy”. Hasła będące pojedynczymi tematami lub określnikami nazywane są hasłami prostymi, natomiast hasła złożone z tematu i określnika, albo też z kilku połączonych określników nazywane są hasłami rozwiniętymi.

3.2 Słownik haseł przedmiotowych

Jądrem i najważniejszą częścią kartoteki haseł przedmiotowych jest słownik haseł przedmiotowych (shp), zwany także słownikiem języka haseł przedmiotowych lub kartoteką haseł wzorcowych. Wszystkie trzy wyrazy: „słownik”, „języka” i „wzorcowych” podpowiadają rolę tej kartoteki. Słownik haseł przedmiotowych stanowi leksykę dla języka haseł przedmiotowych. Stanowi on zasób słownictwa (pojedynczych tematów i określników), na podstawie którego można tworzyć hasła rozwinięte. Jest on więc jak słownik języka naturalnego, na podstawie którego można tworzyć zdania, używając do tego języka naturalnego. Do shp należy też mała część wszystkich haseł rozwiniętych (tak zwane hasła rozwinięte słownikowe). Powód ich należenia do shp zostanie podany później. Natomiast do reszty khp (to znaczy bez shp) należą jedynie hasła rozwinięte. Przykładowo hasło „Korupcja w policji” należy do shp, ale już hasło „Korupcja w policji – Stany Zjednoczone” należy jedynie do khp. Wyraz „wzorcowych” w nazwie „kartoteka haseł wzorcowych” oznacza, że hasła te stanowią pewne wzorce, z których można budować hasła rozwinięte.

W czerwcu 2006 roku słownik haseł przedmiotowych zawierał około 112 tysięcy haseł, podczas gdy w 2000 roku – jedynie około 40 tysięcy haseł. Słownik haseł przedmiotowych zawiera dość bogate słownictwo, co widać chociażby po liczbie haseł w nim zawartych.

Analizując cechy kartoteki haseł przedmiotowych wystarczy ograniczyć się do jej słownika oraz języka. Słownik zostanie opisany bardziej szczegółowo poniżej, natomiast język – w rozdziale czwartym.

Hasła przedmiotowe są słownictwem kontrolowanym. Kontrolowanie słownictwa oznacza, że w opisie książek oraz w hasłach rozwiniętych można używać tylko tych haseł, które wcześniej zostały zatwierdzone przez odpowiednią instytucję zarządzającą słownikiem. Dla katalogu KABA taką instytucją jest NUKAT (Narodowy Uniwersalny Katalog Centralny) [8]. Każde kontrolowane hasło musi mieć jednoznaczną i niepowtarzalną definicję. Jeśli użyty wyraz w języku naturalnym ma więcej niż jedno znaczenie, należy po nim dodać tak zwane dopowiedzenie, czyli nazwę dziedziny w nawiasach, tak aby hasło jednoznacznie identyfikowało pojęcie. Na przykład w słowniku KABA istnieją następujące hasła: „Rak (choroba)” dla określenia nowotworu oraz „Rak (astrologia)” dla określenia gwiazdozbioru (zwierzęta określone są przez hasło „Raki”). Jeśli dane hasło może być określone także przez inne wyrazy, to podaje je się jako warianty hasła (tak zwane terminy odrzucone) i zapisuje w jego rekordzie. Na przykład wariantem hasła „Rak (choroba)” jest termin odrzucony „Nowotwory złośliwe”.

Rekord hasła zawiera także następujące informacje:

- typ hasła: hasło może być nazwą pospolitą, miejsca (geograficzną), nazwą osoby, instytucji, wydarzenia (tak zwanej imprezy) lub tytułem publikacji (nazw

pospolitych i geograficznych jest najczęściej), a także jednym z czterech określników: rzeczowym, geograficznym, chronologicznym lub formy,

- opis znaczenia hasła w języku naturalnym,
- hasło synonimiczne w słowniku LCSH i/lub RAMEAU,
- relacje skojarzeniowe do innych haseł,
- hiperonimy i hiponimy nie wynikające z języka haseł przedmiotowych,
- czasem tekst w sformalizowanym języku naturalnym wskazujący na niektóre hiponimy nie podane jawnie (tak zwany odsyłacz orientacyjny uzupełniający).

Rekordy haseł są zapisywane w formacie MARC [10, 5].

Słownik KABA jest kompatybilny ze swoim amerykańskim oraz francuskim odpowiednikiem, co oznacza, że ma on analogiczną strukturę. Dzięki temu dla wielu haseł KABA istnieją powiązania z synonimami w słowniku LCSH i RAMEAU.

Relacje skojarzeniowe oraz hiperonimy i hiponimy (zwane też terminami szerszymi i węższymi) kierują do haseł semantycznie skojarzonych z aktualnym hasłem. Jest to pomocne w przypadku, gdy wybraliśmy hasło które nie odpowiada dokładnie interesującemu nas tematowi albo też w przypadku, gdy w ogóle nie byliśmy w stanie znaleźć przy pomocy alfabetycznego indeksu odpowiedniego hasła. Przykładowo hasło „Policjanci” posiada termin szerszy „Policja”, węższy: „Policjanci – deontologia” oraz skojarzony: „Detektywi”.

Należy zwrócić uwagę na fakt, że w rekordzie zapisane są jedynie hiperonimy i hiponimy, które nie wynikają z języka haseł przedmiotowych. Na przykład w hasle „Policja” nie znajdziemy informacji o tym, że posiada ono termin węższy „Policja – uprawnienia”. Wynika to z faktu, że ostatnie hasło jest słownikowym hasłem rozwiniętym i relacja do niego wynika z jedynej dotąd opisanej reguły języka haseł przedmiotowych.

Hasła zapisane w słowniku powiązane są przez około 70 tysięcy hiperonimii.

Wspomniane przed chwilą hasło rozwinięte jest dobrym przykładem hasła rozwiniętego słownikowego. Hasła rozwinięte dodawane są do słownika wtedy, gdy wiążą się semantycznie z jednym z haseł w sposób nie wynikający z języka haseł przedmiotowych. Przykładowo hasło „Policja – uprawnienia” jest terminem szerszym hasła słownikowego „Kontrola tożsamości” oraz terminem węższym hasła „Bezpieczeństwo państwa”. Ponieważ relacje te nie wynikają z jhp, to hasło zostało dodane do słownika mimo tego, że jest rozwinięte.

3.2.1 Typy haseł

Każde hasło ma określony typ w celu łatwiejszego określenia jego znaczenia. Hasło może być tematem, tematem rozwiniętym jednym lub większą liczbą określników, a także pojedynczym określnikiem lub hasłem złożonym z kilku określników. Określniki należące do shp nazywane są określnikami swobodnymi, gdyż można je

łączyć z wieloma hasłami. W przeciwieństwie do nich niektóre hasła mogą być rozwinięte określnikami specyficznymi dla tych haseł rozwiniętych i nie występującymi w innych hasłach. Określniki takie nazywane są określnikami związanymi z pewnym hasłem rozwiniętym. Zdefiniowane są one nie w osobnym rekordzie, a w rekordzie hasła rozwiniętego w którym występują.

Z kolei temat może zawierać jeden lub więcej podtematów (a więc składać się z tematu nadrzędnego i tematów podrzędnych). Przykładem jest hasło „Polska. Polskie Siły Powietrzne. 316 Dywizjon Myśliwski Warszawski” określające jeden z dywizjonów Wojska Polskiego. Elementy tematu złożonego oddzielone są od siebie kropkami.

Typ całego hasła określany jest w następujący sposób:

- jeśli hasło zawiera temat (złożony lub pojedynczy), to typ hasła jest równy typowi ostatniego elementu tematu,
- jeśli hasło zawiera tylko określniki, to typ hasła równy jest typowi pierwszego określnika.

Słownik haseł przedmiotowych zawiera około 112 tysięcy haseł, w tym około 49 tysięcy nazw pospolitych (łącznie ze słownikowymi hasłami rozwiniętymi określnikiem geograficznym lub chronologicznym), 19 tysięcy nazw geograficznych, 10 tysięcy nazw instytucji, 27 tysięcy nazw osobowych, 500 nazw wydarzeń oraz 2 tysiące tytułów najważniejszych publikacji. Zawiera również około 4500 określników w tym: około 3500 określników rzeczowych, jeden określnik geograficzny, około 300 określników chronologicznych oraz 760 określników formy. Powód występowania jedynie jednego określnika geograficznego zostanie podany podczas opisu języka haseł przedmiotowych.

W tabeli 3.1 podano nieoczywiste informacje o poszczególnych typach oraz przykładowe hasła.

3.2.2 Zawartość słownika

Hasła tworzone są przeważnie wtedy, gdy istnieje książka o odpowiednim temacie. Oznacza to, że zawartość słownika odpowiada aktualnemu stanowi rozwoju wiedzy. Każde hasło znajdujące się w słowniku jest na tyle ważne, że można mu poświęcić przynajmniej część książki.

Hasła opisują zarówno tematy ogólne – na przykład „Nauki społeczne” jak i szczegółowe – na przykład „Wykrywacze kłamstw”. Hasła szczegółowe powinny być powiązane relacjami hierarchicznymi poprzez hasła pośrednie z hasłami na tematy ogólne. Na rysunku 3.1 przedstawiono około 50 haseł powiązanych tematycznie z hasłem „Policja” oraz hiperonimy tego hasła aż do hasła ogólnego „Nauki społeczne”.

Tabela 3.1: Typy haseł

Typ hasła	Uwagi	Przykładowe hasła
nazwa pospolita		Pancerniki (ssaki)
nazwa geograficzna		Polska
nazwa korporacyjna	nazwa własna opisująca instytucję; temat może być złożony z nazw korporacyjnych nadrzędnych i podrzędnych np. „Polska. Polskie Siły Powietrzne. 316 Dywizjon Myśliwski Warszawski”; pierwszy element nazwy korporacyjnej może być elementem geograficznym	Politechnika Poznańska
nazwa wydarzenia	temat może być złożony z wydarzeń nadrzędnych i podrzędnych	Kongres kardiologów
nazwa osobowa		Sienkiewicz, Henryk
tytuł	tytuł dzieła; może być tematem prostym, ale najczęściej jest złożonym – ostatni element jest tematowy a pierwszy osobowy, korporacyjny lub wydarzenia	Sienkiewicz, Henryk. Potop; Biblia
określnik rzeczowy	zawęza dziedzinę pojęcia określonego w temacie	– zastosowanie w przemyśle
określnik chronologiczny	zawęza czas pojęcia określonego w temacie	– 1890-1930
określnik formy	określa, że w hasle interesują nas publikacje określonego typu z podanej w temacie dziedziny	– encyklopedie
określnik geograficzny	zawęza miejsce pojęcia określonego w temacie	– Polska

3.3 Systemy organizacji wiedzy

W punkcie tym zostaną przedstawione następujące systemy organizacji wiedzy (SOW, ang. Knowledge Organization Systems): słownictwo kontrolowane, taksonomia, tezaurus i ontologia.

Definicja słownictwa kontrolowanego została podana już wcześniej. Słownictwo kontrolowane jest najprostszym z wymienionych SOW. Jego głównymi cechami są jednoznaczność słownictwa oraz relacje jedynie do synonimów. Taksonomia to słownictwo kontrolowane posiadające relacje hiperonimii. Tezaurus to taksonomia posiadająca także inne semantyczne relacje, na przykład skojarzeniowe. Ontologia jest tezaurusem, w którym hasła reprezentują poszczególne pojęcia realnego świata w ten sposób, że opisują także ich cechy. Na przykład dla pojęcia jabłka w ontologii podany jest także jego kształt. Ontologia powinna być w miarę wiernym odzwierciedleniem wiedzy o realnym świecie.

Na podstawie powyższych definicji widać, że słownik haseł przedmiotowych jest jedynie słownictwem kontrolowanym. Choć posiada on wszystkie rodzaje relacji występujących w teaurusie, to jedynie część hiperonimii jest w nim zapisana w postaci jawnej. Jednak ponieważ pozostałe hiperonimie wynikają z języka haseł przedmiotowych, to czytelnik przeglądający katalog przedmiotowy postrzega słownik jako teaurus, tworząc samemu niejawne hiperonimie.

3.4 Koncepcja rozwiązania

Jak wcześniej napisano tworzenie przez czytelnika niejawnych hiperonimii jest czasochłonne. Dlatego bardzo dużym ułatwieniem byłoby automatyczne budowanie niejawnych hiperonimii przez komputer i tym samym automatyczne tworzenie teaurusu ze słownictwa kontrolowanego.

Niejawne hiperonimie wynikają z reguł języka haseł przedmiotowych. Większość z tych reguł została ustalona w ten sposób, aby ułatwić wyszukiwanie haseł w indeksie alfabetycznym poprzez grupowanie w nim haseł o zbliżonym znaczeniu. W następnym rozdziale zostanie omówiony język haseł przedmiotowych oraz jego reguły pozwalające na zbudowanie niejawnych hiperonimii.

Po zbudowaniu teaurusu takiego jak ten przedstawiony na rysunku 3.1, mógłby on być prezentowany czytelnikowi i wykorzystany w szeregu zastosowań.

W wyszukiwarce książek teaurus można by zastosować w następujący sposób. Na początku przy pomocy teaurusu znajdowałyby się wszystkie terminy węższe (rekurencyjnie) hasła którym interesuje się czytelnik biblioteki. Następnie dla oryginalnego hasła oraz wszystkich znalezionych jego terminów węższych znajdowałyby się listę książek opisanych nimi. Listy te scalałoby się w jedną listę, usuwając książki duplikujące się. Scaloną listę prezentowano by czytelnikowi. Dodatkowo książki na tej liście można by posortować malejąco według częstości ich wypożyczeń w przeszłości. Dzięki temu uzyskalibyśmy automatycznie listę interesujących czytelnika książek, na początku której znalazłyby się najpopularniejsze z nich, czyli prawdopodobnie także najciekawsze dla czytelnika.

3.5 Inne możliwe zastosowania teaurusu

Oprócz wykorzystania teaurusu w wyszukiwarce książek można go także wykorzystać w dwóch innych zastosowaniach. Po pierwsze teaurus może być wykorzystany do hierarchicznego przeglądania słownika KABA i tym samym semantycznego wyboru interesującego nas hasła w przeciwieństwie do obecnie stosowanego wyboru alfabetycznego przy pomocy indeksu. Drugim zastosowaniem teaurusu mogłoby być przeprowadzanie analiz statystycznych tematów książek przechowywanych w

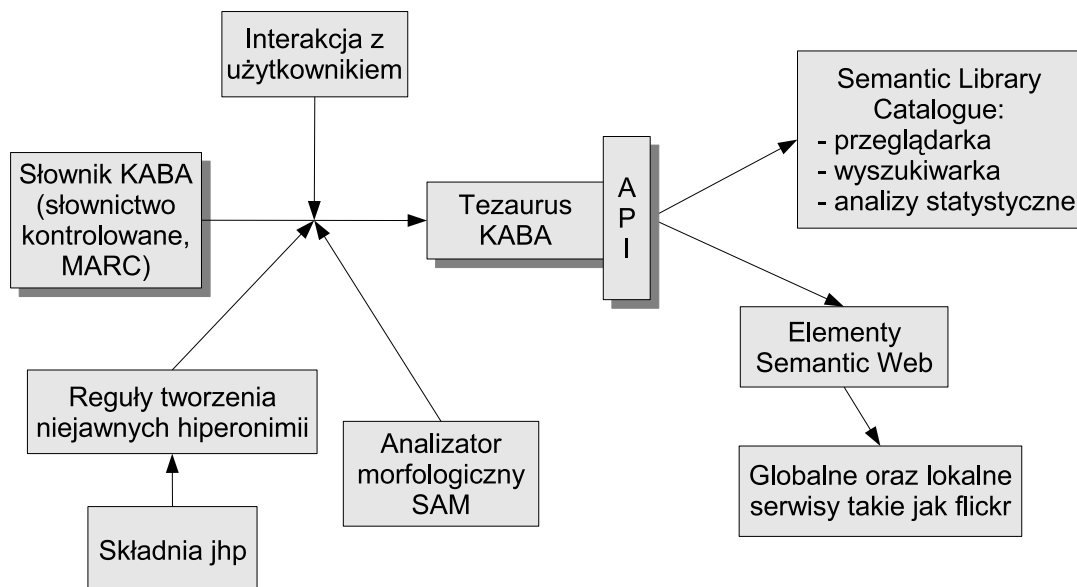
Rysunek 3.1: Fragment słownika KABA z około 50 hasłami tematycznie związanymi z pojęciem „Policja”. Pokazano także wszystkie terminy szersze hasła „Policja”. Słownik został przedstawiony jako tezaurus.

Terminy odrzucone podano kursywą pod terminami przyjętymi. Pogrubiono hasła będące głównymi pojęciami. Po nazwach haseł będących określnikami rzeczowymi dodano tekst „(general subdiv)”, pozostałe hasła są nazwami pospolitymi. Wielokropki oznaczają, że istniało więcej haseł ale ich podawanie nie było istotne dla tego przykładu.

Hiperonimie podane jawnie w rekordach haseł oznaczone są pogrubionymi liniami. Podano także przykładową relację skojarzeniową między hasłem „Policjanci” i „Detektywi”. Relacje skojarzeniowe nie są obecnie wykorzystywane.

Należy zauważyć, że relacje oznaczone cienkimi liniami i literą także mogą być uważane za hiperonimie. Hiperonimie te nie są jednak podane w rekordach jako terminy szersze lub węższe haseł, gdyż wynikają z języka haseł przedmiotowych. Literą oznaczono regułę języka haseł przedmiotowych z której relacja wynika: l – na podstawie leksemów (czyli tematów lub określników), a – na podstawie dopowiedzeń, d – na podstawie nazw zależnościowych, b – na podstawie początków nazw haseł, vx – na podstawie wariantu (na przykład vl na podstawie leksemów wariantu). Reguły te zostaną omówione i wykorzystane w dalszych rozdziałach. Dotychczas wspomniano tylko o jednej regule, która oznaczona jest tutaj literą 'l'.

Usunięto prawie wszystkie relacje wynikające z przechodności innych, choć może to czasem powodować pewną nieczytelność tezaury. Pozostawiono jedynie relację wynikającą z dopowiedzenia, aby zilustrować to źródło relacji.



Rysunek 3.2: Schemat tworzonego systemu komputerowego. Utworzenie systemu będzie polegało na konwersji słownika KABA do tezaurusa, a następnie wykorzystaniu tego ostatniego w szeregu zastosowań.

bibliotece. Na przykład moglibyśmy analizować na jakie tematy jest najwięcej książek w bibliotece lub na jakie tematy wypożycza się najwięcej książek. Powyższe zastosowania tezaurusa zostaną szerzej omówione w następnym rozdziale.

Tezaurus mógłby być zastosowany nie tylko do organizacji treści w bibliotekach, ale także innych zbiorów wiedzy na przykład całego Internetu jak i jego części na przykład w serwisie flickr.

Na rysunku 3.2 przedstawiono koncepcję rozwiązania niedoskonałości obecnych katalogów przedmiotowych oraz zastosowania projektowanego katalogu.

Rozdział 4

Projekt

W tym rozdziale zostanie przeprowadzona analiza możliwości utworzenia niejawnych hiperonimii, przewidywanych problemów z tym związanych, a także zostaną wybrane najlepsze z możliwych rozwiązań. Aby tego dokonać na początku potrzebna będzie szczegółowa analiza języka haseł przedmiotowych KABA. Należy przeanalizować zarówno reguły języka jawnie podane w literaturze, jak i te nie opisane do tej pory a wynikające jedynie z praktyki Centrum NUKAT. Te ostatnie reguły niestety musiały być odkrywane indukcyjnie podczas analizy statystycznej słownika KABA.

Na końcu zostaną przedstawione projekty zastosowań tezaury KABA.

4.1 Język haseł przedmiotowych KABA

Język haseł przedmiotowych (jhp) określa zasady budowy tekstu terminu przyjętego oraz terminów odrzuconych hasła, czyli tak zwanych nagłówków hasła [3].

Nagłówek hasła składa się z tematu hasła i opcjonalnie z określników wyszczególniających znaczenie hasła. Określniki mogą być rzeczowe (nazwy pospolite), geograficzne, chronologiczne oraz formy. Zarówno tematy jak i określniki mogą posiadać dopowiedzenia objaśniające dziedzinę pojęcia, gdy nie jest ona oczywista. Dopowiedzenie może być nazwą pospolitą lub geograficzną. Temat może być złożony z kilku podtematów. Każdy z tematów (podtematów) oraz określników nazywany jest leksemem.

Poniżej podano przykład hasła z dwoma określnikami (geograficznym i chronologicznym) oraz tematem zawierającym dopowiedzenie pospolite:

Dowód (prawo karne) – Polska – 19 w.

Hasło to dotyczy tematyki dowodów używanych w śledztwach kryminalnych w Polsce w XIX wieku. Innym przykładem jest hasło z jednym określnikiem rzeczowym:

Policjanci – kształcenie

dotyczące tematyki kształcenia policjantów. Użyte w hasłach tematy i określniki powinny być terminami przyjętymi, natomiast dopowiedzenia mogą być także ter-

minami odrzuconymi.

Określniki. W haśle może występować kilka określników rzeczowych i formy, jednak nie więcej niż jeden określnik geograficzny i chronologiczny. Wyjątkiem jest hasło: „Ukraińcy – Polska – przesiedlenie – Związek Radziecki”, w którym występują dwa określniki geograficzne określające odpowiednio miejsce początkowe i docelowe. Określniki powinny być podane w następującym porządku: określniki rzeczowe, określnik geograficzny, określnik chronologiczny, określniki formy.

Kompaktowanie dwustopniowych określników geograficznych. Na potrzeby pracy uproszczono język KABA, który oryginalnie zapisywał nazwy geograficzne w postaci dwóch określników geograficznych. Taki hierarchiczny określnik, w którym naprzód występowała nazwa kraju, a potem nazwa obiektu geograficznego był pomocny w alfabetycznym grupowaniu haseł. W przeglądarce semantycznej funkcja grupowania alfabetycznego powinna być zastąpiona grupowaniem semantycznym. Dzięki temu można uprościć język i zamiast pisać „Kościoły – Polska – Poznań” można napisać „Kościoły – Poznań (Polska)”. Jest to o tyle wygodne, że w słowniku istnieje hasło „Poznań (Polska)”, a nie „Polska – Poznań”.

Hasła rozwinięte, określniki związane. Nagłówek nazywany jest prostym, jeśli składa się tylko z tematu lub określników; natomiast jeśli składa się z tematu i określników, to nazywany jest rozwiniętym.

Określniki swobodne definiowane są w osobnym rekordzie i mogą być łączone z tematami wielu lub jednej kategorii semantycznej. Natomiast określniki związane mogą być połączone jedynie z tematem hasła rozwiniętego w którym zostały zdefiniowane. Określniki związane nie są definiowane we własnym rekordzie.

Dopowiedzenia. Dopowiedzeniem jest część tekstu leksemu nagłówka umieszczona w nawiasach okrągłych. Nagłówek może mieć jeden lub wiele leksemów, a każdy z leksemów może mieć jedno lub kilka dopowiedzeń lub nie mieć ich wcale. Dopowiedzenia mogą być kwalifikujące, lokalizujące geograficzne lub lokalizujące czasowe. W jednym leksemie może być kilka dopowiedzeń lokalizujących geograficznych, jeśli leksem dotyczy więcej niż jednej lokalizacji geograficznej. Dopowiedzenie lokalizujące geograficzne najczęściej jest hierarchiczne, to znaczy składa się z dwóch lub trzech nazw geograficznych będących ze sobą w relacji hierarchicznej.

Poniżej dla przykładu podano hipotetyczne jednoleksemowe hasło korporatywne zawierające jedno dopowiedzenie lokalizujące czasowe, dwa hierarchiczne dopowiedzenia lokalizujące geograficzne oraz jedno dopowiedzenie kwalifikujące:

Word Peace Congress (1949 ; Paryż, Francja / Praga, Czechosłowacja ; kongres)

Dopowiedzenia mogą mieć rolę rozróżniającą znaczenia haseł wieloznacznych lub też objaśniającą znaczenia wyrażen jednoznacznych, jednak mogących budzić wątpliwości. Na przykład w hasle „Pancerniki (ssaki)” dopowiedzenie ma rolę rozróżniającą, natomiast w hasle „Siuksowie (Indianie)” ma rolę objaśniającą.

Nazwy zależnościowe. Sam temat może czasem przedstawiać związek między dwoma innymi tematami. Na przykład hasło „Policja i prasa” oznacza relacje między policją i prasą; hasło „Telewizja i dzieci” stosuje się, gdy ma się na myśli wzajemny wpływ telewizji i dzieci; a hasło „Literatura i informatyka” stosuje się do tematu wpływu informatyki na twórczość literacką. Hasła zależnościowe składają się z dwóch haseł (np. „Policja” i „Prasa”) połączonych spójnikiem „i”.

Nazwy wielowyrazowe. Tematy oraz określniki haseł mogą być kilkuwyrazowe. Często mają one postać rzeczownika z następującą po nim przydawką będącą przymiotnikiem, rzeczownikiem lub wyrażeniem przyimkowym, na przykład: „Pisarze polscy”, „Żandarmeria wojskowa”, „Nauki społeczne”. Ważne jest, że przeważnie pierwszym wyrazem jest wyraz określany, a następnymi wyrazami są wyrazy określające. Kolejność ta (naturalna dla języka polskiego dla utartych wyrażen) jest bardzo ważna w katalogu alfabetycznym i tym samym przy określaniu niejawnych hiperonimii.

Hasła o podwójnej funkcji. Jak napisano w poprzednim rozdziale, hasła oraz nagłówki ich terminów przyjętych posiadają typ. Niektóre z haseł będących tematem lub hasłem rozwiniętym mogą być używane w innych hasłach rozwiniętych także jako określniki. Mówi się o nich, że są przeznaczone do pełnienia podwójnej funkcji. W praktyce dotyczy to jedynie nazw geograficznych, których większość może pełnić funkcję określnika geograficznego. Co więcej w słowniku tylko jeden określnik geograficzny jest samodzielnie zdefiniowany, a więc wszystkie określniki geograficzne w hasłach rozwiniętych z wyjątkiem jednego są tak naprawdę formami tematowej nazwy geograficznej.

Język nagłówka. Terminy przyjęte haseł przeważnie podawane są w języku polskim. Jednak nazwy obiektów położonych na terenie krajów o tak zwanych językach kongresowych podawane są w tych ojczystych językach kongresowych. W takim przypadku nazwa w języku polskim podana będzie jako termin odrzucony. W terminach odrzuconych niejednokrotnie podawane są także nazwy w innych językach niż polski (na przykład w łacinie, języku angielskim lub francuskim). W takich przypadkach w oryginalnym słowniku KABA po nagłówku obcojęzycznym podawany jest w nawiasach kwadratowych kod języka. W projektowanym systemie oznaczenie języka będzie usuwane z tekstu nagłówka i umieszczane w specjalnym polu.

4.2 Budowanie niejawnych hiperonimii

Jak już wcześniej napisano, hiperonimie nie wynikające z języka haseł przedmiotowych są podawane jawnie w rekordach haseł. W celu przekształcenia słownika KABA do tezaurusa należy dodatkowo utworzyć hiperonimie wynikające z jhp. Analizując poszczególne elementy jhp możemy dojść do wniosku, że wynikają z niego następujące hiperonimie:

- a) Hasło złożone z tematu i określników jest hiponimem hasła złożonego z dowolnego jego podciągu, na przykład tylko z tematu, tylko z określnika lub z tematu i jednego z określników. Na przykład hasło „Dowód (prawo karne) – Polska” ma dwa hiperonimy: „Dowód (prawo karne)” oraz „Polska”, natomiast hasło „Żydzi – Bawaria (Niemcy) – historia” ma dwa bezpośrednie hiperonimy: „Żydzi – historia” i „Bawaria (Niemcy) – historia” oraz trzy wynikające z przechodniości: „Żydzi”, „Bawaria (Niemcy)” i „– historia”. „Żydzi – Bawaria (Niemcy)” nie jest hiperonimem, gdyż nie ma takiego hasła w słowniku.
- b) Hasło proste (temat bez określników) zawierające dopowiedzenia jest hiponimem hasła określającego samo dopowiedzenie. Na przykład hasło „Dowód (prawo karne)” ma hiperonim „Prawo karne”.
- c) Hasło zależnościowe jest hiponimem swoich haseł składowych. Na przykład hasło „Policja i prasa” ma dwa hiperonimy: „Policja” oraz „Prasa”.
- d) Hasła proste (bez określników) zaczynające się innym hasłem są *prawie zawsze* jego hiponimami. Na przykład hasło „Pisarze” posiada hiponimy: „Pisarze polscy”, „Pisarze angielscy” i tak dalej. Natomiast hasło „Aerodynamika” posiada dwa hiponimy: „Aerodynamika przepływów przydźwiękowych”, „Aerodynamika przepływów naddźwiękowych”, a także przechodnio: „Aerodynamika przepływów hipersonicznych”.

Łatwo zauważyć, że trzy pierwsze reguły wynikają z rozkładania nagłówka terminu przyjętego na poszczególne elementy języka haseł przedmiotowych. Czwarta reguła wynika z rozkładu leksemu hasła na wyrazy języka naturalnego, a nie elementy jhp. Dlatego też czwarta reguła jest heurystyczna, ale ze względu na sformalizowanie jhp prawie zawsze poprawna. Przedstawione reguły to cztery podstawowe reguły tworzenia niejawnych hierarchii. Wynik ich działania został zilustrowany na rysunku 3.1.

Dodatkowo reguły te można by stosować nie tylko do terminów przyjętych, ale także do terminów odrzuconych haseł. Przykładowo dzięki temu, że hasło „Postępowanie karne” posiada termin odrzucony „Prawo karne – postępowanie” wiemy, że jego terminem szerszym jest hasło „Prawo karne”. Nie utworzylibyśmy tej hiperonimii, gdybyśmy analizowali jedynie terminy przyjęte haseł.

W rzeczywistości między hasłami istnieje też pewna ilość hiperonimii, które ani nie są jawnie podane ani nie wynikają z jhp. Dzieje się tak dlatego, że informacja o

nich podana jest w odsyłaczach orientacyjnych uzupełniających hasła.

W dalszej części punktu zostanie przedstawiona analiza możliwości utworzenia hierarchii niejawnych na podstawie wyżej wymienionych czterech reguł, a także zostaną przedstawione przewidywane problemy z tym związane oraz zostaną wybrane najlepsze z możliwych rozwiązań. Zostaną także omówione hiperonimie zapisane w odsyłaczach orientacyjnych uzupełniających.

4.2.1 Wiązanie haseł z ich leksemami

Najistotniejszym elementem reguły jest wiązanie nagłówków rozwiniętych z ich tematami. Przydatnym może się także okazać wiązanie nagłówków rozwiniętych z określnikami w nich występującymi. Jeśli nagłówek rozwinięty składa się z kilku określników, to nie tylko powinno się próbować wiązać go z poszczególnymi pojedynczymi leksemami, ale także próbować wiązać z hasłami kilkuleksemowymi będącymi podzbiorem leksemów oryginalnego hasła. Przykładowo hasło „Żydzi – Bawaria (Niemcy) – historia” ma także następujący hiperonim: „Bawaria (Niemcy) – historia”. Ogólną ideą reguły jest wiązanie hasła z hiperonimami powstałymi poprzez usunięcie z hasła dowolnego pojedynczego leksemu.

Jak już wcześniej wspomniano niektóre hasła korporatywne, wydarzeń oraz będące tytułami mogą mieć temat złożony z kilku leksemów. W takim przypadku należy poszukiwać hiperonimów powstałych poprzez usunięcie ostatniego elementu tematu. Przykładowo hiperonimem tytułu „Sienkiewicz, Henryk. Potop” jest hasło osobowe „Sienkiewicz, Henryk”, a hasła korporatywnego „Polska. Polskie Siły Powietrzne. 316 Dywizjon Myśliwski Warszawski” hasło korporatywne „Polska. Polskie Siły Powietrzne” oraz przechodnio hasło geograficzne „Polska”.

Wszystkie pojedyncze leksemy powinny być terminami przyjętymi zapisanymi w słowniku KABA; natomiast hasła kilkuleksemowe, będące podciągami leksemów oryginalnego hasła, niekoniecznie muszą być obecne w słowniku KABA. Dlatego też jeśli jeden z podciągów oryginalnego hasła nie jest hasłem słownikowym, to powinniśmy szukać hiperonimów rekurencyjnie pośród podciągów tego ostatniego hasła aż do chwili, gdy wszystkie leksemy zostaną rozpoznane.

Wyjątek stanowią określniki związane, które nie posiadają własnego hasła, a więc nie mogą być rozpoznane. Duża część określników związanych posiada o tym informację zapisaną w rekordzie hasła rozwiniętego. Oprócz tego przyjmuje się, że określnikami związanymi są określniki chronologiczne nie posiadające własnego rekordu.

4.2.2 Wiązanie haseł z ich dopowiedzeniami

W tej regule tworzenia relacji niejawnych wystarczy analizować hasła proste. Dla haseł rozwiniętych znajdziemy hiperonimy poprzez rozpoznanie ich leksemów i następnie będziemy mogli wiązać te leksemy z ich dopowiedzeniami.

Rozpoznawanie znaczenia dopowiedzeń może być kłopotliwe. Rozpatrzmy następujące hasło z jednym dopowiedzeniem kwalifikującym: „Bismarck (pancernik)”. Czy dopowiedzenie „pancernik” powinniśmy rozpoznać jako formę hasła „Pancerniki (okręty wojenne)”, czy też hasła „Pancerniki (ssaki)”? Oczywiście powinniśmy wybrać pierwsze hasło, jednak nie możemy tego wniosku wyciągnąć z zawartości słownika KABA, gdyż brakuje w nim tej informacji. Aby temu zaradzić, moglibyśmy posłużyć się zewnętrzną ontologią polskojęzyczną lub angielskojęzyczną, jednak podczas pisania pracy magisterskiej ontologia języka polskiego jeszcze nie była dostępna. Dlatego też dopowiedzenie „pancerniki” zostanie rozpoznane niejednoznacznie. W rezultacie w celu nie wprowadzania niejednoznacznych hiperonimii hasło to nie zostanie powiązane z żadnym hiperonimem.

Należy zwrócić uwagę także na inną trudność, która na szczęście może być rozwiązana. Otóż dopowiedzenia są często podane w innej liczbie gramatycznej niż oryginalne hasło. W celu ich rozpoznania często należy więc dokonać konwersji liczby gramatycznej na przeciwną. Konwersję można wykonać metodą relacyjną, stosując reguły deklinacji lub też metodą słownikową, korzystając z form zapisanych w słowniku języka polskiego lub podobnym. W tworzonym systemie wybrano drugą możliwość.

W hasłach można rozpoznawać nie tylko dopowiedzenia kwalifikujące, ale także lokalizujące geograficzne. Dzięki nim powiązemy hasła z hiperonimami będącymi nazwami geograficznymi. Rozpoznawanie dopowiedzeń lokalizujących chronologicznych nie ma większego sensu, między innymi z tego powodu, że w słowniku jest mało tematów oznaczających czas.

4.2.3 Wiązanie nazw zależnościowych

Ponieważ hasło zależnościowe dotyczy obu haseł składowych, to można przyjąć, że jest ono ich terminem węższym. Jeśli pewna książka jest o relacjach policji z prasą, to także dotyczy zarówno samej policji jak i samej prasy.

Nie wszystkie leksemy zawierające dwa człony powiązane spójnikiem „i” są nazwami zależnościowymi. Taką formę przyjmują także nazwy zbiorowe (sumaryczne). Nazwy zbiorowe stanowią sumę mnogościową znaczeń elementów składowych, podczas gdy nazwy zależnościowe są iloczynem mnogościowym znaczeń elementów składowych. Hasło zbiorowe jest tworzone, gdy nie jest opłacalne tworzenie dwóch osobnych haseł dla obu elementów z tego względu, że znaczenia obu elementów są

bardzo bliskie albo dlatego, że nie ma książek na poszczególne tematy składowe. Natomiast hasło zależnościowe tworzone jest dla haseł o odległych znaczeniach, jednak mimo to mających ze sobą element wspólny, na który właśnie wskazuje nazwa zależnościowa. Przykładem nazwy zbiorowej jest hasło „Emigracja i imigracja”, a nazwy zależnościowej hasło „Ojciec i dziecko” oraz hasło „Kościół i państwo”.

Nazwy zależnościowe są hiponimami swoich elementów, natomiast nazwy zbiorowe nie. Wynika z tego, że jedynie nazwy zależnościowe powinniśmy łączyć relacją hierarchiczną z ich elementami. Jednak w jaki sposób możemy odróżnić nazwę zależnościową od zbiorowej skoro mają one taką samą budowę?

W słowniku KABA istnieje reguła dodawania dla nazw zależnościowych terminu odrzuconego takiego samego jak termin przyjęty, jednak ze składowymi zamienionymi miejscami. Natomiast dla nazw zbiorowych dodaje się termin odrzucony równy drugiemu członowi nazwy zbiorowej. Tak więc hasło „Kościół i państwo” posiada termin odrzucony „Państwo i Kościół”, a hasło „Emigracja i imigracja” posiada termin odrzucony „Imigracja”. Reguły te można byłoby zastosować do rozróżniania nazw zależnościowych od zbiorowych gdyby nie to, że nie zawsze są one stosowane.

Jednak istnieje inny prosty sposób odróżnienia nazw zależnościowych od zbiorowych. Obie składowe nazwy zależnościowej powinny istnieć w słowniku, podczas gdy dla nazwy zbiorowej nie powinny one być samodzielnymi hasłami słownika. Dlatego też dla obu nazw zależnościowych i zbiorowych rozpoznajemy ich składowe. Jeśli obie składowe zostały rozpoznane, to traktujemy hasło oryginalne jako nazwę zależnościową. Jeśli żadna składowa nie została rozpoznana, to traktujemy hasło jako nazwę zbiorową. A jeśli została rozpoznana tylko jedna składowa, to generujemy ostrzeżenie o tym, że hasło jest podejrzewane o niepoprawną budowę.

W składowych nazw zależnościowych podobnie jak w dopowiedzeniach używa się liczby gramatycznej, która najlepiej odzwierciedla związek między obiema składowymi. Liczba składowej może być różna od liczby oryginalnego hasła reprezentowanego przez składową. Na przykład w hasle „Ojciec i dziecko” występują składowe w liczbie pojedynczej, natomiast oryginalne hasła występują w liczbie mnogiej: „Ojcowie”, „Dzieci”. Nie sprawdzono jednak czy zmiana liczby gramatycznej występuje często. Ponieważ szukanie haseł we wszystkich liczbach może powodować większą niejednoznaczność rozpoznania, to nie wiadomo czy jest to opłacalne.

Składowe nazwy zależnościowej muszą być terminami przyjętymi, to znaczy nie stosuje się terminów odrzuconych.

Nazwy zależnościowe występują prawie zawsze tylko w tematach nazw pospolitych. Z powodu założenia o przechodniości relacji termin szerszy – termin węższy wystarczy rozpatrywać jedynie nierozwinięte nazwy pospolite. Rozwinięte nazwy pospolite zawierające w swoim temacie nazwę zależnościową i tak zostaną powiązane z ich tematem, a temat ten zostanie ostatecznie powiązany z jego składowymi.

4.2.4 Wiązanie hasła z hasłami zaczynającymi się nim

Jak już wspomniano podczas omawiania języka haseł przedmiotowych, tematy oraz określniki haseł mogą być kilkuwyrazowe. Przeważnie pierwszym wyrazem jest wyraz określany, a następnymi wyrazami są wyrazy określające. Często mają one postać rzeczownika z następującą po nim przydawką będącą przymiotnikiem, rzeczownikiem lub wyrażeniem przyimkowym. Dlatego też takie kilkuwyrazowe leksemy przeważnie są hiponimami rzeczownika którym się zaczynają pod warunkiem, że rzeczownik ten jest samodzielnym hasłem KABA.

Wiązanie hasła kilkuwyrazowego z innymi wyrazami niż pierwsze o wiele częściej prowadziłoby do powstania błędnych niż poprawnych hiperonimii. Tekst haseł dobierany jest w ten sposób, aby hasła o zbliżonych znaczeniach były blisko siebie w indeksie alfabetycznym. Oznacza to, że język haseł przedmiotowych wpływa na sformalizowanie znaczenia jedynie pierwszych, a nie wszystkich wyrazów w hasle. Dlatego też sens ma jedynie rozpoznawanie pierwszych wyrazów wielowyrazowego leksemu.

Leksemy kilkuwyrazowe nie zawsze mają postać rzeczownika z przydawką. W takim przypadku prawdopodobnie pierwszy wyraz nie zostanie rozpoznany. Na przykład w hasle „Krażenie krwi” wyraz „Krażenie” nie stanowi samodzielnego hasła słownika KABA. Podobnie samodzielnym hasłem nie jest pierwszy wyraz następujących haseł: „Zaburzenia mowy”, „Szeregi Fouriera”.

Wiele leksemów kilkuwyrazowych jest tak zwanymi nazwami inwersyjnymi, w których szyk naturalny został przestawiony w celu łatwiejszego wyszukiwania hasła w indeksie. Na przykład zamiast utworzyć hasło „Zbrodnia katyńska” utworzono inwersyjne hasło „Katyń, Zbrodnia”. Wydaje się, że hasła takie można by także wiązać z pierwszym wyrazem.

4.2.5 Relacje równoważności

Jak już wcześniej wspomniano, wiele hiperonimii można by otrzymać analizując terminy odrzucone haseł. Terminy odrzucone swą zawartością często przenoszą podobną informację co terminy przyjęte, jednak wyrażając ją w odmiennej postaci. Postać ta może być łatwiejsza do analizowania. Równie często termin odrzucony może przenosić dodatkową informację.

Na przykład hasło „Krażenie krwi” posiada termin odrzucony „Krew – krążenie”. Analizując termin przyjęty trudno by było powiązać go z hasłem „Krew”. Natomiast powiązanie takie jest oczywiste dzięki temu, że termin odrzucony jest podzielony na temat i określnik. Innym przykładem jest hasło „Fizyka”. Dzięki temu że posiada ono termin odrzucony „Nauki fizyczne”, jesteśmy w stanie powiązać go z hasłem „Nauka”.

4.2.6 Odsyłacze orientacyjne uzupełniające

W polu 360 rekordu MARC hasła wzorcowego przechowywane są teksty nazywane odsyłaczami orientacyjnymi uzupełniającymi. Pojedyncze hasło może mieć jeden odsyłacz, kilka odsyłaczy lub nie mieć ich wcale. Odsyłacze informują czytelnika o tym, że aktualne hasło ma hiponimy o cechach podanych w opisie tekstowym odsyłacza.

Analizy statystyczne słownika wykazały, że odsyłacze mogą mieć jeden z pięciu typów:

1. odsyłacz wskazujący na hiponimy zawierające pewien podany określnik,
2. odsyłacz wskazujący na hiponimy zawierające pewne podane połączenie kilku określników,
3. odsyłacz wskazujący na hiponimy będące hasłami zaczynającymi się pewnym podanym wyrazem lub wyrażeniem,
4. odsyłacz wskazujący na hiponimy będące hasłami zawierającymi pewien podany wyraz lub wyrażenie,
5. odsyłacz wskazujący na hiponimy odwołując się przy tym do wiedzy semantycznej czytelnika, na przykład wskazując na „poszczególne nazwy hiponimów hasła” nie podając algorytmicznej metody ich odnalezienia.

Oto przykłady określników poszczególnych typów:

- a) w hasle „Literatura” znajduje się określnik typu pierwszego informujący o tym, że hasła zawierające określnik „- literatury” po nazwach „kontynentów lub regionów” są hiponimami aktualnego hasła,
- b) w hasle „Patologia molekularna” znajduje się określnik typu drugiego informujący o tym, że hasła zawierające połączenie określników „- choroby – aspekt molekularny” po nazwach „organizmów żywych, żyjących i wymarłych oraz po nazwach narządów, układów wewnętrznych i części ciała” są hiponimami aktualnego hasła,
- c) w hasle „Literatura” znajduje się określnik typu trzeciego informujący o tym, że hasła zaczynające się wyrazem „Literatura” takie jak „Literatura katolicka” są hiponimami aktualnego hasła,
- d) w hasle „Metale” znajduje się określnik typu czwartego informujący o tym, że hasła zawierające wyraz „metale” takie jak: „Alergia na metale” i „Metaloterapia” są hiponimami aktualnego hasła,
- e) w hasle „Metale” znajduje się określnik typu piątego informujący o tym, że „poszczególne metale i grupy metali” takie jak hasło „Miedź” są hiponimami aktualnego hasła.

Odsyłacz orientacyjny uzupełniający przeważnie zawiera informację o hiperoniach niejawnych, która nie wynika z języka haseł przedmiotowych. W pewnym sensie wyjątek stanowią odsyłacze typu trzeciego, gdyż hiponimy przez nie wskazywane wynikają z reguły łączenia hasła z pierwszym wyrazem.

Odsyłacze pierwszego i drugiego typu są do siebie podobne, więc zostaną omówione wspólnie. Analiza ich zawartości jest kłopotliwa z tego względu, że można by tworzyć na ich podstawie relacje do hiponimów tylko wtedy, gdybyśmy potrafili stwierdzić czy temat hasła będącego kandydatem do bycia hiponimem należy do klasy słownie opisanej w odsyłaczu. Ponieważ opis klasy jest podany w języku naturalnym, to nie jesteśmy w stanie tego zrobić. Dlatego też informacja o hiponimach zapisana w odsyłaczach pierwszego i drugiego typu zostanie utracona.

Potwierdzeniem tego że należy analizować treść klasy opisanej w odsyłaczu jest następujący odsyłacz typu pierwszego w hasle „Remont”: „zobacz też hasła zawierające określnik «– konserwacja i restauracja» po nazwach obiektów architektury i zabytkowych”. Będziemy otrzymywać poprawne relacje do hiponimów dopóki ich tematy będą należały do klasy budowli. Jednak jeśli zaczniemy wiązać hasło „Remont” z dowolnymi hasłami zawierającymi określnik „– konserwacja i restauracja”, to powiążemy go także z hasłem „Książki – konserwacja i restauracja”. Jednak oczywistym jest, że książek nie remontuje się, więc hasło to nie powinno być hiponimem hasła „remont”.

Zostanie także utracona informacja zawarta w odsyłaczach piątego typu, gdyż nie posiadamy wiedzy semantycznej wymaganej do utworzenia hiperonimii na podstawie takich odsyłaczy. Moglibyśmy tworzyć hiperonimie na podstawie odsyłaczy tego typu tylko wtedy, gdybyśmy wykorzystali zewnętrzną ontologię na przykład dla języka polskiego. Innym rozwiązaniem byłoby ręczne dodawanie przez NUKAT jawnych hiperonimii do haseł wskazywanych przez odsyłacze typu piątego. Pewnym pocieszeniem jest fakt, że część hiperonimii wynikających z odsyłaczy piątego typu zostanie dodanych w wyniku rozpoznania dopowiedzeń hiponimów.

Jak wspomniano przed chwilą informacja zawarta w odsyłaczach trzeciego typu raczej nie jest przydatna podczas tworzenia hiperonimii niejawnych.

Implementacja tworzenia hierarchii niejawnych na podstawie odsyłaczy czwartego typu raczej nie sprawiłaby dużych problemów, jednak nie warto ich analizować ze względu na to, że są one rzadko używane w słowniku KABA.

Podsumowując odsyłacze orientacyjne uzupełniające w obecnym stanie rozwoju systemu nie będą wykorzystywane.

4.2.7 Inne reguły

W tym punkcie zostaną wspomniane niektóre z pozostałych reguł, które można by zastosować do utworzenia hiperonimii.

Określniki chronologiczne

Określniki chronologiczne można by wiązać w hierarchie wynikające z zawierania się ich zakresów. Na przykład określnik „– 1800-1900” powinien być hiperonimem określnika „– 1810-1840”.

We wcześniejszej wersji języka haseł przedmiotowych KABA określniki chronologiczne były poprzedzane w nagłówku rozwiniętym określnikiem „– historia” dla uwydatnienia faktu, że hiperonimem określnika chronologicznego jest określnik „– historia”. Obecnie można by było tą informację wywnioskować wiążąc hasła zawierające określnik chronologiczny z hasłem które posiada taką samą formę z tym wyjątkiem, że określnik chronologiczny zastąpiony jest określnikiem „– historia”.

4.3 Zastosowania hierarchii tezauryusa

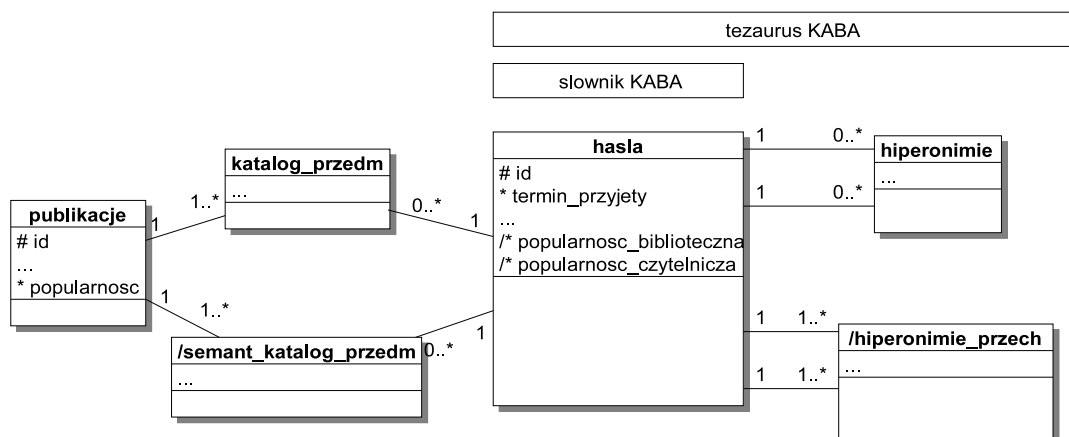
W obecnym punkcie zostaną przedstawione projekty zastosowań wspomnianych w rozdziale trzecim. Na początku jednak zostanie przedstawiony projekt schematu bazy danych umożliwiającej efektywne zaimplementowanie tych zastosowań.

4.3.1 Schemat bazy danych

Gdy już będzie dostępny gotowy tezaurus, trzeba będzie stworzyć dla niego strukturę w bazie danych umożliwiającą efektywnie jego wykorzystanie w semantycznym katalogu przedmiotowym. Potrzeba stworzenia takiej struktury jest oczywista, jeśli uświadomimy sobie, że będziemy musieli dla każdego hasła wykonywać rekurencyjnie operację znajdowania wszystkich hiponimów, to znaczy także tych otrzymywanych dzięki przechodniości relacji hiperonimii. Dzięki omawianej strukturze wystarczy, że wykonamy raz taką operację. Baza danych katalogu przedmiotowego zawierającego gotowy tezaurus KABA przedstawiona jest na rysunku 4.1.

Na tezaurus KABA składa się tabela **hasła** oraz tabela **hiperonimie** łącząca rekordy tej pierwszej ze sobą. W tabeli **publikacje** przechowywane są książki opisane hasłami słownika KABA dzięki tabeli **katalog_przedm**. Tabele oznaczone znakiem '/' umożliwiają efektywne wykonywanie operacji semantycznych i nie przechowują żadnych informacji, które nie wynikałyby z pozostałych tabel. Więcej informacji odnośnie implementacji bazy danych znajduje się w opisie rysunku 4.1.

Semantyczne informacje z bazy danych można odczytywać w następujący sposób. Publikacje opisane pewnym hasłem lub jego hiponimami uzyskujemy przechodząc od krotki w tabeli **hasła** przez tabelę **semant_katalog_przedm** do tabeli **publikacje**. Najpopularniejsze hasła uzyskujemy zwracając określoną liczbę krotek z tabeli **hasła** o największej wartości pola **popularnosc_biblioteczna** albo o największej wartości pola **popularnosc_czytelnicza**. Jeśli chcemy się ograniczyć



Rysunek 4.1: Schemat bazy danych semantycznego katalogu przedmiotowego zawierającego gotowy tezaurus KABA. Część tezaurusowa bazy danych nie służy do tworzenia tezaurusu ze słownika KABA, a jedynie do jego efektywnego stosowania w semantycznym katalogu przedmiotowym.

Oryginalny tezaurus KABA, zawierający hasła powiązane hiperonimiami z możliwością przechodzenia do bezpośrednich hiponimów i hiperonimów, przechowywany jest w tabelach `hasla` oraz `hiperonimie`. Na polu `termin_przyjety` założony jest indeks. Pojedynczy rekord tabeli `hiperonimie` reprezentuje pojedynczą hiperonimie łączącą hiponim z hiperonimem.

Ze względów efektywnościowych z tabeli `hiperonimie` wyznaczana jest tabela `hiperonimie_przech` („przechodnie hiperonimie”), co oznaczane jest znakiem `'/'` przed jej nazwą. Rekordy tej tabeli łączą dowolne hasło z samym sobą oraz ze wszystkimi jego bezpośrednimi lub pośrednimi hiponimami (lub hiperonimami, jeśli tabela odczytywana jest w drugą stronę). Tak więc każde hasło powiązane jest przynajmniej ze sobą. Tabela umożliwia odpowiedź na pytania czym jest hasło (jest sobą i hiperonimami) oraz co jest hasłem (hasłem jest ono samo oraz jego hiponimy).

W tabeli `publikacje` przechowywane są informacje o publikacjach w bibliotece. Dla każdej publikacji przechowywana jest jej popularność wśród czytelników. Tabela `katalog_przedm` („katalog przedmiotowy”) opisuje każdą publikację jednym lub kilkoma hasłami KABA. Umożliwia ona także uzyskanie listy publikacji opisanych pewnym hasłem.

Podobnie jak wcześniej ze względów efektywnościowych z tabel `katalog_przedm` i `hiperonimie_przech` wyznaczana jest tabela `semant_katalog_przedm` w ten sposób, że dzięki niej dla dowolnego hasła można uzyskać publikacje opisane tym hasłem lub dowolnym jego hiponimem.

Na końcu z tabeli `semant_katalog_przedm` (i `publikacje`) wyznaczane są pola `popularnosc_biblioteczna` i `popularnosc_czytelnicza`. Pierwsze jest liczbą książek opisanych danym słowem kluczowym lub jego hiponimem (przechodnio). Drugie jest summaryczną liczbą wypożyczeń wyżej wymienionych książek.

jedynie do najpopularniejszych haseł będących hiponimami pewnego hasła, to wykorzystujemy w tym celu tabelę `hiperonimie_przech` i postępujemy analogicznie jak poprzednio.

Hasła KABA można wybierać dzięki indeksowi pola `termin_przyjety` albo przeglądając hierarchię dzięki tabeli `hiperonimie`.

4.3.2 Przeglądarka tezauryasa

Określenie, zlokalizowanie oraz przejrzenie zawartości wszystkich hiponimów aktualnego hasła zajmuje dużo czasu, co pokazano już w rozdziale drugim. Wydaje się, że bardzo dużym ułatwieniem byłoby wyświetlanie dla aktualnie wybranego hasła wszystkich bezpośrednich hiponimów (jawnych i niejawnych) w postaci hiperłączy. Jeśli pewne hasło miałyby dużo bezpośrednich hiponimów, to można by je grupować semantycznie, na przykład według typu hasła lub dziedziny hasła (jeśli hasła takie miałyby także drugi, inny hiperonim).

Oprócz relacji hierarchicznych powinny być także wyświetlane w postaci hiperłączy relacje skojarzeniowe.

4.3.3 Wyszukiwarka książek

Zastosowanie tezauryasa w wyszukiwarce książek zostało już przedstawione w rozdziale trzecim. W tym podpunkcie przypomnimy je jeszcze raz, opisując bardziej szczegółowo projekt działania wyszukiwarki, także z wykorzystaniem bazy danych zaprezentowanej w punkcie 4.3.1.

Na początku, w celu przyspieszenia działania wyszukiwarki, dla każdego słowa kluczowego określamy i zapisujemy w bazie danych powiązania do wszystkich jego hiperonimów – nie tylko do tych bezpośrednich, ale także do wszystkich które można osiągnąć korzystając z przechodniości hiperonimii. Dzięki temu będziemy mogli szybko dowiedzieć się, że książka opisana hasłem „Korupcja w policji” dotyczy także następujących tematów: „Korupcja”, „Policjanci – deontologia”, „– deontologia (general subdiv)”, „Policjanci”, „Policja – uprawnienia”, „Policja” ... „Prawo”, „Nauki społeczne”. Powiązanie takiej książki z tematami najbardziej ogólnymi jest słabsze, choć nadal istnieje. Strukturę taką w bazie danych można oczywiście odczytywać także w drugą stronę, to znaczy odczytać wszystkie (bezpośrednie i pośrednie) hiponimy. W zaprezentowanej bazie danych powiązania te przechowywane są w tabeli `hiperonimie_przech`.

Jeśli teraz czytelnik chciałby uzyskać listę wszystkich książek na dany temat, to powinien wybrać słowo kluczowe go reprezentujące. Może tego dokonać przy pomocy przeglądarki hierarchii, indeksu alfabetycznego lub korzystając z obu metod. Następnie tworzy się listę haseł złożoną z wybranego przez czytelnika hasła oraz

wszystkich jego hiponimów (bezpośrednich i pośrednich). Dla każdego takiego hasła znajduje się listę wszystkich książek nim opisanych, a następnie listy te scala się, uważając aby na wynikowej liście książki nie powtarzały się. Całą listę sortuje się malejąco według częstości wypożyczeń książek. Należy zwrócić uwagę, że na początku listy znajdują się prawdopodobnie książki na najbardziej ogólne tematy, gdyż są one prawdopodobnie najczęściej wypożyczane. Korzystając z zaprezentowanej bazy danych całą operację można wykonać jeszcze prościej, wykorzystując proste zapytanie do tylko dwóch tabel, mianowicie do tabeli `semant_katalog_przedm` i tabeli `publikacje` (do tej ostatniej tylko w celu posortowania względem popularności).

Książki na liście można także pogrupować, na przykład według bezpośrednich hiponimów wybranego przez czytelnika hasła. Grupowanie mogłoby zacieemnić informację o tym, która książka jest najciekawsza. Dlatego opcjonalnie można byłoby zachować oryginalną kolejność książek, oznaczając przynależność książki do grupy w sposób graficzny, na przykład za pomocą ikony lub koloru.

Przedstawioną listę książek można utożsamić z bibliografią danego tematu. Należy zwrócić uwagę na to, że informacja o popularności książek, a więc prawdopodobnie także o jej przydatności dla czytelnika, jest obecnie nieosiągalna w żaden sposób.

Jeśli czytelnik chciałby wyszukać książki będące jednocześnie na kilka interesujących go tematów i nie znalazłby odpowiedniego słowa kluczowego (hasła rozwiniętego takiego jak „Policjanci – kształcenie”, nazwy zależnościowej takiej jak „Policja i prasa” lub hasła z przydawkami takiego jak „Pisarze polscy”), to może wybrać kilka słów kluczowych. Komputer dla każdego słowa kluczowego wykona opisaną wyżej procedurę, a następnie zwróci iloczyn mnogościowy zbiorów książek. Na przykład można spróbować wyszukać książki o żabach żyjących w Wielkopolsce, wybierając dwa słowa kluczowe: „Żaby” i „Wielkopolska (Polska ; region)”. Jeśli nie istnieje żadna książka na te dwa tematy, to można rozszerzyć zakres dowolnego ze słów kluczowych wybierając jego hiperonim. Na przykład można poszukać książek o wszystkich płazach w Wielkopolsce lub o żabach w całej Polsce. W przypadku pustego wyniku wyszukiwania, czynność rozszerzenia zakresów mogłaby być także wykonywana automatycznie przez komputer. Wyniki automatycznego rozszerzenia zakresów można by także prezentować, gdyby takie książki miały o wiele większą popularność od książek z oryginalnej listy.

Jak widać, zastosowanie tezauryusa w wyszukiwarkach rodzi wiele możliwości.

4.3.4 Badania statystyczne tematyki książek

Interesującym pytaniem jest, na jakie tematy jest najwięcej książek w bibliotece, to znaczy jakie tematy są najbardziej popularne? Dla każdego słowa kluczowego można określić liczbę książek o odpowiadającym mu temacie lub o temacie będą-

cym bezpośrednim lub pośrednim hiponimem (pole `popularnosc_biblioteczna`). Następnie można zwrócić listę tematów o największej liczbie książek. Tematy te mogą pochodzić z całego słownika lub można wyświetlić tylko tematy na najwyższym poziomie hierarchii.

Powyższą analizę można przeprowadzić nie tylko dla całej biblioteki, ale także dla wybranego tematu (słowa kluczowego). Na przykład można zadać pytanie na jakie podtematy tematu „Ssaki” jest najwięcej książek (jakie tematy są najbardziej popularne w bibliografii).

Zadanie można rozszerzyć na kilka słów kluczowych, określając na przykład jakie podtematy tematów „Polska” i „Rosja” są najbardziej popularne. W tym ostatnim przypadku podtematy zostaną wybrane spośród haseł będących hiponimami zarówno jednego jak i drugiego słowa kluczowego.

Innym rozszerzeniem byłyby taka modyfikacja powyższych pytań, że nie pytano by się o tematy na które istnieje najwięcej książek, ale o tematy z których najczęściej są *wypożyczane* książki. Tak więc zamiast dla każdego hasła zliczać liczbę książek, liczyłyby się sumę wypożyczeń przez czytelników po wszystkich książkach (pole `popularnosc_czytelnicza`).

Popularność bezpośrednich hiponimów można by wykorzystać do sortowania hiponimów w przeglądarce tezauryusa.

Rozdział 5

Implementacja katalogu przedmiotowego

W rozdziale tym zostanie zaprezentowany proces oraz wyniki implementacji struktur umożliwiających przechowywanie katalogu przedmiotowego i wykonywanie na nim operacji. Na biblioteczny katalog przedmiotowy składa się kartoteka haseł wzorcowych zwana też słownikiem oraz zbiór publikacji opisanych hasłami powyższej kartoteki. Na początku zostaną omówione zastosowane technologie oraz ogólna architektura katalogu przedmiotowego. Następnie zostanie omówiona kartoteka haseł wzorcowych i oferowane przez nią funkcje, jednak z pominięciem jej importu z formatu MARC. Omówienie implementacji reprezentacji publikacji oraz ich importu z systemu dLibra zostanie pominięte.

5.1 Zastosowane technologie

Podstawową zastosowaną technologią był język programowania Java SE 5.0. Zastosowano następujące standardowe elementy języka Java: serializację, XML (w tym DOM, XPath i walidację XML Schema) oraz JDBC ze sterownikiem do bazy Oracle ojdbc14 (JDBC Thin Oracle 10g R2). Oprócz tego zastosowano dwie niestandardowe biblioteki napisane w języku Java: dLibra i MARC4J. Biblioteka dLibra 2.2, napisana przez Poznańskie Centrum Superkomputerowo-Sieciowe, umożliwia dostęp do zawartości bibliotek cyfrowych opartych na platformie dLibra [2]. Wykorzystano ją do importu opisu publikacji z Wielkopolskiej Biblioteki Cyfrowej. Natomiast bibliotekę MARC4J 2.0 [6], umożliwiającą dostęp do rekordów bibliotecznych zapisanych w formacie MARC, wykorzystano do importu słownika KABA.

Oprócz języka Java stosowano także inne technologie. Słownik KABA ze względów efektywnościowych zdecydowano się przechowywać w pamięci RAM. Jednak już zbiór publikacji, ze względu na ich potencjalnie bardzo dużą liczbę, przechowywane się w bazie danych Oracle 10g. Do konwersji liczby gramatycznej rzeczowników

na przeciwną zastosowano analizator morfologiczny SAM-95 [12, 13]. Zastosowanie SAM-a wiązało się z koniecznością wykorzystania słowników Ispella i Aspella dla języka polskiego oraz programowej analizy rezultatów SAM-a za pomocą LEX-a, YACC-a i poleceń powłoki Linuksa. W celu ułatwienia importu słownika KABA z formatu MARC korzystano także z formatu MARC XML [7].

Do poprawiania rekordów KABA zapisanych w formacie MARC XML w pojedynczym pliku o rozmiarze 250 MB używano edytora Vim 7.0 dla Windows (nie jego graficznej wersji gVim!). Vim wykonuje płynnie operacje edycji na takim pliku. Operacje wyszukiwania wykonywane są w czasie do 20 sekund¹ (czas wyszukiwania zależy liniowo od ilości informacji do przeszukania), a operacje zapisu i odczytu na dysku wykonywane są w czasie około 15 sekund. Narzędzie takie okazało się wystarczające do poprawienia kilkudziesięciu błędów w słowniku KABA uniemożliwiających jego dalsze przetwarzanie. Do przeglądania zawartości słownika KABA zapisanej w pliku tekstowym używano programu „Lister 32” z menedżera plików Total Commander.

Podczas implementacji korzystano ze środowiska programistycznego Eclipse 3.2.0. Do optymalizacji zajętości pamięci przez słownik KABA oraz czasu jego importu wykorzystano dwa profilery w postaci wtyczek do środowiska Eclipse: JProfiler 4.2.1 i YourKit Java Profiler 5.5.6.

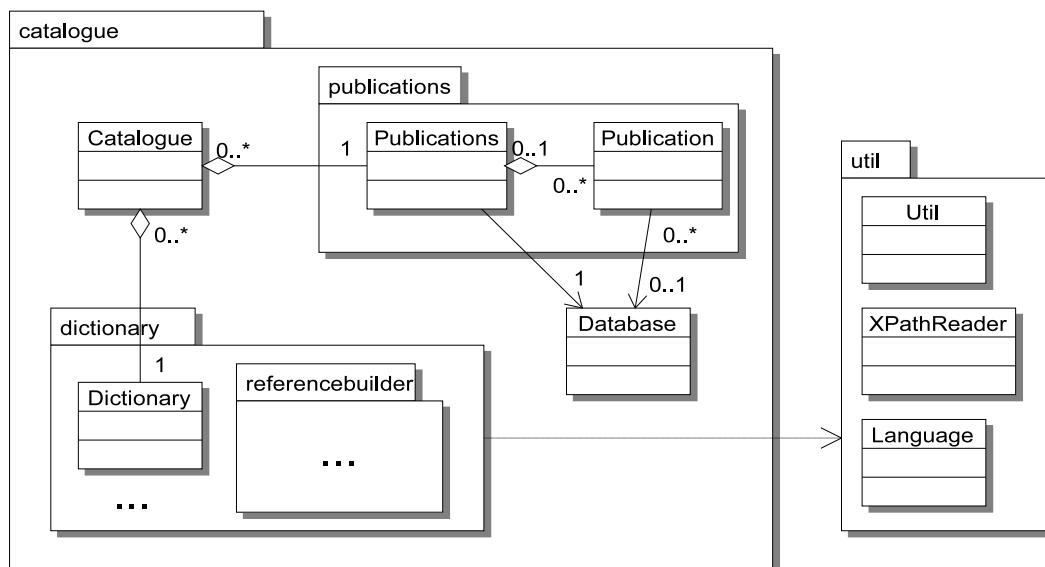
5.2 Architektura systemu

Ponieważ podstawową zastosowaną technologią był język programowania Java, to architekturę systemu najlepiej przedstawić opisując poszczególne pakiety programu napisane w tym języku. Na rysunku 5.1 przedstawiono diagram ilustrujący wszystkie pakiety systemu oraz ich najważniejsze klasy.

Głównym pakietem zaimplementowanego systemu jest `catalogue`. Implementuje on biblioteczny katalog przedmiotowy, a sam katalog reprezentowany jest przez klasę `Catalogue`. Katalog przedmiotowy składa się ze słownika haseł przedmiotowych oraz zbioru publikacji opisanych hasłami tego słownika. Słownik haseł przedmiotowych zaimplementowany jest w pakiecie `dictionary`, natomiast zbiór publikacji w pakiecie `publications`. Oba pakiety mogą być używane osobno jak też razem, tworząc katalog przedmiotowy publikacji. W ostatnim przypadku klasą integrującą oba pakiety jest wspomniana `Catalogue`.

Pakiet publikacji składa się z dwóch klas: `Publications` i `Publication`. Pierwsza klasa implementuje listę publikacji, natomiast druga klasa implementuje same publikacje. Lista publikacji przechowywana jest w bazie danych. Dostęp do bazy danych utworzonej na serwerze Oracle realizowany jest poprzez klasę `Database`. Każda

¹na komputerze z procesorem Athlon XP 2500+ pracującym z częstotliwością 1,84 GHz



Rysunek 5.1: Diagram pakietów wykonanego systemu. Katalog przedmiotowy implementowany przez pakiet `catalogue` składa się z listy publikacji implementowanej przez pakiet `publications` oraz słownika KABA implementowanego przez pakiet `dictionary`. Ostatni pakiet wykorzystuje metody użytkowe z pakietu `util`.

z publikacji może istnieć niezależnie od listy `Publications` i tym samym od bazy danych. Rozwiązanie takie istnieje by umożliwić tworzenie obiektu reprezentującego publikację w oderwaniu od listy publikacji. Każdy z obiektów publikacji można dodać do listy publikacji i tym samym do reprezentującej ją bazy danych. Pojedyncza publikacja opisana jest identyfikatorem bibliotecznym, tytułem, listą słów kluczowych oraz popularnością.

Publikacje mogą być importowane z systemu `dLibra` lub pliku tekstowego oraz eksportowane do pliku tekstowego.

Najbardziej złożonym pakietem katalogu przedmiotowego jest pakiet `dictionary` reprezentujący słownik oraz tezaurus haseł przedmiotowych. Słownik i tezaurus przechowywane są w tych samych strukturach danych, z tym że jeśli hasła przedmiotowe powiązane są hiperonimiami, to słownik staje się automatycznie tezauresem. Podstawową klasą pakietu jest klasa `Dictionary`. Dzięki niej można odwoływać się do poszczególnych elementów słownika.

Klasy pakietu `dictionary` umożliwiają reprezentację oraz analizę słownika KABA. Możliwy jest import słownika z pliku w formacie wymiennym MARC lub MARC XML z rozbudowaną analizą poprawności danych, a także eksport słownika do pliku tekstowego. Trwałość danych uzyskana jest dzięki możliwości serializacji całego słownika.

Konwersję słownika KABA do tezaurusa poprzez utworzenie hiperonimii zaimplementowano w pakiecie `referencebuilder`. Pakiet ten można traktować niez-

leżnie od głównego pakietu jako opcjonalny pakiet analizy słownika.

Pakiet `dictionary` korzysta z pakietu `util`. W pakiecie tym w trzech klasach zostały zaimplementowane funkcje użytkowe niezależne od katalogu przedmiotowego, choć przydatne w jego implementacji.

W klasie `Util` znajdują się ogólnie przyteczne metody takie jak: wyświetlanie komunikatów, walidacja pliku XML według XML Schema oraz operacje na tablicach i łańcuchach znaków. W klasie `XPathReader` zaimplementowano odczyt elementów pliku XML przy pomocy `XPath`.

W klasie `Language` zaimplementowano wybrane operacje na językach naturalnych. Dla każdego języka zaimplementowane są operacje na alfabecie takie jak: tworzenie obiektu klasy `Collator` używanego do sortowania zbiorów i map łańcuchów znaków, porównywanie łańcuchów znaków używane do sortowania klas zaimplementowanych własnoręcznie oraz operacje na małych i dużych literach alfabetu. Dodatkowo dla języka polskiego zaimplementowano odmianę rzeczowników w mianowniku przez liczbę gramatyczną. Ostatnia funkcja jest opisana w następnym punkcie.

Zastosowania tezauryasa w katalogu przedmiotowym mogą być implementowane jako podpakiety pakietu `catalogue`. Pakiety te mogą być nazwane na przykład w ten sposób: `browser`, `searchengine`, `apriori`.

5.3 Odmiana rzeczowników przez liczbę

Jak już napisano w projekcie, podczas budowy niejawnych hiperonimii często zachodzi konieczność uzyskania dla rzeczownika lub wyrażenia rzeczownikowego w mianowniku formy w przeciwnej liczbie gramatycznej. Dokładniej mówiąc, dla rzeczownika lub wyrażenia rzeczownikowego w liczbie pojedynczej należy określić jego postać w liczbie mnogiej, a dla rzeczownika lub wyrażenia rzeczownikowego w liczbie mnogiej należy określić jego postać w liczbie pojedynczej. Zmianę liczby można wykonać metodą relacyjną stosując reguły deklinacji lub też metodą słownikową korzystając z informacji zapisanych w słowniku języka polskiego lub podobnym. W budowanym systemie wybrano drugą możliwość. Dla prostoty ograniczono możliwość określania liczby przeciwnej jedynie do rzeczowników. W punkcie tym zostaną opisane cechy zadania oraz pobieżna metoda implementacji zmiany liczby.

Idea

Na początek przyjrzyjmy się postawionemu zadaniu i teoretycznej możliwości jego pełnej realizacji. Załóżmy, że dysponujemy dwukolumnową tabelą odmian rzeczowników. W pierwszej kolumnie znajdują się rzeczowniki w mianowniku liczby pojedynczej, a w drugiej znajdują się formy jakie może przyjmować wybrany rzeczownik w mianowniku liczby mnogiej.

Dlaczego dany rzeczownik może posiadać kilka form w liczbie mnogiej? Po pierwsze należy zwrócić uwagę, że w języku polskim zdarza się, że istnieją dwa rzeczowniki o tym samym tekście ale o różnym znaczeniu i *różnej deklinacji*. Z powodu różnej deklinacji mogą one mieć dwie różne formy w liczbie mnogiej. W postawionym zadaniu nie jest znane znaczenie rzeczownika, a więc nie możemy ich rozróżnić.

Po drugie mała liczba rzeczowników męskoosobowych posiada dwie powszechnie używane formy liczby mnogiej – deprecjatywną i niedeprecjatywną. Forma niedeprecjatywna jest formą podstawową i prawie zawsze używaną. Natomiast forma deprecjatywna używana jest, gdy chcemy użyć formy deprecjonującej grupę osób, to znaczy obniżającej jej wartość. Przykładowo rzeczownik „chłop” posiada formę niedeprecjatywną „chłopi” oraz deprecjatywną „chłopy”. Natomiast rzeczownik „druid” posiada formę niedeprecjatywną „druidzi” oraz deprecjatywną „druidowie”. W budowanym systemie używany jest język formalny, a więc form deprecjatywnych można nie dodawać do tabeli.

Jak można wykorzystać przedstawioną tabelę do znalezienia formy w liczbie przeciwnej? Jeśli oryginalny rzeczownik jest w liczbie pojedynczej, to wystarczy zwrócić przypisaną mu listę. Natomiast jeśli rzeczownik jest w liczbie mnogiej, to należy go poszukać w prawej kolumnie i dla wszystkich znalezionych pozycji zwrócić formę z lewej kolumny. Z powodu efektywności należy z przedstawionej tabeli utworzyć tabelę, która będzie umożliwiała dla liczby mnogiej szybkie znalezienie liczby pojedynczej. Ponieważ nie wiadomo czy podany w zadaniu rzeczownik jest w liczbie pojedynczej czy liczbie mnogiej, to powinniśmy poszukać go zarówno wśród form dla liczby pojedynczej jak i mnogiej, a następnie zwrócić odpowiednią listę.

Może się tak zdarzyć, że będą istniały dwa rzeczowniki, z których jeden będzie miał liczbę pojedynczą o podanej w zadaniu postaci, a drugi będzie miał liczbę mnogą o podanej w zadaniu postaci. W takim przypadku należy zwrócić sumę mnogościową list form liczb przeciwnych przyjmowanych przez oba rzeczowniki. Szczególnym przypadkiem takiej sytuacji jest sytuacja, w której pewien rzeczownik posiada liczbę mnogą wyglądającą identycznie jak liczba pojedyncza. W takim przypadku oczywiście zostanie zwrócona forma oryginalna.

Dla rzeczowników posiadających jedynie liczbę pojedynczą lub mnogą zostaną zwrócone puste listy.

Wykorzystanie analizatora SAM

W celu utworzenia przedstawionej tabeli zdecydowano się wykorzystać analizator morfologiczny SAM-95 [12, 13].

SAM umożliwia dla dowolnej formy pochodnej każdego wyrazu znajdującego się w jego słowniku określenie jego formy podstawowej. Na przykład dla rzeczownika w celowniku liczby mnogiej „sprzedawcom” SAM określi jego formę w mianowniku

liczby pojedynczej „sprzedawca”. Dla formy podstawowej między innymi określany jest jej tekst (na przykład „sprzedawca”) oraz jaką jest częścią mowy (na przykład rzeczownikiem). Dodatkowo SAM określa charakterystykę gramatyczną podanej formy pochodnej – dla rzeczownika jest to jego liczba oraz przypadek. Przykładowo SAM określi, że wyraz „sprzedawcom” jest w celowniku liczby mnogiej. Może się zdarzyć, że podany wyraz jest formą pochodną kilku form podstawowych. W takim przypadku SAM podaje wszystkie formy podstawowe.

Tabela odmian tworzona jest w poniżej opisany sposób.

Aby wykorzystać SAM-a należało wydobyć z niego informację dla wszystkich zapisanych w nim wyrazów. Listę wyrazów języka polskiego uzyskano z połączenia słowników Ispella i Aspella dla języka polskiego. Każdy wyraz z tej listy rozpoznawano analizatorem morfologicznym SAM. Następnie z listy rozpoznanych wyrazów wybrano rzeczowniki w mianowniku liczby mnogiej. Rzeczowniki te posiadały informacje dodane przez SAM-a o tekście ich formy podstawowej, to znaczy w liczbie pojedynczej. Dzięki temu uzyskano mapowanie rzeczowników w liczbie mnogiej na rzeczowniki w liczbie pojedynczej, czyli drugą tabelę odmian przez liczbę. Tabelę pierwszą uzyskano w prosty sposób z tabeli drugiej.

Przekształcenie pliku wynikowego SAM-a do postaci tabeli odmian wykonano przy pomocy LEX-a i YACC-a. Omówienie szczegółów tego dość skomplikowanego przekształcenia zostanie pominięte. Analiza wyników SAM-a była trudna z podobnego powodu co analiza słownika KABA. SAM był przeznaczony do wykorzystania przez ludzi, a nie przez komputery. Tym samym wynikowy format był trudno parsowalny. SAM zawierał kilka nieudokumentowanych zachowań oraz sporą liczbę błędów formatowania. Większą część prac nad komputerowym wykorzystaniem SAM-a wykonano poza pracą magisterską w ramach zajęć studenckich.

W trakcie testów poprawności tabeli odmian rzeczowników przez liczbę okazało się, że SAM stosuje oznaczenie form deprecjatywnych i nieprecjatywnych niezgodne z jego dokumentacją. Z tego powodu w tabeli często znajdowały się formy deprecjatywne, a brakowało nieprecjatywnych. Dlatego zdecydowano się tak zmodyfikować tworzenie tabel odmian, aby dla rzeczowników męskoosobowych podawane były zarówno formy nieprecjatywne jak i deprecjatywne. Większość rzeczowników męskoosobowych posiada formy deprecjatywne, choć najczęściej nie są one używane. SAM nie umożliwia określenia tego, czy dana forma deprecjatywna jest w użyciu, to znaczy czy na przykład dla rzeczownika „abstrakcjonista” używa się formy „abstrakcjonistowie” czy jedynie formy „abstrakcjonisci”. Dlatego obecnie w tabeli znajdują się obie formy, nawet jeśli forma deprecjatywna nie jest używana w języku. Nie stanowi to problemu podczas budowania hiperonimii niejawnych, gdyż formy deprecjatywne nie występują w słowniku KABA.

Z powodów implementacyjnych zdecydowano się dla rzeczowników bez liczby

pojedynczej oraz dla rzeczowników bez liczby mnogiej dodawać do tabeli wiersz o dwóch takich samych formach.

Analiza wyników

W celu określenia poprawności utworzonej tabeli odmian przeanalizowano 100 pierwszych jej wierszy. Zakres obejmowany przez tabelę jest w około 90% zgodny z zawartością trzytomowego *Słownika języka polskiego PWN* pod redakcją Mieczysława Szymczaka. Odmiana wyrazów zawartych w tabeli jest poprawna z jednym wyjątkiem. Osiem spośród przejranych rzeczowników obecnych w tabeli posiada liczbę mnogą, mimo tego że w rzeczywistości są rzeczownikami jej nieposiadającymi. Przykładowo rzeczowniki: „abstrakcyjność”, „absurdalność”, „adekwatność” według tabeli posiadają następujące formy liczby mnogiej: „abstrakcyjności”, „absurdalności”, „adekwatności”. Błąd ten wynikający z cechy SAM-a nie wpływa na działanie reguł tworzenia niejawnych hiperonimii z tego samego powodu co poprzednio.

Tabela przechowywana jest w pliku tekstowym w formie umożliwiającej szybkie uzyskanie odwzorowań w obu opisanych kierunkach. Każda para liczba pojedyncza – liczba mnoga zapisana jest jako osobny wiersz pliku. Jeśli na przykład pewien rzeczownik ma dwie formy w liczbie mnogiej, to zapisany jest w dwóch wierszach. W kolumnie pierwszej będzie występowała dwa razy ta sama forma liczby pojedynczej, a w kolumnie drugiej – raz jedna a raz druga forma liczby mnogiej.

Na rysunku 5.2 przedstawiono 30 pierwszych wierszy tabeli odmian. Wszystkie wyrazy znajdujące się przedstawionej części tabeli odmian występują w *Słowniku języka polskiego*. Natomiast wyraz „abakan” znajdujący się w *Słowniku języka polskiego* nie występuje w tabeli odmian, gdyż nie został rozpoznany przez SAM-a. Pośród wierszy tabeli znajdują się podwojone wiersze dla dwóch rzeczowników posiadających dwie formy w liczbie mnogiej – jedne z form są deprecjatywne i niepoprawne z wyżej opisanego powodu. Rzeczownikami tymi są „abderyta” i „abolicjonista”.

5.4 Kartoteka haseł wzorcowych

W punkcie tym zostanie opisana implementacja kartoteki haseł wzorcowych. Na początku zostaną przedstawione jej funkcje oraz opis klas umożliwiający zrozumienie cech implementacji słownika i sposobu posługiwania się nim. W dalszej części punktu zostaną opisane rozwiązania trudniejszych problemów i szczegóły ich implementacji. Na końcu zostanie przedstawiony format reprezentacji tekstowej słownika, dzięki któremu można zapoznać się z jego zawartością oraz sposób wywołania programu.

abakus	abakusy
abazja	abazje
abażur	abażury
abażurek	abażurki
abcug	abcugi
abderyta	abderyci
abderyta	abderytowie
abdykacja	abdykacje
abdykowanie	abdykowania
abecadło	abecadła
aberracja	aberracje
abiogeneza	abiogenezy
abisynka	abisynki
abiturient	abiturienci
abiturientka	abiturientki
abiudykacja	abiudykacje
ablacja	ablacje
ablaktacja	ablaktacje
ablaktowanie	ablaktowania
ablegier	ablegry
ablucja	ablucje
abnegat	abnegaci
abolicja	abolicje
abolicjonista	abolicjonistowie
abolicjonista	abolicjoniści
abonament	abonamenty
abonent	abonenci
abonentka	abonentki
abonowanie	abonowania
abordaż	abordaże

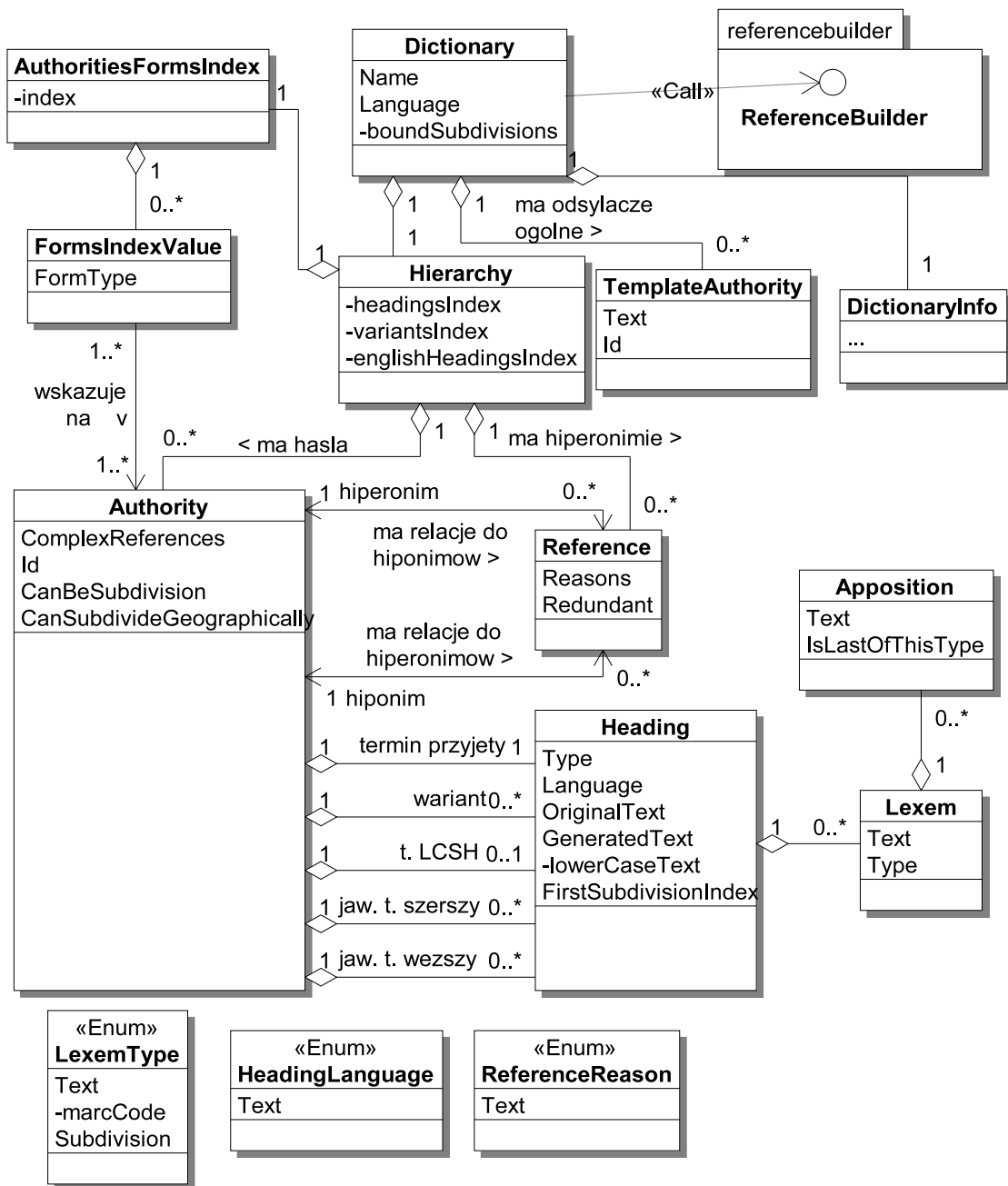
Rysunek 5.2: Tabela odmian rzeczowników przez liczbę. Pokazano 30 pierwszych wierszy pliku przechowującego tabelę.

5.4.1 Opis funkcji kartoteki i jej klas

Jak już wcześniej napisano, implementacja kartoteki haseł wzorcowych w pakiecie `dictionary` umożliwia przechowywanie słownika, jego serializację, import, eksport oraz analizę poprawności importowanych danych. Dodatkowo zostały zaimplementowane elementy umożliwiające analizę zawartości słownika. Analiza słownika umożliwia rozpoznawanie znaczenia jego elementów szczególnie pod kątem budowania niejawnych hiperonimii. W punkcie tym zostaną opisane dane przechowywane przez kartotekę oraz sposób posługiwania się nią.

Podobnie jak poprzednio najłatwiej będzie zapoznać się z zawartością pakietu `dictionary` opisując kolejno jego klasy.

Pakiet składa się z jedenastu klas oraz trzech typów wyliczeniowych zilustrowa-



Rysunek 5.3: Diagram klas pakietu dictionary implementującego słownik haseł przedmiotowych. Główną klasą pakietu jest klasa Dictionary. Słownik przechowuje między innymi hasła wzorcowe (Authority) powiązane w teaurusie hiperonimiami (Reference). Dostęp do wybranych haseł wzorcowych można uzyskać przy pomocy szeregu indeksów. Analiza zawartości słownika możliwa jest między innymi dzięki klasie Heading oraz indeksowi AuthoritiesFormsIndex. Słownik można przekonwertować do teaurusu wywołując metodę z pakietu referencebuilder.

nych na rysunku 5.3. Poniżej zostaną opisane wszystkie klasy z pominięciem metod służących do importu ich zawartości z formatu MARC. Zostaną opisane dane przechowywane przez klasy oraz metody służące do ich obsługi i analizy.

Ogólne cechy

Wspólną cechą wszystkich klas jest sortowanie alfabetyczne zawartości wszystkich kolekcji. Językiem sortowania jest język natywny słownika, czyli język polski. Podczas sortowania nie są brane pod uwagę: wielkość liter, spacje, a także znak „-” na początku haseł będących określnikami. Istnieją nieliczne i nieistotne wyjątki od tej zasady w przypadku gdy sortowanie w języku natywnym uniemożliwiałoby serializację słownika. W takim przypadku struktury są sortowane alfabetycznie w języku domyślnym języka programowania Java.

Dodatkowe wyjaśnienie należy się w przypadku sortowania haseł. Jeśli dwa hasła mają ten sam tekst, to są one sortowane według ich typu, a na końcu według ich języka. Język brany jest pod uwagę jedynie podczas sortowania listy wariantów, gdyż zakłada się, że wszystkie terminy przyjęte są w języku natywnym.

Klasa Dictionary

Podstawową klasą pakietu jest klasa `Dictionary`. Reprezentuje ona cały słownik i umożliwia przejście do jego elementów. Słownik opisany jest nazwą – w przypadku słownika KABA jest to „KABA”.

W niektórych przypadkach niezbędne jest przetwarzanie zawartości słownika zapisanej w natywnym języku naturalnym, to jest domyślnym języku słownika. Dla słownika KABA językiem natywnym jest język polski. Z tego powodu słownik KABA ma przypisany obiekt klasy `util.Language` reprezentujący język polski.

W prywatnym polu `boundSubdivisions` przechowywana jest posortowana alfabetycznie lista określników związanych. W tym przypadku określnikami związanymi nazywamy określniki zadeklarowane jako związane w polu 667 rekordu MARC. Lista ta używana jest podczas budowania hiperonimii na podstawie leksemów hasła.

Zawartość słownika KABA może być importowana z pliku w formacie MARC lub MARC XML metodami `importFrom[Xml]Marc`. Słownik KABA może być serializowany do pliku metodą `serializeTo` oraz deserializowany metodą `deserializeFrom`. Zawartość słownika można zapisać w postaci pliku testowego metodą `writeToFile`. Więcej informacji na ostatni temat znajduje się w punkcie 5.4.6.

W klasie zaimplementowane są dwie statyczne metody umożliwiające konwersję słownika z formatu wymiennego MARC do formatu MARC XML i odwrotnie. Metody te to: `convertMarcToXmlMarc` oraz `convertXmlMarcToMarc`. Metody zaimplementowane są przy użyciu biblioteki MARC4J. Pierwsza z metod po wykonaniu konwersji sprawdza poprawność słownika względem schematu XML Schema

dla MARC XML. Druga metoda nie działa do końca poprawnie, gdyż przekonwertowanego pliku nie można z powrotem przekonwertować do formatu MARC XML. Wydaje się, że jest to spowodowane błędem w bibliotece MARC4J lub słowniku KABA.

Klasa `DictionaryInfo`

W obiekcie klasy `DictionaryInfo` wydzielone zostały pozostałe informacje konfiguracyjne słownika. Oznacza to, że informacje tam przechowywane nie dotyczą bezpośrednio zawartości słownika (tak jak pole `boundSubdivisions`) a jego cech (tak jak pole `Name` lub `Language`). Cechy tam przechowywane używane są podczas importu oraz analizy zawartości słownika i dotyczą formatu nagłówek hasła, odsyłaczy orientacyjnych uzupełniających lub uwag zapisanych w hasłach. Informacjami używanymi podczas analizy są następujące pola prywatne:

- a) Lista dopowiedzeń kwalifikujących używanych w nazwach geograficznych. Przykładami takich dopowiedzeń są: „województwo”, „góra”, „rzeka”. Lista ta potrzebna jest podczas kompaktowania dwustopniowych określników geograficznych.
- b) Wzorzec służący do rozpoznawania liczbowych wyrażień chronologicznych takich jak: „1968”, „ca 1200 a.C.” (około 1200 roku przed naszą erą), „12 w.” (XII wiek), „1910-1980”.
- c) Lista słownych wyrażień chronologicznych: „starożytność”, „średniowiecze”, „renesans”.

Dwa ostatnie elementy służą do rozpoznawania wyrażień chronologicznych w nagłówkach haseł, a dokładniej dopowiedzeń lokalizujących czasowych. Dopowiedzenia lokalizujące czasowe rozpoznawane są podczas kompaktowania dwustopniowych określników geograficznych oraz podczas tworzenia hiperonimii na podstawie dopowiedzeń.

Powyższe pola wykorzystywane są w dwóch publicznych metodach pobierających łańcuch znaków i zwracających wartość logiczną. Są to następujące metody:

- `isGeographicObjectClassName` – metoda sprawdza czy podany tekst jest jednym z dopowiedzeń kwalifikujących używanych w nazwach geograficznych,
- `isTimeExpression` – metoda sprawdza czy podany tekst jest liczbowym lub słownym wyrażeniem chronologicznym.

Klasa `TemplateAuthority`

W obiektach klasy `TemplateAuthority` przechowywane są odsyłacze całkowite orientacyjne. Nie należy mylić tych odsyłaczy z odsyłaczami orientacyjnymi uzupełniającymi. Oba typy odsyłaczy odsyłają z bieżącego hasła do innego hasła, jednak na tym kończy się ich podobieństwo. Każdy odsyłacz całkowity zdefiniowany jest w

osobnym rekordzie MARC, jak wskazuje na to jego nazwa. Warto dodać, że istnieją dwa rodzaje rekordów w słowniku KABA: rekordy zwykłych haseł wzorcowych oraz omawianych właśnie odsyłaczy całkowitych.

Odsyłacz całkowity przeważnie ma następującą budowę:

<Element stały> [<wyszczególniający element zmienny>]

Poniżej podano przykłady odsyłaczy całkowitych wraz z przykładami haseł na które wskazują:

muzycy [przymiotnik od nazwy narodowości] → Muzycy węgierscy, Muzycy włoscy

zwierzęta [środowisko] → Zwierzęta bagien, Zwierzęta dżungli, Zwierzęta głębinowe

W pewnym sensie odsyłacze całkowite są podobne do odsyłaczy uzupełniających trzeciego i czwartego typu. Różnicę stanowi fakt, że odsyłacze uzupełniające wskazują na hiperonimie między zwykłymi hasłami, natomiast odsyłacze całkowite wskazują na format budowy zwykłych haseł. Rolą odsyłaczy całkowitych jest utrzymanie sformalizowania tekstów nagłówków.

Klasa `Dictionary` przechowuje posortowaną alfabetycznie listę odsyłaczy całkowitych orientacyjnych. Elementy listy są typu `TemplateAuthority`. Każdy z odsyłaczy opisany jest jego identyfikatorem w kartotece oraz tekstem nagłówka.

Obecnie odsyłacze całkowite nie są wykorzystywane w programie.

Klasa `Hierarchy`

Dzięki klasie `Hierarchy` otrzymujemy dostęp do zwykłych haseł wzorcowych zawartych w słowniku. Dzięki iteratorowi klasy można przeglądać wszystkie hasła wzorcowe, natomiast dzięki różnorodnym indeksom można wyszukiwać hasła o określonym nagłówku. Kolekcja `References` przechowuje wszystkie hiperonimie istniejące między hasłami w hierarchii.

Podstawowymi indeksami słownika haseł są: `headingsIndex`, `variantsIndex` i `englishHeadingsIndex`. Umożliwiają one wyszukiwanie haseł odpowiednio po nagłówkach ich: terminów przyjętych, terminów odrzuconych oraz terminów w języku angielskim (najczęściej z kartoteki LCSH). Zamiast podawać nagłówek jako jeden obiekt można osobno podać tekst i typ nagłówka. Wielkość liter w podanym tekście nie ma znaczenia. Dodatkowo indeks `headingsIndex` umożliwia wyszukiwanie wszystkich haseł o podanym tekście terminu przyjętego niezależnie od jego typu, a także wyszukiwanie wszystkich haseł o podanym początku tekstu terminu przyjętego. Indeksy wariantów oraz terminów angielskojęzycznych są niejednoznaczne co oznacza, że dla dowolnego szukanego nagłówka mogą zwrócić więcej niż jedno hasło wzorcowe. Oczywiście może się też zdarzyć, że zostanie zwrócony zbiór pusty. Będzie to oznaczało, że podany nagłówek nie jest terminem odrzuconym żadnego hasła (terminem angielskim w przypadku indeksu angielskojęzycznego). Wyszuki-

wanie haseł po podaniu jedynie tekstu terminu przyjętego także jest potencjalnie niejednoznaczne. Poniżej podano składnię metod służących do wyszukiwania haseł:

Authority ²	getBy	$\left\{ \begin{array}{l} \text{Heading}[\text{Text}[\text{Beginning}]] \\ \text{Variant} \\ \text{EnglishHeading} \\ \text{Heading/AuthorityForm} \end{array} \right. \left(\begin{array}{l} \text{heading} \\ \text{headingText } [, \text{headingType}] \end{array} \right)$
Collection<Authority>		
Collection<Authority>		
FormsIndexValue		

Kolekcja zwracana przez metodę `getByHeadingTextBeginning` może zawierać bardzo dużo haseł wzorcowych. Jednak mimo to czas jej działania jest niezależny od tej liczby. Wynika to z faktu, że zwracana kolekcja jest jedynie widokiem oryginalnego indeksu.

Metody `getByHeadingForm` i `getByAuthorityForm` są jedynie wywołaniami metod klasy `AuthoritiesFormsIndex` i zostaną omówione podczas omawiania tej klasy.

Wszystkie indeksy implementowane są jako posortowana alfabetycznie mapa. Sortowanie umożliwia alfabetyczne wyświetlenie ich zawartości.

Klasa `Authority`

Klasa `Authority` reprezentuje hasło wzorcowe słownika KABA. Najważniejsze informacje o hasle wzorcowym przechowywane są w nagłówkach i kolekcjach nagłówków tworzonych dla każdego hasła. Na diagramie klas przedstawione są one jako pięć powiązań z klasą `Heading`. Kolejne relacje między obiema klasami odpowiadają następującym polom w klasie `Authority`:

- `Heading` – nagłówek terminu przyjętego hasła,
- `Variants` – kolekcja nagłówków terminów odrzuconych hasła,
- `EnglishHeading` – nagłówek terminu w języku angielskim lub `null`, gdy hasło nie ma takiego nagłówka,
- `BroaderHeadings` – kolekcja nagłówków jawnych terminów szerszych podanych w polach 5XX rekordu MARC,
- `NarrowerHeadings` – kolekcja nagłówków jawnych terminów węższych podanych w polach 5XX rekordu MARC.

Pierwsze i trzecie pole jest typu `Heading`, pozostałe są typu `Collection<Heading>`. Typ hasła wynika z typu nagłówka jego terminu przyjętego.

Jeśli pole `CanBeSubdivision` nie jest puste, to oznacza to, że hasło nieokreślone może pełnić także drugą funkcję – funkcję określnika. W takim przypadku typ określnika zapisany jest jako wartość tego pola. W KABA takimi hasłami są niektóre nazwy geograficzne mogące pełnić rolę określnika geograficznego.

²Wynikiem metody `getByHeading` jest `Authority` (`null` w przypadku negatywnego wyniku wyszukiwania), natomiast wynikiem pozostałych metod jest `Collection<Authority>`.

Wartość pola `CanSubdivideGeographically` typu znakowego określa, czy hasło może być rozwinięte przy pomocy określnika geograficznego. Wartość `'a'` oznacza, że jest taka możliwość; wartość `'b'` oznacza, że jest taka możliwość ale tylko, gdy hasło jest używane w funkcji tematu; natomiast wartości `'@'` i `'n'` oznaczają, że nie ma takiej możliwości.

W polu `Id` przechowywany jest identyfikator hasła zaimportowany z rekordu w formacie MARC.

W polu `ComplexReferences` przechowywana jest kolekcja przetworzonych odsyłaczy orientacyjnych uzupełniających. Obecnie przechowywane są jedynie odsyłacze trzeciego typu. Odsyłacze zostały w ten sposób przetworzone, że na początku występuje cyfra oznaczająca typ odsyłacza, a następnie po dwukropku istotna treść odsyłacza. Dla odsyłaczy trzeciego typu jest to tekst którym powinny zaczynać się wskazywane przez odsyłacz hasła. Przeważnie będzie to tekst terminu przyjętego, choć może to być także na przykład termin przyjęty w przeciwnej liczbie gramatycznej. Informacja odczytywana jest z pola 360 rekordu MARC.

W polach `BroaderReferences` i `NarrowerReferences` przechowywane są kolekcje hiperonimii odpowiednio do terminów szerszych i węższych. Przechowywane są zarówno jawne jak i niejawne hiperonimie. Polom tym odpowiadają na diagramie klasy relacje do klasy `Reference`. Przed konwersją do tezauryśa kolekcje te są puste.

Klasy `Heading`, `Lexem` i `Apposition`

Klasa `Heading` reprezentuje nagłówek hasła. W hasła mogą występować następujące nagłówki:

- nagłówek terminu przyjętego czyli nagłówek hasła,
- nagłówki terminów odrzuconych,
- nagłówek terminu w języku angielskim,
- nagłówki jawnie podanych terminów szerszych oraz węższych.

Wszystkie z wymienionych nagłówek mają prawie identyczną budowę i dlatego mogą być reprezentowane przez tą samą klasę.

Nagłówek opisany jest przede wszystkim tekstem, typem i językiem. Tekst odczytywany jest z pola MARC i zapisywany w polu `OriginalText`. Jeśli nagłówek ma typ określnikowy, to jego zapisany tekst nie zaczyna się znakami „-”, czyli nie są one dodawane przed właściwym tekstem. Z tekstu usuwana jest kropka kończąca nagłówek, co ułatwi jego analizę.

Z tekstu nagłówek usuwana jest także ewentualna informacja o języku nagłówek. Zapisywana jest ona w polu `Language`. Język inny niż natywny mogą mieć jedynie terminy odrzucone i w języku kompatybilnym z LCSH. Te ostatnie są zawsze w języku angielskim.

Typ nagłówka wyznaczany jest na podstawie dwóch ostatnich cyfr etykiety pola nagłówka w rekordzie MARC. W słowniku KABA nazwami osobowymi, korporacyjnymi i imprezami nazywane są także tytuły dzieł napisanych przez wymienionych „autorów”. Wynika to z zasady stosowanej w kartotece KABA która mówi, że typ wieloelementowego tematu jest równy typowi jego pierwszego elementu. Zasada taka prowadzi do przekłamań i dlatego w tworzonym systemie wprowadzono zasadę mówiącą, że typ wieloelementowego tematu równy jest ostatniemu elementowi. Dlatego typy podanych przed chwilą nagłówków poprawiane są na tytuły. Nagłówki terminów w języku angielskim nie mają określonego typu, ale mimo to można przypuszczać, że jest on taki sam jak typ nagłówka terminu przyjętego.

W celu umożliwienia analizy znaczenia nagłówków dzielone są one na jednostki języka haseł przedmiotowych zwane leksemami. Leksemami są tematy, określniki, a także elementy złożonych tematów hierarchicznych. Leksemy nagłówka przechowywane są w kolekcji obiektów klasy *Lexem*. Nagłówek dla którego została przeprowadzona analiza jego elementów posiada przynajmniej jeden leksem. Analiza taka przeprowadzana jest dla wszystkich nagłówków z wyjątkiem nagłówków terminów w języku angielskim. Postępowanie takie jest uzasadnione tym, że język LCSH może się różnić w pewnym stopniu od języka KABA. Poza tym analiza nagłówków angielskich nie jest wykorzystywana w budowaniu niejawnych hiperonimii. Możliwe jest usuwanie wybranych leksemów z nagłówka, co jest wykorzystywane w regule budowania hiperonimii na podstawie leksemów.

Leksemy oprócz tekstu określone są także typem analogicznym do typu całego nagłówka. Typ leksemów określany jest na podstawie kodów podpól nagłówka w rekordzie MARC. Określenie typu określników jest proste, gdyż każdy typ określnika opisany jest innym kodem podpola. Typ pierwszego leksemu nagłówka równy jest typowi całego nagłówka w rekordzie MARC (nie w tworzonym systemie!) z tym wyjątkiem, że jeśli złożona nazwa korporacyjna zaczyna się nazwą geograficzną, to pierwszy leksem ma poprawy typ, czyli typ geograficzny. Niektóre tematy złożone z kilku leksemów zapisywane są w rekordzie MARC w większej liczbie podpól niż liczba leksemów. Dzięki analizie takich tematów udało się wyodrębnić kody podpól które zaczynają nowe leksemy. Kody te są rozpoznawane między innymi jako tytuły i podtytuły oraz podrzędne nazwy zbiorowe i podrzędne nazwy wydarzeń.

Podobnie jak poprzednio w celu umożliwienia analizy leksemów wydzielane są z nich dopowiedzenia. Każdy leksem może mieć kilka dopowiedzeń, które są zapisywane w postaci listy obiektów klasy *Apposition*. Dopowiedzenie posiada tekst oraz informację o tym, czy jest ostatnim dopowiedzeniem pewnego typu. Dopowiedzenia mogą być następujących typów: dopowiedzenia lokalizujące czasowe, dopowiedzenia lokalizujące geograficzne oraz dopowiedzenia kwalifikujące. W oryginalnym tekście nagłówka w rekordzie MARC dopowiedzenia różnych typów oddzielane są średni-

kiem, podczas gdy dopowiedzenia tego samego typu oddzielane są ukośnikiem.

Tekst dopowiedzenia może być hierarchiczny tak jak w następującym przykładzie: „Polska, województwo lubuskie”. Elementy dopowiedzenia hierarchicznego nie są wydzielane, gdyż bez ich analizy semantycznej nie wiadomo czy są to elementy dopowiedzenia hierarchicznego, czy też jest to niehierarchiczne dopowiedzenie zawierające przecinek. Jednak elementy dopowiedzenia hierarchicznego mogą być łatwo określone przy pomocy metody `parseHierarchy`.

Możliwe jest dodawanie nowych dopowiedzeń do leksemu, co jest wykorzystywane podczas kompaktowania dwustopniowych określników geograficznych.

Pole `FirstSubdivisionIndex` jest numerem pierwszego określnika w kolekcji leksemów. Pole jest wykorzystywane między innymi podczas uzyskiwania widoków tematów oraz określników metodami: `getSubjects`, `getSubdivisions`.

Tekst wygenerowany z leksemów

Po rozpoznaniu elementów nagłówka generowany jest z tych elementów tekst nagłówka i zapisywany jest w polu `GeneratedText`. Tekst ten powinien być identyczny lub podobny do tekstu `OriginalText`. Różnica może wynikać z trzech przyczyn.

Po pierwsze niektóre elementy nagłówka zawarte w nawiasach i nie będące w rzeczywistości dopowiedzeniami mogą zostać uznane za dopowiedzenia. Ten błąd algorytmu mógłby zostać poprawiony pewnym nakładem pracy. Jednak występuje on dość rzadko i jak wykazano podczas analiz wpływa on negatywnie na algorytmy budowania hiperonimii niejawnych jedynie dla kilku hiperonimii.

Drugim powodem różnic w obu tekstach są błędy formatowania w oryginalnym tekście nagłówka hasła zapisanego w rekordzie MARC. Przykładowymi błędami są: spacje umieszczone w niewłaściwym miejscu lub brak spacji, brak kropki oddzielającej leksemu tematu złożonego lub brak nawiasu zamykającego grupę dopowiedzeń. Ze względu na rozmiar słownika można zaryzykować stwierdzenie, że w formatowaniu elementów nagłówków występują błędy prawie wszystkich typów mogących teoretycznie wystąpić.

Po trzecie w wygenerowanym tekście dwustopniowe określniki geograficzne są skompaktowane.

Ponieważ wyszukiwanie nagłówków opiera się na wyszukiwaniu ich tekstów, to teksty te powinny być pozbawione błędów formatowania. Dlatego wszędzie gdzie wykorzystuje się tekst nagłówka wykorzystywany jest tekst wygenerowany z leksemów. Dokładne analizy rozpoznawania leksemów oraz generowania tekstu świadczą o tym, że tekst ten dla wszystkich nagłówków powinien przynosić tę samą informację co tekst oryginalny, zatem jego stosowanie jest bezpieczne. Odczytując tekst nagłówka powinno się zawsze używać metody `getText`. Metoda ta może zwracać tekst oryginalny lub wygenerowany – obecnie metoda zwraca tekst wygenerowany.

Typy wyliczeniowe **LexemType** i **HeadingLanguage**

Typ **LexemType** używany jest do określania typów nagłówków oraz ich leksemów. Możliwe typy podane są w tabeli 5.1 na stronie 74. Leksemy nie mogą mieć typu „nieznany”. Typ **HeadingLanguage** używany jest do określenia języka nagłówka. Możliwe języki podane są w tabeli 5.2.

Klasa **Reference**

Klasa **Reference** przechowuje informacje o każdej hiperonimii łączącej hasła. Osobna klasa stosowana jest przede wszystkim z tego powodu, że dzięki temu możliwe jest przechowywanie informacji o tym z jakiej reguły wynika dana hiperonimia. Hiperonimia oczywiście może wynikać z kilku reguł, na przykład może być jawnie podana w rekordzie oraz wynikać z leksemów hasła. Lista reguł dzięki którym hiperonimia została utworzona przechowywana jest w polu **Reasons**. Lista ta posortowana jest zgodnie z kolejnością definicji typów reguł w typie wyliczeniowym **ReferenceReason**.

Pole **Redundant** typu logicznego w zamysłu ma służyć do oznaczania relacji wynikających z przechodniości hiperonimii. Pole to może być ustawione w ostatniej fazie budowania hiperonimii, podczas szukania relacji nadmiarowych.

Typ wyliczeniowy **ReferenceReason**

Typ **ReferenceReason** używany jest do określania reguł tworzenia hiperonimii. Możliwe typy reguł podane są w tabeli 5.3.

Klasy **AuthoritiesFormsIndex** i **FormsIndexValue**

Słownictwo kontrolowane posiada dwie ważne cechy. Pierwszą cechą jest to, że istnienie synonimów rozwiązywane jest przez dodawanie do słownictwa terminów odrzuconych. Po drugie jeśli pewien wyraz języka naturalnego jest wieloznaczny, to do słownictwa kontrolowanego dodawany jest on z dopowiedzeniami.

Używanie terminów odrzuconych oraz dopowiedzeń sprawia, że słownictwo kontrolowane jest jednoznaczne. Jednak istnienie osobnych indeksów dla terminów przyjętych i odrzuconych utrudnia wyszukiwanie, gdyż przy każdym wyszukiwaniu musielibyśmy szukać haseł w obu indeksach. Istnienie w hasłach dopowiedzeń także utrudnia wyszukiwanie, gdyż aby znaleźć takie hasła musimy znać ich dopowiedzenia.

Z powyższego powodu wynika, że słownictwo kontrolowane powinno posiadać indeks, w którym po podaniu tekstu i typu nagłówka można by było znaleźć wszystkie hasła o podanym typie, posiadające termin przyjęty lub odrzucony o podanym tekście lub podanym tekście z ewentualnymi dopowiedzeniami.

Taki indeks został zaimplementowany w klasie `AuthoritiesFormsIndex`. Indeks ten nazywany jest indeksem form haseł, gdyż indeksuje wszystkie formy jakie może przyjmować hasło w innych hasłach. Indeks dla podanego nagłówka zwraca wszystkie hasła, które mogą przyjmować formę określoną przez ten nagłówek. Dodatkowo indeks określa jaką formą haseł jest podany nagłówek, to znaczy na przykład czy jest on terminem przyjętym, terminem odrzuconym, czy też terminem bez dopowiedzeń. Wynik szukania w indeksie jest obiektem klasy `FormsIndexValue`. Obiekt ten zawiera wspomnianą listę haseł (pole `Authorities`) i typ formy (pole `FormType`).

Z indeksu można korzystać dzięki dwóm metodom. Metoda `getByAuthorityForm` działa w opisany przed chwilą sposób. Działanie metody `getByHeadingForm` różni się tym, że metoda zwraca jedynie te hasła, dla których podany nagłówek jest formą terminu przyjętego (terminem przyjętym lub terminem przyjętym bez dopowiedzeń). Oznacza to, że nie są zwracane hasła, dla których podany nagłówek jest formą terminu odrzuconego.

Indeks form jest szczególnie przydatny podczas rozpoznawania elementów nagłówków haseł wzorcowych. Szczegóły dotyczące budowania i tym samym działania indeksu form zostaną podane w punkcie 5.4.5.

W klasie `Hierarchy` zostały dla wygody umieszczone metody wywołujące powyższe metody.

5.4.2 Błędy w słowniku KABA

Słownik KABA dostarczony przez NUKAT zawierał dużą liczbę różnorodnych błędów. Wynika to z faktu, że słownik nigdy nie był sprawdzany przez program komputerowy pod względem poprawności wprowadzonych danych. Dlatego zdecydowano, aby każdy z implementowanych elementów systemu wykonywał gruntowną walidację poprawności przetwarzanych danych. Każdy z elementów systemu dokładnie testowano, aby upewnić się, że działa poprawnie. Dzięki temu wydaje się, że podczas importu słownika KABA oraz tworzenia hiperonimii wykryto prawie wszystkie błędy, które mogłyby wpłynąć na niepoprawną reprezentację w systemie danych kartoteki albo też na niepoprawne działanie reguł tworzenia hiperonimii. Wszystkie błędy wykryte w wersji słownika KABA z czerwca 2006 roku zostały przekazane do Centrum NUKAT celem ich poprawienia.

Rekordy zawierające błędy są pomimo to przetwarzane. Oddziaływanie błędów jest lokalne, to znaczy jeśli na przykład w dopowiedzeniu wystąpi literówka, to ma ona jedynie wpływ na budowanie hiperonimii na podstawie tego jednego dopowiedzenia. Z tego powodu wpływ błędów istniejących w słowniku jest łatwo przewidywalny. Wpływ błędów na zachowanie systemu jest minimalizowany, to znaczy algorytmy przetwarzania rekordów KABA są tak pisane, aby były jak najmniej wrażliwe na błędy.

5.4.3 Nazwy geograficzne

Większość nazw geograficznych posiada dopowiedzenia uściślające lub objaśniające lokalizację opisywanych obiektów na świecie. Przykładowo miasto Olsztyn ma następującą nazwę przyjętą: „Olsztyn (Polska, województwo warmińsko-mazurskie)”. Oznacza ona, że miejscowość znajduje się w Polsce. Drugie dopowiedzenie używane jest tylko wtedy, gdy w danym kraju istnieją dwa obiekty tej samej klasy (tutaj dwie miejscowości), a jego zadaniem jest szczegółowe wskazanie obiektu – w tym przypadku mamy na myśli Olsztyn w województwie warmińsko-mazurskim.

Użycie nazwy geograficznej w funkcji określnika

W KABA przyjęto, że jeśli chcemy użyć nazwy geograficznej jako określnika, to musimy użyć formy dwustopniowego określnika geograficznego, która dla omawianego przykładu będzie wyglądała następująco: „– Polska – Olsztyn (województwo warmińsko-mazurskie)”. Przykładem hasła wykorzystującego tę formę jest hasło: „Dzielnice miast – Polska – Olsztyn (województwo warmińsko-mazurskie)”. W określnikach używa się innej formy, aby ułatwić wyszukiwanie oraz grupowanie hasła w indeksie alfabetycznym. Dzięki grupowaniu możemy łatwo przejrzeć wszystkie hasła dotyczące danego kraju. W przeciwnym wypadku musielibyśmy szukać nazwy miasta pośród wszystkich miast świata. Mogłoby to sprawiać problem, na przykład jeśli nie pamiętalibyśmy dokładnie jego nazwy.

Jednak w jaki sposób można automatycznie rozpoznać taki dwustopniowy określnik geograficzny, to znaczy jak znaleźć oryginalne hasło, którego formą jest ten określnik?

Dwustopniowy określnik należy sprowadzić do formy jednostopniowej poprzez jego skompaktowanie. Kompaktowanie polega na usunięciu leksemu lokalizującego i wstawieniu go do drugiego leksemu jako dopowiedzenie na pierwszej pozycji. Dopowiedzenie wstawiane jest na pierwszej pozycji z dwóch powodów. Po pierwsze dopowiedzenia lokalizujące zawsze podawane są na początku. Należy jeszcze rozpatrzyć sytuację, gdy drugi leksem zawierał już dopowiedzenie lokalizujące (na przykład „województwo warmińsko-mazurskie”). W takim przypadku należy utworzyć dopowiedzenie hierarchiczne. Jednak czy dodawane dopowiedzenie należy dodać po istniejącym dopowiedzeniu czy przed nim? W nazwach geograficznych (a takie kompaktujemy) elementy hierarchicznego dopowiedzenia lokalizującego podawane są w kolejności od najogólniejszych do najbardziej szczegółowych. Pierwszy leksem określnika dwustopniowego jest zawsze najogólniejszym lokalizatorem, więc należy go podać na pierwszej pozycji dopowiedzenia hierarchicznego.

Ciągle jednak pozostał do rozstrzygnięcia pewien szczegół. Czy kompaktując „– Polska – Olsztyn (województwo warmińsko-mazurskie)” do „Olsztyn (Polska,

województwo warmińsko-mazurskie)” należy oddzielić nowe dopowiedzenie od starych przecinkiem czy średnikiem, czyli czy utworzyć hasło: „Olsztyn (Polska, województwo warmińsko-mazurskie)”, czy też hasło: „Olsztyn (Polska; województwo warmińsko-mazurskie)”. Przecinkami oddzielane są elementy dopowiedzenia lokalizującego geograficznego. Jednak jeśli nazwa geograficzna opisuje inny obiekt niż miasto, to na końcu listy dopowiedzeń umieszcza się tak zwane dopowiedzenie kwalifikujące oddzielając je od wcześniejszych średnikiem, na przykład „... ; jezioro)”.

Wynika z tego, że jeśli pierwsze dopowiedzenie drugiego leksemu jest dopowiedzeniem lokalizującym, to należy je poprzedzić przecinkiem; a jeśli jest dopowiedzeniem kwalifikującym, to średnikiem. W opisywanym przykładzie należy użyć przecinka, gdyż „województwo warmińsko-mazurskie” jest dopowiedzeniem geograficznym a nie kwalifikującym. Ostatecznie problem rozwiązano dzięki utworzeniu listy określników kwalifikujących używanych w nazwach geograficznych. Jeśli dopowiedzenie w hasle znajduje się na tej liście to wiadomo, że kompaktując dwustopniowy określnik należy postawić przed tym dopowiedzeniem średnik, natomiast jeśli nie ma go na liście, to powinno postawić się przecinek.

Na początku planowano kompaktować dwustopniowe określniki geograficzne do nazwy podstawowej bezpośrednio przed ich rozpoznawaniem w indeksie. W międzyczasie podczas szukania rozwiązania prostszego w implementacji okazało się, że skoro w pisany programie hasła znajdowało się będzie dzięki hiperonimom zbudowanym przez komputer, a nie dzięki alfabetycznemu indeksowi haseł, to tak naprawdę dwustopniowe formy określnikowe nie są potrzebne. Dlatego dwustopniowe określniki geograficzne kompaktowane są już w czasie importu słownika i dzięki temu w całym systemie (także w indeksie) występują jednolite formy nazw geograficznych, co oczywiście ułatwia proces ich wyszukiwania. Tak więc w opisywanym przykładzie Olsztyna hasła w systemie wyglądają następująco:

Olsztyn (Polska, województwo warmińsko-mazurskie)

Dzielnice miast – Olsztyn (Polska, województwo warmińsko-mazurskie)

Dwustopniowe określniki geograficzne po imporcie istnieją w słowniku jedynie w oryginalnych tekstach nagłówków.

Użycie nazwy geograficznej w funkcji dopowiedzenia

W KABA przyjęto, że jeśli chcemy użyć nazwy geograficznej w roli dopowiedzenia (dokładniej dopowiedzenia lokalizującego geograficznego), to musimy użyć formy dopowiedzenia hierarchicznego, która dla opisywanego przykładu będzie wyglądała następująco: „Olsztyn, Polska, województwo warmińsko-mazurskie”. Przykładem jest hasło: „Ratusz (Olsztyn, Polska, województwo warmińsko-mazurskie)”. Możliwe, że w dopowiedzeniach używa się tej formy, aby nie zagnieźdzać nawiasów. Po przeanalizowaniu różnych typów haseł znajdujących się w słowniku okazało się, że

pomijając kilka wyjątków które nie wpływają negatywnie na działanie algorytmów tworzenia hiperonimii, hasła z dopowiedzeniami geograficznymi mają następującą budowę:

Leksem geograficzny (Kraj [,jednostka_podrzędna_kraju])

Leksem niegeograficzny (Miasto, Kraj [,jednostka_podrzędna_kraju])

Pierwszy format dopowiedzeń używany jest dla nazw geograficznych to jest takich, z którymi można skojarzyć pewien większy obszar (miejscowość, gmina, region, rzeka). Przykładowo Olsztyn jest nazwą geograficzną, a więc gdy chcemy dopowiedzieć gdzie na świecie znajduje się Olsztyn, używamy formatu pierwszego: „Olsztyn (Polska, województwo warmińsko mazurskie)”. W formacie w nawiasach kwadratowych podano dopowiedzenia używane tylko wtedy, gdy poprzednie dopowiedzenia nie lokalizują jednoznacznie opisywanego obiektu. Drugi format używany jest dla nazw niegeograficznych. Na przykład ratusz miejski nie jest nazwą geograficzną ale nazwą budowli, więc w tym przypadku hasło będzie wyglądało następująco: „Ratusz (Olsztyn, Polska, województwo warmińsko-mazurskie)”.

Wydaje się, że używanie w KABA dwóch różnych formatów wynika z następującego powodu. Format pierwszy jest rozsądnym podejściem. Kraj dość dobrze lokalizuje nazwy geograficzne, gdyż są dużymi obszarami. Jeśli podanie kraju nie byłoby wystarczające, to podaje się jednostkę wyszczególniającą. Natomiast dla obiektów niegeograficznych (czyli mniejszych obiektów znajdujących się prawie zawsze w miastach) naturalnym i najważniejszym dopowiedzeniem lokalizującym, które z tego powodu powinno być podane pierwsze, jest miasto.

Chcąc w podanym leksemie rozpoznać dopowiedzenia lokalizujące, należy rozpoznać czy jest to leksem geograficzny czy niegeograficzny (leksemy geograficzne to tematy geograficzne oraz określniki geograficzne). Następnie, gdy już wiemy jaki format mają dopowiedzenia, należy je skompaktować jednym z następujących algorytmów:

- dla leksemów geograficznych: Kraj, j_podrz → j_podrz (Kraj)
- dla leksemów niegeograficznych: Miasto, Kraj [,j_podrz]) → Miasto (Kraj [,j_podrz])

Przykłady:

- W haśle „Olsztyn (Polska, województwo warmińsko-mazurskie)” kompaktujemy dopowiedzenie hierarchiczne „Polska, województwo warmińsko-mazurskie” do jego postaci podstawowej: „województwo warmińsko-mazurskie (Polska)”.
- W haśle „Ratusz (Olsztyn, Polska, województwo warmińsko-mazurskie)” dopowiedzenie „Olsztyn, Polska, województwo warmińsko-mazurskie” będące dopowiedzeniem hierarchicznym zostanie skompaktowane do jego postaci podstawowej: „Olsztyn (Polska, województwo warmińsko-mazurskie)”.

Kompaktowanie dopowiedzeń hierarchicznych następuje *bezpośrednio przed ich rozpoznawaniem i wyszukiwaniem w indeksie*, a więc później niż kompaktowanie

form określnikowych. Kompaktowanie dopowiedzeń w czasie importu byłoby złym rozwiązaniem, gdyż zanim semantycznie rozpoznamy dopowiedzenie nie jesteśmy w stanie określić czy dopowiedzenie zawierające przecinki jest dopowiedzeniem hierarchicznym, czy też jest zwykłym wyliczeniem rzeczowników tak jak w haśle: „Konrad I Mazowiecki (książę mazowiecki, kujawski, sieradzki, łęczycki i krakowski ; ca 1187-1247)”. Dlatego też, w przeciwieństwie do metody kompaktującej określniki dwustopniowe będącej częścią pakietu `dictionary`, metoda kompaktująca dopowiedzenia hierarchiczne jest częścią pakietu `referencebuilder`.

5.4.4 Liczba gramatyczna haseł

W słowniku KABA w rekordzie definiującym hasło wzorcowe używa się liczby gramatycznej, która najtrafniej kojarzy się nam z definiowanym obiektem. A więc jeśli chcemy utworzyć hasło opisujące pojęcie rodzic/rodzice, to użyjemy liczby mnogiej, gdyż nie mamy na myśli konkretnego rodzica ale każdego z rodziców. Natomiast wydaje się, że pojęcie rodzina/rodziny najtrafniej opisuje słowo w liczbie pojedynczej: „rodzina”. Czasem wybór liczby nie jest prosty.

Natomiast jeśli chcemy użyć nazwy pospolitej jako dopowiedzenie, musimy użyć jej w liczbie takiej samej jak liczba leksemu do którego dodajemy dopowiedzenie. Jeśli leksem określa jeden obiekt, to dopowiedzenie powinno mieć liczbę pojedynczą; jeśli leksem określa więcej niż jeden obiekt – liczbę mnogą.

Przykładowo w słowniku KABA znajduje się hasło: „Samochody” opisane także takimi trzema wariantami: „Auta”, „Automobile”, „Samochody osobowe”. Jak widać w definicji pojęcia użyto liczby mnogiej. Poniżej znajdują się przypadki użycia tego hasła w roli dopowiedzenia:

Kabriolety (samochody osobowe)

Honda (samochody osobowe)

Honda Civic (samochód osobowy)

Mercedes 230 (samochód osobowy)

Pierwsze hasło określa typ samochodów. W opisywanym przez leksem zbiorze znajduje się więcej niż jeden obiekt, gdyż istnieje więcej niż jeden model kabrioletu. W przykładzie drugim opisujemy wszystkie modele Hondy. Natomiast w przykładzie trzecim i czwartym opisujemy tylko po jednym modelu – a więc tutaj użyjemy liczby pojedynczej.

Widać więc, że chcąc rozpoznać hasło użyte w dopowiedzeniu nie możemy posługiwać się jego liczbą – oryginalne hasło może mieć równie dobrze liczbę przeciwną.

Początkowo planowano dla wszystkich nazw pospolitych występujących w słowniku KABA, dla których znana jest forma w mianowniku przeciwnej liczby (a więc pojedynczej lub mnogiej), zindeksować w indeksie form także formę dla tej drugiej liczby gramatycznej. Oznaczałoby to, że na przykład dla istniejącego w słowniku ha-

sła „Rodzice” dodano by do indeksu także tekst „Rodzic”, który także wskazywałby na hasło „Rodzice”. Ponieważ w indeksie byłyby formy obu liczb, to znaleźlibyśmy w nim hasło bez względu na to, formy w jakiej liczbie byśmy szukali. Jednak dodawanie do indeksu formy w liczbie przeciwnej prawdopodobnie byłoby błędem z opisanego poniżej powodu.

W KABA dość często istnieją dwa różne hasła różniące się jedynie liczbą gramatyczną. Przykładem są hasła: „Brąz” (jako stop metalu) oraz „Brązy” (jako obiekty z brązu). W takim przypadku w indeksie mielibyśmy następujące powiązania tekstów i wskazywanych haseł:

„brąz” → „Brąz” lub „Brązy” (niejednoznaczna odpowiedź indeksu)

„brązy” → „Brąz” lub „Brązy” (niejednoznaczna odpowiedź indeksu)

Jeśli chcielibyśmy znaleźć oryginalne hasło dopowiedzenia użytego w hipotetycznym haśle „Posejdon z Göteborgu (brąz)”, to szukając w indeksie tekstu „brąz” dowiedzielibyśmy się, że tak naprawdę nie wiadomo jakie hasło zostało użyte jako dopowiedzenie. Jest to poprawna odpowiedź, gdyż komputer nie wie czy Posejdon jest nazwą stopu czy też obiektu ze stopu.

Jednak jeśli użyjemy jednego z tych haseł w haśle rozwiniętym w roli tematu lub określnika, to musimy go użyć w dokładnie takiej samej liczbie gramatycznej w jakiej został zdefiniowany. Jeśli teraz chcielibyśmy rozpoznać elementy składowe hasła rozwiniętego, na przykład oryginalne hasło użyte jako temat w haśle „Brąz – przewodnictwo ciepłe”, to naturalnie chcielibyśmy wyszukując w indeksie otrzymać jednoznaczny wynik – hasło „Brąz”, a nie niejednoznaczną odpowiedź: Brąz lub Brązy. Widać więc, że lepiej będzie nie wprowadzać do indeksu niejednoznaczności poprzez dodawanie do niego form liczby przeciwnej.

Lepszym rozwiązaniem, aktualnie stosowanym, jest następujące podejście. Chcąc rozpoznać dopowiedzenie znajdujemy wszystkie jego formy w mianowniku (w liczbie pojedynczej i mnogiej), ponieważ tak naprawdę nie wiemy w jakiej liczbie jest oryginalne hasło. Następnie szukamy wszystkich tych form w indeksie. Wyniki zwrócone przez indeks sumujemy mnogościowo otrzymując końcowy wynik. Jeśli wynikowy zbiór ma jeden element, to rozpoznanie jest jednoznaczne; jeśli zero elementów, to nie udało się rozpoznać dopowiedzenia; jeśli więcej niż jeden element, to rozpoznanie jest niejednoznaczne (jak w przypadku dopowiedzenia „brąz”).

Z przeprowadzonej w tym punkcie analizy wynika, że zmiana liczby gramatycznej będzie stosowana bezpośrednio przed rozpoznawaniem dopowiedzeń, a więc podobnie jak kompaktowanie dopowiedzeń hierarchicznych będzie zaimplementowana w pakiecie `referencebuilder`.

Obecnie zmiana liczby gramatycznej na przeciwną zaimplementowana jest jedynie dla rzeczowników znajdujących się w słowniku języka polskiego. Zmiana liczby nie jest wykonywana ani dla wyrażen rzeczownikowych, ani dla rzeczowników nie

występujących w słowniku języka polskiego. Zmiana liczby wykonywana jest za pomocą analizatora morfologicznego SAM-95 i została już opisana w punkcie 5.3.

5.4.5 Indeks form haseł

W tym punkcie zostanie dokładniej opisany wcześniej przedstawiony indeks form. Indeks ten powinien umożliwić rozpoznawanie wszystkich form, jakie może przyjmować hasło wzorcowe w innych złożonych z niego hasłach. Formy te to:

- termin przyjęty hasła (główny opis),
- termin odrzucony hasła (warianty, zwane też synonimami),
- termin przyjęty hasła ale z typem określnikowym, gdyż część haseł wzorcowych zapisanych w słowniku jako tematy może występować w hasle rozwiniętym także jako określnik – mają one taki sam tekst, a różnią się jedynie typem,
- forma określnika rzeczowego dla terminów przyjętych będących określnikiem formy,
- wszystkie powyższe formy bez dopowiedzeń, jeśli formy oryginalne mają dopowiedzenia.

Przedstawione formy mogą występować jako określniki oraz dopowiedzenia. Przykładowo termin podstawowy czasem podany jest w formie inwersyjnej. Forma ta, zawierająca nienaturalny dla języka polskiego szyk wyrazów, tworzona jest w celu umiejscowienia na początku hasła ważniejszego wyrazu i tym samym ułatwienia wyszukiwania hasła w indeksie alfabetycznym. Hasła takie posiadają terminy odrzucone z formą nieinwersyjną. Natomiast nie ma żadnego powodu aby w dopowiedzeniu używać szyku inwersyjnego, dlatego używa się w nich prostszego nieinwersyjnego terminu odrzuconego. W dopowiedzeniach często używa się także form haseł bez ich własnych dopowiedzeń, aby nie prowadzić do zbędnego zagnieżdżania nawiasów.

Zastosowanie formy bez dopowiedzeń zostanie przedstawione w poniższym przykładzie. W słowniku znajdują się następujące dwa hasła: „Pancerniki (okręty wojenne)” oraz „Bismarck (pancernik)”. Dzięki dodaniu do indeksu form bez dopowiedzeń jesteśmy w stanie powiązać Bismarcka z hasłem „Pancerniki (okręty wojenne)”. Szukając w indeksie tekstu „pancernik” (dokładniej wszystkich form mianownikowych tego rzeczownika a więc form „pancernik” oraz „pancerniki”) dostaniemy w odpowiedzi, że tekst „pancerniki” identyfikuje hasło „Pancerniki (okręty wojenne)”. Na marginesie należy dodać, że w rzeczywistości w słowniku jest także hasło: „Pancerniki (ssaki)”. Spowoduje to, że indeks zwróci dwa hasła, a więc nie będziemy mogli dokładnie rozpoznać co oznacza dopowiedzenie „pancernik”.

Podczas rozpoznawania tekstu w indeksie form mogą zajść następujące sytuacje:

- tekst zostanie rozpoznany jednoznacznie,
- tekst zostanie rozpoznany, ale niejednoznacznie, to znaczy wskazuje on na kilka haseł (tak jak tekst „pancerniki”),

- tekst nie będzie pasował do żadnego hasła.

Dlatego też indeks form nazywany jest indeksem niejednoznacznym.

Należy się zastanowić, co zrobić jeśli dwa różne hasła mają taką samą formę. Czy zawsze forma ta powinna wskazywać niejednoznacznie na oba hasła? Jak na przykład postąpić, gdy termin przyjęty po usunięciu dopowiedzenia ma taki sam tekst jak pewien termin odrzucony? W tym przypadku wybrano, że termin odrzucony (jako mniej ważny) nie będzie indeksowany, a więc szukając w indeksie rozpatrywanego tekstu znajdziemy hasło opisane terminem przyjętym, a nie odrzuconym.

Rozwiązaniem ogólnym jest wprowadzenie priorytetów form. Poszczególne typy form mają następujące priorytety (od najważniejszych do najmniej ważnych):

- termin przyjęty,
- termin przyjęty z typem określnikowym, jeśli hasło było oznaczone jako mogące przyjmować formę określnikową,
- termin przyjęty bez dopowiedzeń,
- terminy odrzucone,
- terminy odrzucone bez dopowiedzeń.

Jeśli dwa różne hasła mają takie same formy o tym samym priorytecie, to forma ta będzie wskazywała (niejednoznacznie) na oba hasła. Natomiast jeśli dwa różne hasła mają takie same formy o różnym priorytecie, to forma ta będzie wskazywała jedynie na hasło o wyższym priorytecie.

Każda forma hasła posiada w indeksie dwuznakowe oznaczenie jej pochodzenia, czyli sposobu utworzenia z formy podstawowej. Nagłówki będące częścią haseł wzorcowych mają kod „ ” (dwie spacje), jeśli są terminami przyjętymi lub „v ” (‘v’ i spacja), jeśli są terminami odrzuconymi. Pochodzące od nich formy nagłówków mają inne kody. Pierwszy znak kodu określa, czy nagłówek jest terminem przyjętym (ewentualnie po zmianie typu na określnikowy), czy też jest terminem odrzuconym:

- ‘ ’ – termin przyjęty,
- ‘s’ – forma określnikowa terminu przyjętego,
- ‘v’ – termin odrzucony.

Drugi znak określa, czy nagłówek powstał poprzez usunięcie dopowiedzeń z nagłówka oryginalnego:

- ‘ ’ – nie usuwano dopowiedzeń,
- ‘a’ – forma z usuniętymi dopowiedzeniami.

Informacja o typie formy potrzebna jest w przypadku, gdy chcemy znaleźć hasło w indeksie form utworzonych jedynie na podstawie terminów przyjętych, terminów odrzuconych lub też dowolnych terminów ale bez usuwania dopowiedzeń. Ma to zastosowanie na przykład podczas znajdowania hiperonimii na podstawie leksemów. Leksemy mogą być jedynie formami terminu przyjętego, natomiast nie mogą być formami terminu odrzuconego. Jeśli więc w indeksie zwrócone zostanie hasło z in-

formacją, że szukany tekst jest jedynie terminem odrzuconym tego hasła, to nie powinniśmy leksemu wiązać z takim hasłem.

Przyjrzyjmy się tworzeniu indeksu na przykładzie nazw geograficznych o leksemie „Lubowidz”. W słowniku istnieją trzy nazwy geograficzne o tym tekście. Są to:

- Hasło „Lubowidz (Polska, województwo mazowieckie)” reprezentujące miasto. Hasło to posiada wariant „Lubowidz”. Termin przyjęty hasła może pełnić funkcję określnika geograficznego.
- Hasło „Lubowidz (Polska, województwo pomorskie)” reprezentujące miasto. Hasło to posiada wariant „Lubowidz”. Termin przyjęty hasła może pełnić funkcję określnika geograficznego.
- Hasło „Lubowidz (Polska ; gmina)” reprezentujące gminę. Hasło to posiada dwa warianty: „Gmina Lubowidz (Polska)” oraz „Lubowidz (gmina)”. Termin przyjęty hasła może pełnić funkcję określnika geograficznego.

Jaka struktura zostanie zbudowana w indeksie form dla tych haseł?

Na początku należy zbudować formy pochodzące od terminów przyjętych. Po pierwsze w indeksie znajdują się terminy przyjęte wszystkich haseł. Następnie do indeksu zostaną dodane formy określnikowe wszystkich haseł (różniące się jedynie typem). Teraz należy dodać formy bez dopowiedzeń. Każde z haseł ma formę bez dopowiedzeń, która ma postać: „Lubowidz”. Ponieważ forma taka jeszcze nie jest obecna w indeksie form, to zostanie ona do niego dodana i będzie wskazywała niejednoznacznie na *wszystkie* hasła. Dodatkowo zostanie dodana określnikowa forma bez dopowiedzeń wskazująca tak samo na wszystkie hasła.

Następnie należy dodać formy pochodzące od terminów odrzuconych. Terminy odrzucone dwóch pierwszych haseł nie zostaną dodane do indeksu, gdyż są już one w nim obecne (z większym priorytetem). Natomiast zostaną do indeksu dodane dwa terminy odrzucone trzeciego hasła, gdyż żadne z nich nie jest jeszcze obecne w indeksie. Pierwszy z terminów odrzuconych trzeciego hasła, czyli nagłówek „Gmina Lubowidz (Polska)”, po usunięciu dopowiedzenia ma formę „Gmina Lubowidz”. Forma ta nie jest obecna w indeksie, a więc zostanie dodana do niego. Drugi z terminów odrzuconych trzeciego hasła, czyli nagłówek „Lubowidz (gmina)”, po usunięciu dopowiedzenia ma formę „Lubowidz”. Forma ta *jest* obecna w indeksie, a więc *nie zostanie* do niego dodana.

Na rysunku 5.4 znajduje się reprezentacja tekstowa słownika i indeksu dla omówionego przykładu. Format reprezentacji omówiony jest w następnym punkcie.

Dictionary "KABA-test" contains 3 authorities and 0 tempate authorities.
 Authorities can be found in index containing 11 headings. (...)

Authorities:

 Lubowidz (Polska, województwo mazowieckie) (geographic)
 id: "s 00084002" can be: (geographic subdiv)
 variants: [Lubowidz (geographic)]

Lubowidz (Polska, województwo pomorskie) (geographic)
 id: "s 97053841" can be: (geographic subdiv)
 variants: [Lubowidz (geographic)]

Lubowidz (Polska ; gmina) (geographic)
 id: "s 2004104776" can be: (geographic subdiv)
 variants: [Gmina Lubowidz (Polska) (geographic),
 Lubowidz (gmina) (geographic)]

Index:

 Lubowidz (Polska, województwo mazowieckie) (geographic)
 () -> Lubowidz (Polska, województwo mazowieckie) (geographic)
 -- Lubowidz (Polska, województwo mazowieckie) (geographic subdiv)
 (s) -> Lubowidz (Polska, województwo mazowieckie) (geographic)
 Lubowidz (Polska, województwo pomorskie) (geographic)
 () -> Lubowidz (Polska, województwo pomorskie) (geographic)
 -- Lubowidz (Polska, województwo pomorskie) (geographic subdiv)
 (s) -> Lubowidz (Polska, województwo pomorskie) (geographic)
 Lubowidz (Polska ; gmina) (geographic)
 () -> Lubowidz (Polska ; gmina) (geographic)
 -- Lubowidz (Polska ; gmina) (geographic subdiv)
 (s) -> Lubowidz (Polska ; gmina) (geographic)
 Lubowidz (geographic)
 (a) -> Lubowidz (Polska, województwo mazowieckie) (geographic)
 >> Lubowidz (Polska, województwo pomorskie) (geographic)
 >> Lubowidz (Polska ; gmina) (geographic)
 -- Lubowidz (geographic subdiv)
 (sa) -> Lubowidz (Polska, województwo mazowieckie) (geographic)
 >> Lubowidz (Polska, województwo pomorskie) (geographic)
 >> Lubowidz (Polska ; gmina) (geographic)
 Gmina Lubowidz (Polska) (geographic)
 (v) -> Lubowidz (Polska ; gmina) (geographic)
 Gmina Lubowidz (geographic)
 (va) -> Lubowidz (Polska ; gmina) (geographic)
 Lubowidz (gmina) (geographic)
 (v) -> Lubowidz (Polska ; gmina) (geographic)

Rysunek 5.4: Przykład ilustrujący budowę indeksu form dla haseł zawierających leksem „Lubowidz”.

5.4.6 Reprezentacja tekstowa

Zawartość słownika KABA może być zapisana w postaci pliku tekstowego. Zapisywane są następujące elementy:

- krótki opis i podsumowanie zawartości słownika,
- lista haseł wzorcowych,
- lista określników związanych zdefiniowanych w polu 667 rekordu MARC,
- lista odsyłaczy całkowitych orientacyjnych,
- indeks wszystkich form jakie mogą przyjmować hasła wzorcowe.

Wszystkie wymienione listy haseł posortowane są alfabetycznie według języka natywnego słownika. Więcej informacji o poszczególnych elementach można także znaleźć w punkcie 5.4.1.

Opis i podsumowanie zawartości słownika

Na początku pliku znajduje się krótki opis słownika oraz statystyki poszczególnych jego elementów. Można się z nich dowiedzieć między innymi ile jest haseł poszczególnych typów oraz ile jest relacji utworzonych według danej reguły.

Hasła wzorcowe

Hasła wzorcowe podane są w kolejności alfabetycznej natywnego języka słownika. Dla słownika KABA językiem natywnym jest język polski.

Opis poszczególnych haseł może zawierać następujące elementy:

- a) informacje właściwe o wybranym haśle:
 - nagłówek terminu przyjętego hasła,
 - identyfikator hasła (po etykietce „id”),
 - informacja o typie drugiej funkcji hasła (po etykietce „can be”),
 - informacja o możliwości używania po hasle określnika geograficznego (po etykietce „geogr”),
 - lista nagłówek terminów odrzuconych hasła (po etykietce „variants”),
 - nagłówek w języku kompatybilnym z LCSH (zawsze w języku angielskim) (po etykietce „english”),
- b) referencje hierarchiczne do innych haseł:
 - lista wszystkich terminów szerszych (jawnych i niejawnych) (po etykietce „broader”),
 - lista wszystkich terminów węższych (jawnych i niejawnych) (po etykietce „narrower”),
- c) jawna informacja o referencjach hierarchicznych do innych haseł (po etykietce „Original information”):
 - nagłówki jawnie podanych terminów szerszych (po etykietce „broader”),

- nagłówki jawnie podanych terminów węższych (po etykiecie „narrower”),
- lista haseł podanych w odsyłaczach orientacyjnych uzupełniających (po etykiecie „complex”).

Każde hasło posiada przynajmniej nagłówek terminu przyjętego oraz identyfikator. Pozostałe elementy są opcjonalne.

Format nagłówka. Nagłówki terminów przyjętych, odrzuconych, w języku kompatybilnym oraz terminów szerszych i węższych mają ten sam format reprezentacji tekstowej. Na początku podana jest nazwa nagłówka. Nazwa nagłówka może się nieco różnić od nazwy odpowiedniego nagłówka w rekordzie MARC, gdyż wyświetlana jest nazwa wygenerowana z leksemów. Po nazwie nagłówka podany jest w nawiasach okrągłych jego typ oraz język pod warunkiem, że typ jest inny niż nazwa pospolita, a język jest inny niż natywny. Język podawany jest dodatkowo w nawiasach kwadratowych. W tabelach 5.1 i 5.2 podano używane oznaczenia typów i języków.

Dla nagłówek w języku kompatybilnym z LCSH nie podaje się oznaczenia języka, gdyż nagłówki te w systemie zawsze są w języku angielskim.

Identyfikator hasła. Identyfikator hasła jest taki sam jak ten w rekordzie MARC.

Informacja o typie drugiej funkcji hasła. Niektóre hasła nieokreślnikowe mogą pełnić w hasłach rozwiniętych funkcję określnikową. W KABA takimi hasłami są niektóre nazwy geograficzne. Posiadają one wtedy informację „can be: (geographic subdiv)”.

Informacja o możliwości używania po hasle określnika geograficznego. Jeśli hasło posiada informację „geogr: 'a'/'b'”, to w hasle rozwiniętym może po nim wystąpić określnik geograficzny. Pole nie jest wykorzystywane w pracy magisterskiej.

Nagłówki jawnie podanych terminów szerszych i węższych. Listy zawierają nagłówki terminów szerszych i węższych podane w polach 5XX rekordu MARC. Na liście nie ma relacji skojarzeniowych.

Lista haseł odsyłaczy orientacyjnych uzupełniających. Omawiane tu odsyłacze orientacyjne uzupełniające wskazują na inne hasła zaczynające się pewnym tekstem. Na liście podane są teksty od których powinny zaczynać się te wskazywane hasła.

Listy wszystkich terminów szerszych i węższych. Na listach znajdują się hasła (nagłówki ich terminów przyjętych), do których prowadzą relacje hierarchiczne wykryte przez komputer. Po hasle podana jest lista reguł, na podstawie których

Tabela 5.1: Oznaczenia typów nagłówków haseł

Typ nagłówka	Oznaczenie typu nagłówka
nazwa pospolita	<i>brak</i>
nazwa geograficzna	geographic
nazwa korporatywna	corporate
nazwa imprezy	meeting
nazwa osobowa	personal
tytuł	title
określnik rzeczowy	general subdiv
określnik chronologiczny	chronological subdiv
określnik formy	form subdiv
określnik geograficzny	geographic subdiv
nieznany (dla nagłówków w języku kompatybilnym z LCSH)	unknown

Tabela 5.2: Oznaczenia języków nagłówków haseł.

W ramach pracy magisterskiej nie udało się ustalić nazwy języka o oznaczeniu [m] pojawiającym się w rekordach MARC słownika KABA.

Język nagłówka	Oznaczenie języka nagłówka
natywny język słownika	<i>brak</i>
angielski	[e]
francuski	[f]
łaciński	[l]
transliteracja	[t]
LANG.M	[m]
nierozpoznany (ale inny niż natywny)	[?]

Tabela 5.3: Oznaczenia reguł tworzenia relacji

Reguła tworzenia relacji	Oznaczenie reguły
relacja podana jawnie	expl
relacja utworzona na podstawie leksemów	lexem
relacja utworzona na podstawie dopowiedzeń	appos
relacja utworzona na podstawie nazw zależnościowych	depend
relacja utworzona na podstawie początków tekstów haseł	begin

została utworzona do niego relacja. W tabeli 5.3 opisano używane oznaczenia reguł tworzenia relacji. Lista reguł posortowana jest według kolejności podanej w tabeli. Hasła w listach terminów posortowane są według typu pierwszej reguły wykorzystanej do ustanowienia relacji do nich. Jeśli nie da to rozstrzygnięcia, to listy terminów węższych posortowane są mniej więcej alfabetycznie, natomiast terminów szerszych – w sposób wynikający z algorytmu ich tworzenia³.

Określniki związane

Niektóre określniki posiadają w polu 667 rekordu MARC informację o tym, że są związane. Podczas importu słownika odczytywane są takie informacje i tworzona jest lista określników związanych. Lista ta jest także umieszczana w reprezentacji tekstowej słownika.

Lista odsyłaczy całkowitych orientacyjnych

W reprezentacji tekstowej słownika znajduje się także lista odsyłaczy całkowitych orientacyjnych.

Indeks wszystkich form haseł wzorcowych

Indeks ten umożliwia odnalezienie hasła wzorcowego po podaniu jego typu oraz tekstu dowolnej jego formy. Hasło wzorcowe może przyjmować w innych hasłach różne formy: formę terminu przyjętego, terminu odrzuconego, formę określnikową lub formę bez dopowiedzeń. Formy te są kluczami indeksu. Wartością indeksu jest zbiór wszystkich haseł, które mogą być opisane podaną formą, to znaczy jest ona ich terminem przyjętym, terminem odrzuconym, formą określnikową terminu przyjętego lub formą bez dopowiedzeń. Formy posiadają dwuznakowe oznaczenie ich typu. Dla ułatwienia wyszukiwania kolejne hasła niejednoznacznych wartości indeksu poprzedzone są znakami „>>”.

Przykład

Na rysunku 5.5 przedstawiono reprezentację tekstową przykładowych elementów tezaurusa KABA.

³Sortowanie aktualnie nie jest dokładne, gdyż przeprowadzane jest osobno przez poszczególne algorytmy. I tak relacje utworzone na podstawie leksemów lub dopowiedzeń są sortowane alfabetycznie wspólnie a nie oddzielnie. Natomiast relacje jawne mają kolejność taką samą jak kolejność oryginalnie podana w hasle (na końcu występują posortowane alfabetycznie relacje do tego hasła podane jawnie jedynie na liście terminów szerszych innych haseł). W przyszłości powinno się zaimplementować dokładne sortowanie alfabetyczne zarówno list terminów węższych jak i szerszych. Ewentualnie można zachować oryginalną kolejność jawnie podanych terminów węższych.

Dictionary "KABA" contains 112671 authorities and 256 template authorities. Authorities have following types:

type	number of authorities
topic	48688
general subdiv	3481
chronological subdiv	325
form subdiv	760
geographic	18801
geographic subdiv	1
corporate	10287
meeting	575
personal	27476
title	2277

Authorities can be found in index containing 403567 headings.

Authorities are connected by 140929 hierarchical relations.

Relations are created for these reasons:

reason	number of relations
expl	69970
lexem	43377
appos	16146
depend	1206
begin	13809

Authorities, template authorities and index are listed below in natural order of Polish language.

Authorities:

 Marchew

id: "s 99080696" geogr: 'a'
 variants: [Daucus ([1])]
 broader: [Baldaszkowate [expl]]
 narrower: [Marchew zwyczajna [begin]]
 Original information:
 broader: [Baldaszkowate]
 complex: [3:Marchew]

 Marchew (warzywa)

id: "s 99080698" geogr: 'a'
 variants: [Marchew siewna (warzywa), Marchew siewna -- system korzeniowy]
 english: Carrots (unknown)
 broader: [System korzeniowy [expl], Warzywa [expl, appos]]
 narrower: [Marchew (warzywa) -- produkcja i handel [lexem]]
 Original information:
 broader: [System korzeniowy, Warzywa]
 (...)

Rysunek 5.5: Reprezentacja tekstowa tezauryasa KABA. Wielokropek oznacza, że w pliku znajduje się więcej elementów pewnego typu.

```

Bound subdivisions:
-----
-- aktywność
-- aktywność wody
-- akustyka i fizyka
(...)

Template authorities:
-----
i [nazwa geograficzna]
i [przedmiot]
kolekcje [przedmiot]
(...)

Index:
-----
Marchew                ( ) -> Marchew
Marchew (warzywa)      ( ) -> Marchew (warzywa)
Marchew ogrodowa      (v ) -> Marchew zwyczajna
Marchew pastewna      (v ) -> Marchew zwyczajna
Marchew siewna         (v ) -> Marchew zwyczajna
Marchew siewna (warzywa) (v ) -> Marchew (warzywa)
Marchew siewna -- system korzeniowy (v ) -> Marchew (warzywa)
Marchew zwyczajna     ( ) -> Marchew zwyczajna
(...)

```

Rysunek 5.5: Kontynuacja

5.4.7 Wywołanie programu

W tym punkcie zostanie omówiony sposób wywołania programu importującego słownik KABA oraz transformującego go do tezaurusa. Program będzie też wykonywał przykładowe analizy zawartości tezaurusa. Obecnie rozwój systemu semantycznego katalogu przedmiotowego opartego na tezaurucie jest w stadium budowania tezaurusa, dlatego zaprezentowane zostaną jedynie operacje na słowniku a nie na całym katalogu przedmiotowym. Co więcej ciągle jeszcze trwają prace programistyczne, choć już prawie zostały zakończone. Dlatego nie będzie prezentowany ostateczny program, a jedynie metoda napisana w programie Java.

Na rysunku 5.6 przedstawiono metodę wykorzystującą pakiet `dictionary` do importu słownika KABA, jego transformacji do tezaurusa oraz zaprezentowania trzech przykładowych analiz zawartości słownika KABA.

Słownik KABA udostępniony jest w starszym i bardziej rozpowszechnionym formacie MARC (dokładniej w formacie wymiennym MARC). Elementy tekstowe rekordów są w nim zakodowane w UTF-8. Słownik można importować bezpośrednio z tego formatu, jednak bardziej rozsądna jest jego wcześniejsza konwersja do formatu MARC XML. Dzięki temu będzie można wcześniej sprawdzić zgodność słownika z

```
import dd.catalogue.dictionary.*;

/***** Etap 1. Import słownika i konwersja do tezaurusa *****/

//konwersja z formatu MARC do MARC XML
Dictionary.convertMarcToXmlMarc("files/kaba.mrc", "files/kaba.xml");

//import (potem zapis do pliku tekstowego i serializacja)
Dictionary dic = new Dictionary("files/kaba.config");
dic.importFromXmlMarc("files/kaba.xml");
dic.writeFile("files/kaba.txt", true);
dic.serializeTo("files/kaba.serialize");

//budowanie hiperonimii (ew. z deserializowanego słownika, potem zapis
//do pliku tekstowego i serializacja)
//dic = Dictionary.deserializeFrom("files/kaba.serialize",
//                                "files/kaba.config");
dic.buildReferences();
dic.writeFile("files/thesaurus_kaba.txt", true);
dic.serializeTo("files/thesaurus_kaba.serialize");

/***** Etap 2. Analizy na słowniku i tezausie KABA *****/

//dic = Dictionary.deserializeFrom("files/thesaurus_kaba.serialize",
//                                "files/kaba.config");

//wyświetlenie terminów przyjętych o dwóch określnikach geograficznych
for (Authority authority : dic.getHierarchy()) {
    int geogrSubdivCount = 0;
    for (Lexem subdivision : authority.getHeading().getSubdivisions())
        if (subdivision.getType() == LexemType.GEOGRAPHIC_SUBDIV)
            geogrSubdivCount++;
    if (geogrSubdivCount >= 2)
        System.out.println(authority);
}

//wyświetlenie przyjętych nazw pospolitych o postaci inwersyjnej
//(zawierających przecinek)
for (Authority authority : dic.getHierarchy()) {
    Heading heading = authority.getHeading();
    if (heading.getType() == LexemType.TOPIC) {
        String subjectText = heading.getSubjects().get(0).getText();
        if (subjectText.matches(".*, .*"))
            System.out.println(heading);
    }
}

//wyświetlenie przyjętych nazw pospolitych bez terminów szerszych
for (Authority authority : dic.getHierarchy())
    if (authority.getHeading().getType() == LexemType.TOPIC)
        if (authority.getBroaderReferences().size() == 0)
            System.out.println(authority.getHeading());
```

Rysunek 5.6: Kod programu importującego słownik KABA, transformującego go do tezaurusa i wykonującego przykładowe analizy jego zawartości.

podstawowymi regułami zawartymi w schemacie MARC XML. Dodatkowo uzyska się czytelną postać słownika, która przyda się podczas analizy błędów wykrytych w nim w czasie działania programu. Program operuje na plikach zawartych w katalogu „files”. W tym przypadku plik „kaba.mrc” jest konwertowany do pliku „kaba.xml”.

Następnym etapem jest import słownika KABA z pliku MARC XML. Na początku należy utworzyć puste struktury słownika przygotowane na import. Podczas tego procesu z pliku konfiguracyjnego „kaba.config” wczytywane są cechy słownika niezależne od zawartości. Większość tych cech będzie dostępna poprzez klasę `DictionaryInfo`. Następna metoda importuje dane słownika z formatu MARC XML. Podczas jej działania mogą być generowane komunikaty o ewentualnych błędach w zawartości słownika KABA. Po zakończeniu importu zawartość słownika można zapisać do pliku tekstowego celem zapoznania się z jego zawartością. Plik jest zakodowany w UTF-8. Można także zapisać zawartość słownika do pliku binarnego (tak zwana serializacja). Umożliwi to późniejsze szybkie wczytanie słownika z tego pliku (czyli wykonanie tak zwanej deserializacji). Dzięki temu nie trzeba będzie importować słownika przy każdym korzystaniu z niego.

Po imporcie słownika można go przetransformować metodą `buildReferences` do tezaury. Podczas działania tej metody mogą być generowane komunikaty o ewentualnych błędach w zawartości słownika KABA wykrytych na tym etapie. Tezaurus może być zapisany do pliku tekstowego celem zapoznania się ze zbudowanymi hiperonimiami, a także zserializowany celem zaoszczędzenia czasu podczas jego następnego użycia.

W przedstawionym kodzie zostały także podane metody wykonania trzech przykładowych analiz zawartości słownika KABA. W pierwszym przykładzie zostaną wyświetlone wszystkie terminy przyjęte, których nagłówki zawierają dwa określniki geograficzne. W drugim przykładzie zostaną wyświetlone wszystkie terminy przyjęte, które są nazwami pospolitymi o postaci inwersyjnej. Przykładem nazwy inwersyjnej jest hasło „Katyń, Zbrodnia”. W trzecim przykładzie zostaną wyświetlone wszystkie hasła będące nazwami pospolitymi i nie mające hiperonimów.

Wyświetlane komunikaty

Na rysunku 5.7 przedstawiono komunikaty, które powinien wygenerować przedstawiony kod importu słownika i jego konwersji do tezaury. Dodatkowo mogą być generowane komunikaty o błędach w słowniku KABA oraz inne błędy i ostrzeżenia.

Złożoność pamięciowa i obliczeniowa

Słownik zajmuje dużo miejsca w pamięci RAM. Oprócz tego wirtualna maszyna Javy (JVM) alokuje zawsze trochę więcej miejsca w pamięci dynamicznej niż jest

```
Import haseł wzorcowych ...
10000 haseł wzorcowych zostało zaimportowanych do tej pory
20000 haseł wzorcowych zostało zaimportowanych do tej pory
<...>
90000 haseł wzorcowych zostało zaimportowanych do tej pory
100000 haseł wzorcowych zostało zaimportowanych do tej pory
110000 haseł wzorcowych zostało zaimportowanych do tej pory
Budowanie indeksu dozwolonych form haseł wzorcowych ....
Łącznie zaimportowano 112929 haseł wzorcowych (zwykłych i odsyłaczy
    całkowitych orientacyjnych) w czasie 707 sekund
Serializacja ... wykonana

[Deserializacja ... wykonana w czasie 110 sekund]
Wiązanie haseł relacjami ...
Wiązanie haseł relacjami jawnymi
Wiązanie haseł relacjami na podstawie leksemów oraz dopowiedzeń
Wiązanie haseł relacjami na podstawie nazw zależnościowych
Wiązanie haseł relacjami na podstawie początków tekstów haseł
Utworzono łącznie 140929 powiązań między hasłami w czasie 112 sekund
Serializacja ... wykonana
```

Rysunek 5.7: Komunikaty generowane przez przedstawiony na rysunku 5.6 kod importujący słownik i konwertujący go do tezaurusa.

aktualnie wykorzystywane. Alokacją pamięci dynamicznej sterują trzy argumenty JVM:

- a) Xmx – Określa maksymalną ilość pamięci dynamicznej jaką można zaalokować. Próba alokacji większej ilości pamięci spowoduje błąd programu. Domyślna wartość dla JVM firmy Sun to 64 MB.
- b) Xminf (XX:MinHeapFreeRatio) – Jeśli struktury dynamiczne programu nie mieszczą się w aktualnie zaalokowanej pamięci dynamicznej, to JVM zwiększa ilość zaalokowanej pamięci. Pamięć jest alokowana do takiego poziomu, aby stosunek ilości zaalokowanej pamięci nie zajętej przez struktury do całej zaalokowanej pamięci równy był Xminf. Domyślna wartość to 0,4.
- c) Xmaxf (XX:MaxHeapFreeRatio) – Jeśli duża część zaalokowanej pamięci jest nieużywana, to JVM zwraca część zaalokowanej pamięci do systemu operacyjnego. Wielkość zaalokowanej pamięci jest zmniejszana, gdy stosunek ilości zaalokowanej pamięci nie zajętej przez struktury do całej zaalokowanej pamięci jest większy niż Xmaxf. Wielkość zaalokowanej pamięci jest zmniejszana do takiego poziomu, aby wyżej podany stosunek był równy Xmaxf. Xmaxf powinien być większy od Xminf. Domyślna wartość to 0,7.

Wirtualnej maszynie Javy pozwolono zaalokować nie więcej niż 1300 MB. Aby ograniczyć wielkość alokowanej pamięci ustawiono Xmaxf na 0,3. Podczas deseriali-

zacji alokowane jest dużo dodatkowej pamięci na bufor deserializowanego pliku. Po deserializacji pamięć ta mogłaby być zwolniona. Dlatego ustawiono Xmaxf na niską wartość 0,35. Niestety wydaje się, że implementacja wirtualnej maszyny Javy firmy Sun dla systemu Windows nie działa poprawnie, gdyż dealokuje znacznie mniej pamięci niż powinna.

Struktury słownika w pamięci RAM zajmują około 400 MB. Program był optymalizowany profilerem pod względem złożoności pamięciowej. Utworzenie słownika w wyniku importu powoduje zaalokowanie około 430 MB pamięci. Wykonanie serializacji takiego słownika spowoduje zwiększenie ilości zaalokowanej pamięci do około 730 MB. Utworzenie słownika w wyniku deserializacji spowoduje zaalokowanie aż około 900 MB pamięci, z powodu używania dużego buforu wczytywanego pliku. Zalecane jest tworzenie słownika w wyniku deserializacji, gdyż jest o wiele szybsze pod warunkiem, że nie będzie używany do tego systemowy plik wymiany. Dlatego zaleca się posiadanie więcej niż 1 GB pamięci operacyjnej – 1,5 GB wystarczy do efektywnej pracy. Konwersja słownika z formatu MARC do MARC XML nie używa dużo pamięci RAM.

Import słownika KABA jest wykonywany w ciągu niecałych 12 minut⁴, a hiperonimie budowane są w około 2 minuty. Serializacja wykonywana jest w czasie rzędu minuty, a deserializacja w czasie rzędu 2-3 minut. Analizy na wszystkich hasłach słownika wykonywane są w czasie rzędu kilku sekund.

⁴na komputerze z procesorem Athlon XP 2500+ pracującym z częstotliwością 1,84 GHz

Rozdział 6

Implementacja budowy niejawnych hiperonimii

Zaimplementowano tworzenie hiperonimii jawnych oraz tworzenie hiperonimii niejawnych według czterech przedstawionych w projekcie reguł. Reguły te to tworzenie hiperonimii: na podstawie leksemów, na podstawie dopowiedzeń, dla nazw zależnościowych oraz wiązanie hasła z hasłami zaczynającymi się nim. Dobrym rozwiązaniem jest budowanie hiperonimii w podanej kolejności. Dzięki temu można uniknąć tworzenia wielu nadmiarowych hiperonimii.

Hiperonimie na podstawie dopowiedzeń powinny być budowane, gdy są już znane hiperonimie na podstawie leksemów. Jeśli hiperonim utworzony na podstawie leksemów zawiera to samo dopowiedzenie co hasło rozwinięte, to dla rozwiniętego hasła nie warto tworzyć hiperonimii na podstawie tego dopowiedzenia. Hiperonimia do tego dopowiedzenia i tak zostanie utworzona na podstawie leksemu.

Jeśli dla hasła zależnościowego wyznaczyliśmy hiperonim do jego pierwszej składowej, to nie warto hiperonimii tej dodatkowo oznaczać jako wynikającej z reguły wiązania hasła z hasłami zaczynającymi się nim. Dlatego rozsądnym wydaje się wyznaczanie hiperonimii na podstawie nazw zależnościowych przed wyznaczeniem hiperonimii na podstawie wiązania hasła z hasłami zaczynającymi się nim.

Każda z reguł tworzenia hiperonimii rozpoznaje pewne elementy nagłówków. Na przykład podczas tworzenia hiperonimii na podstawie leksemów rozpoznaje się leksemy lub ich grupy. Jeśli element taki powinien być rozpoznany (i tym samym powinna być utworzona hiperonimia), a mimo to z pewnego powodu nie został rozpoznany, to informacja o tym zapisywana jest w logu. Informacje takie można prześledzić i dzięki temu zapoznać się z powodami niepełnego utworzenia hiperonimii. Sytuacja taka może się zdarzyć z następujących powodów:

- W słowniku KABA może istnieć błąd uniemożliwiający rozpoznanie elementu nagłówka. Błędy takie powinny zostać poprawione przez centrum zarządzające słownikiem.

- Nierozpoznanie może wynikać z błędu w algorytmie lub implementacji budowania hiperonimii. W takim przypadku należy poprawić algorytm lub implementację.
- Nierozpoznanie może wynikać z tego, że do utworzenia hiperonimii potrzebna jest informacja semantyczna, która nie została dołączona do słownika. Nie przeszkadza to w tworzeniu hiperonimii przez czytelnika biblioteki, jednak automatyczne utworzenie hiperonimii jest utrudnione. Sytuacja może być rozwiązana przez dodanie tej informacji do słownika lub ewentualnie zastosowanie zewnętrznej ontologii.

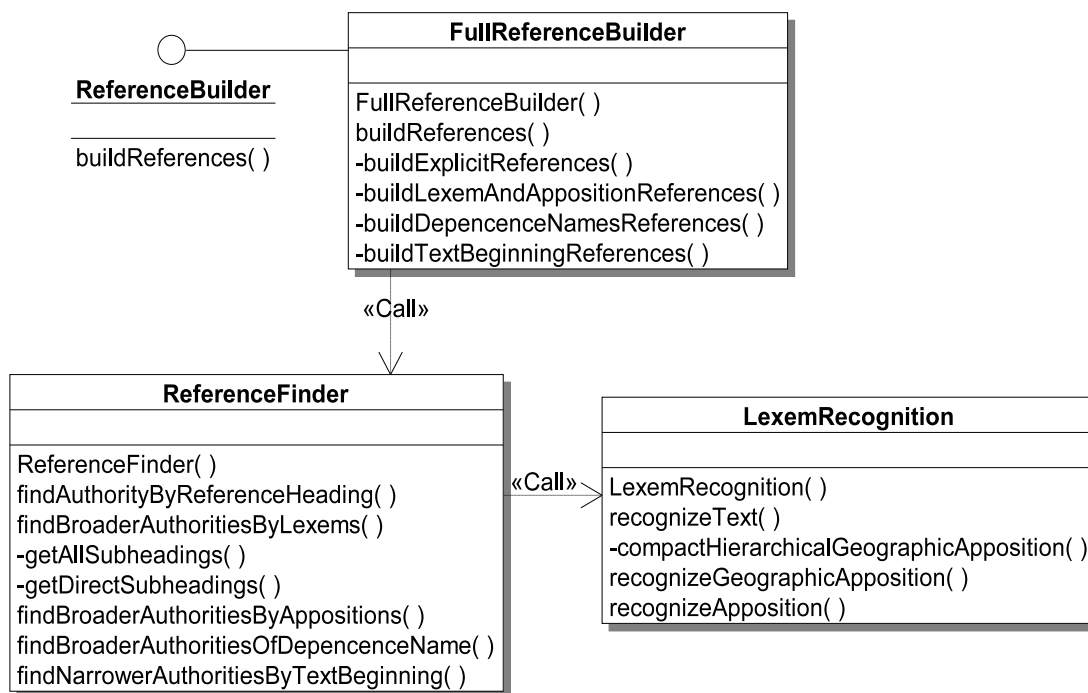
Na początku rozdziału zostanie opisana budowa modułu tworzenia hiperonimii. Następnie zostanie omówiona implementacja poszczególnych reguł w podanej przed chwilą kolejności. Podczas implementacji i analizy działania reguł okazało się, że niektóre z nich można by poprawić. Dlatego zostaną też podane propozycje ulepszeń istniejących oraz wprowadzenia nowych reguł budowy hiperonimii. Po zbudowaniu wszystkich hiperonimii należy zastanowić się globalnie nad siecią połączeń hierarchicznych i odpowiedzieć, czy należałoby coś jeszcze zrobić, aby polepszyć jej strukturę. Rozdział zostanie zakończony analizą jakości budowanego tezaurusa.

6.1 Opis klas i ich metod

Architektura modułu tworzenia hiperonimii `referencebuilder` jest prosta. Działanie modułu polega na wykonaniu szeregu metod. Metody te operują na danych przechowywanych w module `dictionary`. Sam moduł `referencebuilder` nie przechowuje żadnych danych. Z modułu `dictionary` odczytywane są potrzebne informacje, a jedyne dokonywane zmiany ograniczają się do dodawania hiperonimii między hasłami. Metody modułu podzielone są tematycznie na trzy klasy, przy czym klasa pierwsza wywołuje metody klasy drugiej, a klasa druga wywołuje metody klasy trzeciej. Diagram klas modułu przedstawiony jest na rysunku 6.1.

W celu zbudowania hiperonimii (jawnych oraz niejawnych) należy wywołać metodę `buildReferences` interfejsu `ReferenceBuilder`. Interfejs ten jest implementowany przez klasę `FullReferenceBuilder`. Klasa ta odpowiedzialna jest za powiązanie wszystkich haseł hiperonimiami na podstawie wszystkich reguł. Metoda `buildReferences` wywołuje kolejno metody budowania hiperonimii dla poszczególnych reguł. Każda z tych metod wywołuje dla pojedynczych nagłówków odpowiednią metodę z klasy `ReferenceFinder`. Zadaniem metod klasy `ReferenceFinder` jest odnalezienie haseł, do których powinny być utworzone hiperonimie od podanego nagłówka na podstawie rozpatrywanej reguły.

Aby odnaleźć hasło powiązane hierarchicznie należy rozpoznać element hasła właściwy dla reguły tworzenia hiperonimii. Przy budowaniu hiperonimii jawnych, na

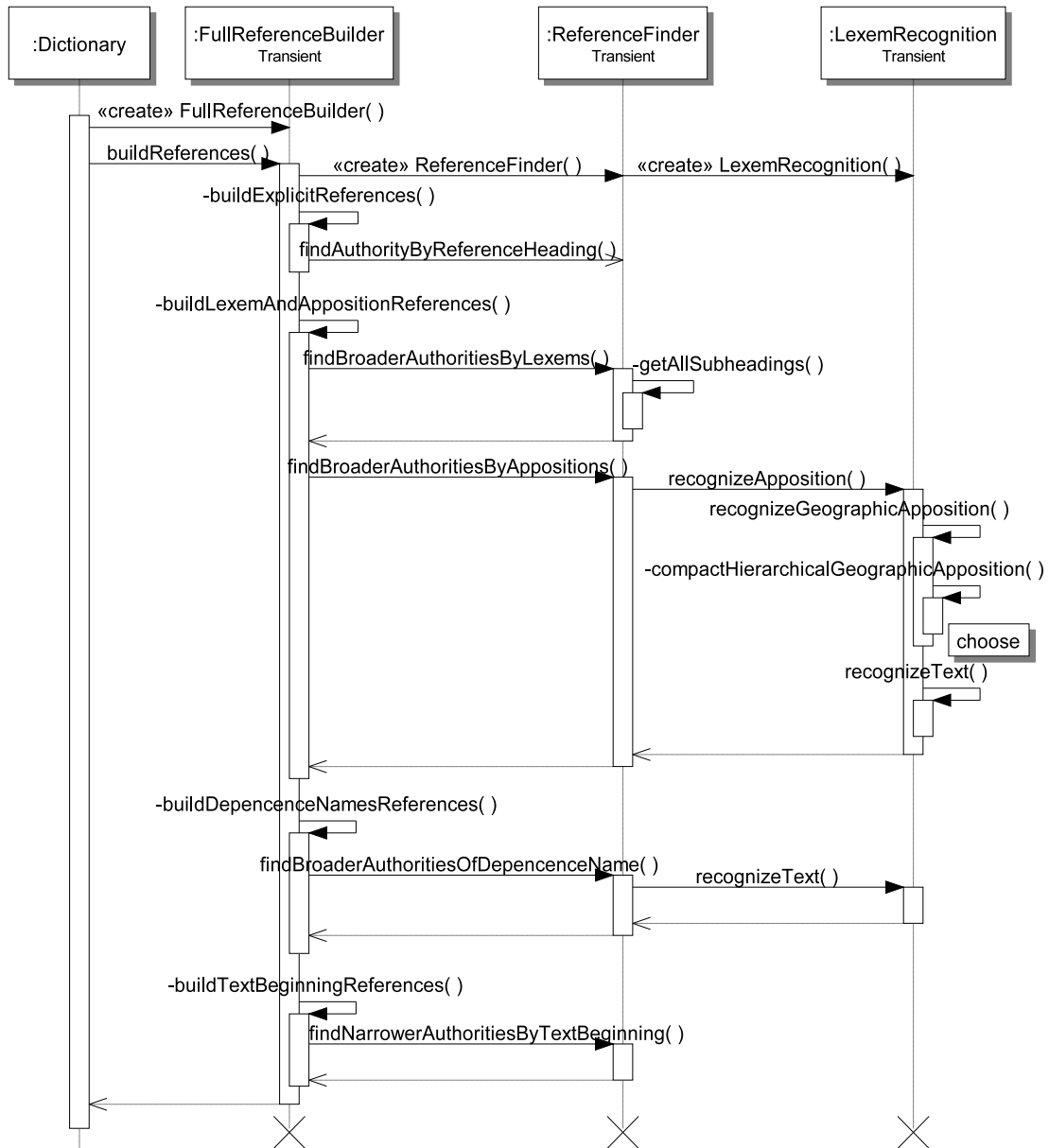


Rysunek 6.1: Diagram klas modułu `referencebuilder` składającego się z jednego interfejsu oraz trzech klas. Klasy kooperują ze sobą w porządku liniowym delegując wykonanie części operacji do klas podrzędnych.

podstawie leksemów oraz na podstawie początku hasła elementy można rozpoznawać bezpośrednio przy pomocy indeksów podstawowych oraz indeksów form. Rozpoznawanie dopowiedzeń oraz składowych nazw zależnościowych jest bardziej skomplikowane i dlatego zostało oddelegowane do klasy `LexemRecognition`. Rozpoznając dopowiedzenie musimy mieć na uwadze, że inaczej należy postąpić podczas rozpoznawania dopowiedzenia geograficznego, a inaczej rozpoznając dopowiedzenie kwalifikujące. W pierwszym przypadku należy wywołać `recognizeGeographicApposition`, a w drugim – metodę `recognizeText`.

Działanie modułu można bardzo dobrze przedstawić przy pomocy diagramu sekwencji z rysunku 6.2.

W następnych punktach zostaną kolejno omówione reguły tworzenia niejawnych hiperonimii.



Rysunek 6.2: Diagram sekwencji modułu `referencebuilder`. Obiekty klas modułu tworzone są tylko na czas działania metody `buildReferences`.

6.2 Wiązanie haseł z ich leksemami

W punkcie tym zostanie przedstawiona metoda budowania hiperonimii na podstawie leksemów hasła, a także wyniki jej działania.

6.2.1 Opis działania

Reguła zaimplementowana jest w trzech metodach:

- `getDirectSubheadings`,
- `getAllSubheadings`,
- `findBroaderAuthoritiesByLexems`.

Pierwsza metoda dla podanego nagłówka znajduje wszystkie nagłówki powstałe poprzez usunięcie jednego z leksemów: określnika lub tematu. Jeśli temat jest wieloleksemowy, to można usuwać tylko jego ostatni leksem. Zwracane nagłówki są posortowane od nagłówka nie zawierającego ostatniego leksemu do nagłówka nie zawierającego pierwszego usuwalnego leksemu. Na przykład dla nagłówka hasła: „Żydzi – Bawaria (Niemcy) – historia – źródła” metoda zwróci następujące nagłówki:

- Żydzi – Bawaria (Niemcy) – historia
- Żydzi – Bawaria (Niemcy) – źródła
- Żydzi – historia – źródła
- Bawaria (Niemcy) – historia – źródła

Nagłówki te mogą być teoretycznie terminami szerszymi oryginalnego nagłówka pod warunkiem, że są one terminami przyjętymi w słowniku KABA (to znaczy, że takie rekordy dodano do słownika).

Metoda druga wykonuje to samo co metoda pierwsza, z tą różnicą że usuwa nie tylko po jednym leksemie, ale także (iteracyjnie) po dwa, trzy leksemy i tak dalej. Wynikiem jej działania będą wszystkie podciągi leksemów nagłówka, które mogą być terminami szerszymi. Zwracane nagłówki są posortowane według następujących dwóch kryteriów (na początku jest brane pod uwagę pierwsze):

- od najbardziej szczegółowych nagłówków (mających najwięcej leksemów) do najmniej szczegółowych,
- od nagłówków nie zawierających leksemów występujących na końcu oryginalnego nagłówka (czyli na początku tematy) do nagłówków zawierających leksemy występujące na końcu oryginalnego nagłówka.

Dla rozpatrywanego przykładu metoda zwróci następujące nagłówki:

- Żydzi – Bawaria (Niemcy) – historia
- Żydzi – Bawaria (Niemcy) – źródła
- Żydzi – historia – źródła
- Bawaria (Niemcy) – historia – źródła

- Żydzi – Bawaria (Niemcy)
- Żydzi – historia
 - Bawaria (Niemcy) – historia
- Żydzi – źródła
 - Bawaria (Niemcy) – źródła
 - historia – źródła
- Żydzi
 - Bawaria (Niemcy)
 - historia
 - źródła

Metoda działa w ten sposób, że wywołuje metodę `getDirectSubheadings` dla podanego nagłówka, generując zbiór nagłówków posiadających o jeden leksem mniej niż oryginalny nagłówek. Następnie metoda `getDirectSubheadings` wywoływana jest dla wszystkich nagłówków znajdujących się w wygenerowanym zbiorze, generując zbiór nagłówków posiadających o dwa leksemy mniej niż oryginalny nagłówek. Postępowanie jest kontynuowane aż do wygenerowania nagłówków jednoleksemowych. Wygenerowane zbiory są scalane w jeden zbiór i zwracane jako wynik. Implementacja iteracyjna zamiast rekurencyjnej zmniejsza liczbę generowanych nagłówków tymczasowych.

Trzecia metoda wywołuje metodę drugą tworząc listę nagłówków do sprawdzenia. Następnie *po kolei* szuka każdego z nagłówków znajdującego się na liście w indeksie form terminów przyjętych. Jeśli nagłówek zostanie rozpoznany, to uznawany jest za termin szerszy. Jeśli nagłówek zostanie rozpoznany niejednoznacznie, to wyświetlane jest ostrzeżenie (kilkadziesiąt przypadków – przeważnie błędy w słowniku KABA). Natomiast jeśli nagłówek nie zostanie rozpoznany to:

- jeśli nagłówek składa się z kilku określników lub określnika i tematu, to ostrzeżenie nie jest generowane (nie ma wymogu istnienia wszystkich podciągów leksemów nagłówka w słowniku KABA),
- jeśli nagłówek jest pojedynczym określnikiem związanym lub chronologicznym, to ostrzeżenie nie jest generowane (prawie wszystkie określniki chronologiczne są związane, ale informacja o tym nie jest umieszczana w słowniku),
- jeśli nagłówek jest nadrzędną nazwą korporacyjną, to ostrzeżenie nie jest generowane (nie ma wymogu dodawania do słownika KABA nadrzędnego członu dla znajdującej się w słowniku wielocłonowej nazwy korporacyjnej),
- w przeciwnym wypadku (pojedynczy określnik niechronologiczny albo też temat pojedynczy lub złożony z kilku elementów) generowane jest ostrzeżenie, gdyż leksemy takie powinny być rozpoznane, a ich nierozpoznanie może być wynikiem błędów w słowniku KABA.

Dodatkowo w przypadku gdy nagłówek zostanie rozpoznany, korzystając z metody `getAllSubheadings` usuwamy z listy nagłówków do sprawdzenia nagłówki będące

podciągami leksemów rozpoznanego nagłówka. Postępujemy w ten sposób, aby nie wprowadzać nadmiarowych hiperonimii, które wynikałyby z hiperonimii już dodanych.

W przedstawionym przykładzie metoda stwierdza, że terminami szerszymi hasła „Żydzi – Bawaria (Niemcy) – historia – źródła” są trzy hasła:

Żydzi – historia

Bawaria (Niemcy) – historia (hasło zostało znalezione poprzez jego formę określnikową „– Bawaria (Niemcy) – historia”)

– historia – źródła (hasło określnikowe złożone z dwóch określników)

Dlaczego terminów szerszych szuka się także pośród hasel rozwiniętych, a nie tylko pośród hasel reprezentujących pojedyncze leksemy? Dzieje się tak dlatego, by z hasła reprezentującego historię Bawarii czyli: „Bawaria (Niemcy) – historia” można było przejść do terminu węższego: „Żydzi – Bawaria (Niemcy) – historia – źródła”. Hiperonimia ta nie zostałaaby dodana, gdyby terminów szerszych nie szukało się pośród hasel rozwiniętych.

Należy zwrócić uwagę, że terminów szerszych szukamy w indeksie form terminów przyjętych. Użyteczne to jest w przypadku, gdy w hasle rozwiniętym zostanie użyty leksem bez dopowiedzenia, który pochodzi od terminu przyjętego mającego dopowiedzenie.

6.2.2 Możliwe udoskonalenia

Terminy odrzucone. Analizując terminy odrzucone na pewno otrzymałoby się bardzo dużo dodatkowych hiperonimii. Jako termin odrzucony do słownika dodawany jest bardzo często nagłówek rozwinięty wykorzystujący alternatywny zasób słownictwa niż to użyte do utworzenia terminu przyjętego. Oznacza to, że budując hiperonimie na podstawie tego słownictwa, można by było znaleźć całkiem nowe hiperonimy. Aktualnie nie analizuje się terminów odrzuconych, gdyż wymaga to przeprowadzenia testów na znacznej liczbie nagłówków. Terminy odrzucone mogą posiadać nieznanne odstępstwa od języka hasel przedmiotowych, które należałoby wykryć, aby dobrze analizować terminy odrzucone.

Szukanie pośród innych typów. W słowniku KABA istnieje dużo nazw korporacyjnych podrzędnych wobec Kościoła katolickiego. Przykładem są złożone nazwy korporacyjne poszczególnych dekanatów na przykład: „Kościół Katolicki. Dekanat Radomski (corporate)”. Niestety nie zostaną one powiązane z hiperonimem „Kościół Katolicki”, gdyż to ostatnie hasło nie jest nazwą korporacyjną tylko nazwą pospolitą. Problem ten należy przeanalizować dokładniej i podjąć odpowiednie kroki w celu utworzenia takich hiperonimii.

6.2.3 Analiza wyników

W słowniku znajduje się łącznie 112671 haseł. Analizowano jedynie ich terminy przyjęte. 84759 haseł ma terminy przyjęte składające się tylko z jednego leksemu (to znaczy prostego tematu lub pojedynczego określnika). Pozostałe 27912 hasła to hasła złożone z kilku leksemów – złożonego tematu (na przykład nazwy aktora i tytułu), kilku określników lub tematu i określników. Poniżej zostaną omówione wyniki rozpoznawania leksemów tych haseł.

Nierozpoznane leksemy

Leksemy będące określnikami oznaczonymi jako związane, określnikami chronologicznymi oraz nadrzędnymi nazwami korporacyjnymi nie muszą być zdefiniowane w słowniku i tym samym ich nierozpoznanie jest dopuszczalne. W analizowanych hasłach znajduje się: około 2 tysięcy nierozpoznanych określników zdefiniowanych jako związane, około 3 tysięcy nierozpoznanych określników chronologicznych oraz około 500 nierozpoznanych nadrzędnych nazw korporacyjnych. Należy jednak zaznaczyć, że przynajmniej część nierozpoznanych nadrzędnych nazw korporacyjnych wynika z błędów w słowniku KABA. Przykładowo w nagłówku „Polska. Sejm (1669)” pierwszy leksem powinien być oznaczony jako geograficzny a nie jest. Z tego powodu nie jest możliwe rozpoznanie nazwy geograficznej „Polska”. Taka sama sytuacja występuje dla hasła „Lwów (Polska). Magistrat”, natomiast analogiczna dla omówionych już korporacyjnych haseł złożonych dotyczących Kościoła katolickiego.

Pozostałe leksemy powinny zostać rozpoznane, jednak 3 tysiące z nich nie zostało rozpoznanych. Poniżej zostaną omówione poszczególne typy leksemów i powody ich nierozpoznania.

Najczęściej nierozpoznanyimi leksemami są określniki rzeczowe. Około 2 tysiące z nich nie zostało rozpoznanych. Większość z nich to określniki związane, które nie są opisane jako związane. Jednak wydaje się, że część z nich nie została rozpoznana z powodu błędów w słowniku KABA.

Jeśli chodzi o określniki formy, to trudno powiedzieć z jakiego powodu nie zostały rozpoznane, jednak takich sytuacji jest tylko kilkanaście. Leksemy tytułowe także nie będą omawiane, bo tylko dwa z nich nie zostały rozpoznane.

Leksemy pozostałych typów zostały nierozpoznane przeważnie z powodu błędów znajdujących się w słowniku KABA. Błędami tymi najczęściej są literówki oraz niepoprawnie podany typ hasła (na przykład oznaczenie nazwy pospolitej jako geograficznej). Dla pewnej niewielkiej liczby haseł trudno powiedzieć czy generowane ostrzeżenie wynika z błędu w słowniku, czy też z wyjątku od języka haseł przedmiotowych. Przykładowo dla leksemów będących tematami pospolitymi $\frac{1}{3}$ nierozpoznań wynika z niepoprawnie podanego typu hasła. Dużo jest też haseł rozwiniętych o te-

macie nie dodanym do słownika! Natomiast dla leksemów będących określnikami geograficznymi w słowniku często istnieje nazwa geograficzna o tym samym tekście. Jednak określnika nie można z nią powiązać, bo nazwa ta nie ma oznaczenia pełnienia podwójnej funkcji (tematu i określnika)!

Do nierozpoznanych elementów nagłówków zaliczają się także rozpoznane niejednoznacznie „podciągi” nagłówków. Część z nich także wynika z błędów w słowniku KABA.

Możliwe, że nierozpoznane określniki rzeczowe można by w niektórych przypadkach wiązać do nazw pospolitych o tym samym tekście. Jednak taki temat powinien być oznaczony jako dwufunkcyjny!

Rozpoznane leksemy

Znaczna część leksemów została rozpoznana. Nie znaleziono żadnego terminu szerszego jedynie dla 596 haseł złożonych z więcej niż jednego leksemu. Dla pozostałych haseł złożonych znaleziono przynajmniej jeden termin szerszy, a średnio 1,58 terminów.

Prawie wszystkie utworzone hiperonimie są poprawne. Poprawności kilku hiperonimii autor nie mógł określić. Na przykład hasło „Meteorologia – tabele, wykresy”, w którym występuje określnik rzeczowy, zostało uznane za hiponim określnika formy „– tabele, wykresy (form subdiv)”. Autor nie wie czy hiperonimia ta jest semantycznie poprawna ze względu na zastosowaną zmianę typu, chociaż sama zmiana typu określnika formy na określnik rzeczowy jest dopuszczalna w języku haseł przedmiotowych. Takie szczegółowe analizy semantycznej poprawności w przypadku określników formy (a więc określających pojęcia stosowane tylko w dziedzinie bibliotekarstwa) można by przeprowadzić razem z Centrum NUKAT.

Wydaje się, że jeśli chodzi o implementację wiązania haseł na podstawie leksemów zrobiono wszystko co można było zrobić bez szczegółowej analizy wyników z Centrum NUKAT.

6.2.4 Przykłady powiązań

Poniżej podano hiperonimy znalezione na podstawie omawianej reguły dla trzech przykładowych haseł. Znak \rightarrow oznacza relację hiperonimii, a kierunek strzałki wskazuje na hiperonim. W nawiasach kwadratowych podany jest zbiór wszystkich hiperonimów oddzielonych przecinkami.

- „Polska. Polskie Siły Zbrojne na Zachodzie – historia (corporate)” \rightarrow [Polska. Polskie Siły Zbrojne na Zachodzie (corporate), Polska – historia (geographic)]
Zostały znalezione dwa terminy szersze. Drugi jest najbardziej interesujący.

Dzięki niemu wiemy, że terminem szerszym „historii polskich sił zbrojnych na zachodzie” jest „historia Polski”.

- „Absurd – w literaturze” → [Absurd (filozofia)]

Dzięki temu, że hasło „Absurd – w literaturze” posiada wariant: „Absurd (filozofia) – w literaturze” wiemy, że chodzi o pojęcie absurdu w sensie filozoficznym, ale nie wykorzystujemy tej wiedzy w algorytmie. Hiperonimia została znaleziona dzięki temu, że w indeksie form istniała forma bez dopowiedzeń hasła „Absurd (filozofia)”.

- „Japonia – cywilizacja – 1185-1333 (geographic)” → [Japonia – cywilizacja (geographic), Japonia – 1185-1333 (Okres Kamakura) (geographic)]

Znaleziono dwa hasła powstałe poprzez usunięcie pojedynczego leksemu z hasła oryginalnego. Drugi hiperonim został znaleziony dzięki temu, że w indeksie form istniała forma bez dopowiedzeń hasła „Japonia – 1185-1333 (Okres Kamakura)”.

6.3 Wiązanie haseł z ich dopowiedzeniami

W punkcie tym zostanie przedstawiona metoda tworzenia relacji hierarchicznych z terminami szerszymi na podstawie dopowiedzeń haseł wzorcowych, a także wyniki jej działania. Dopowiedzenia zostały omówione podczas omawiania języka haseł przedmiotowych.

6.3.1 Opis działania

Dopowiedzenia czasowe obecnie nie są wykorzystywane. Pozostałe dopowiedzenia znajdujące się w rozpatrywanym hasle wzorcowym są rozpoznawane, to znaczy określane jest hasło wzorcowe przez nie reprezentowane. Rozpoznane hasła wzorcowe są następnie łączone relacjami hierarchicznymi z hasłem aktualnym.

Hierarchiczne dopowiedzenia geograficzne są przed rozpoznaniem skompaktowane w sposób opisany w punkcie 5.4.3. Skompaktowany lub oryginalny tekst dopowiedzenia geograficznego jest następnie szukany w indeksie wszystkich form nazw geograficznych, to znaczy zarówno w ich terminach przyjętych jak i odrzuconych, a także w terminach z usuniętymi dopowiedzeniami. Jeśli skompaktowane dopowiedzenie geograficzne nie zostanie znalezione w indeksie, to szukane jest także to dopowiedzenie bez jego własnych dopowiedzeń powstałych podczas kompaktowania.

Rozpoznając dopowiedzenia kwalifikujące szukamy ich w indeksie wszystkich form nazw pospolitych w sposób opisany w punkcie 5.4.4.

Aby poprawnie rozpoznawać wszystkie dopowiedzenia musimy się jeszcze zastanowić w jaki sposób określić czy dane dopowiedzenie jest dopowiedzeniem geograficznym czy kwalifikującym. Jeśli leksem jest geograficzny, a tekst dopowiedzenia

zaczyna się dużą literą, to na początku poszukiwana jest nazwa geograficzna, w przeciwnym wypadku – nazwa pospolita. Jeśli dopowiedzenie nie zostanie rozpoznane, to szukamy nazwy drugiego typu.

Przykładowe hasło podane podczas omawiania dopowiedzeń:

Word Peace Congress (1949 ; Paryż, Francja / Praga, Czechosłowacja ; kongres) powinno zostać powiązane z następującymi trzema terminami szerszymi:

[Paryż (Francja) (geographic), Praga (Czechy) (geographic), Kongresy].

Dopowiedzenie „Paryż, Francja” na początku próbujemy rozpoznać jako nazwę pospolitą. Nie ma jednak takiej nazwy pospolitej, więc po skompaktowaniu do postaci „Paryż (Francja)” szukamy jej pośród nazw geograficznych. Nazwa ta zostanie rozpoznana, a więc uznajemy ją jako termin szerszy hasła korporatywnego. Dopowiedzenie „Praga, Czechosłowacja” także nie zostanie rozpoznane jako nazwa pospolita, a dopowiedzenie „Praga (Czechosłowacja)” nie zostanie rozpoznane jako nazwa geograficzna. Dopiero dopowiedzenie „Praga” zostanie rozpoznane jako forma bez dopowiedzenia hasła „Praga (Czechy)”. A więc terminem szerszym hasła korporatywnego jest także hasło „Praga (Czechy)”¹. Dopowiedzenie „kongres” zostanie po zmianie liczby na mnogą rozpoznane jako hasło „Kongresy”.

Jeśli dla pewnego kilkuleksmowego hasła z dopowiedzeniami termin szerszy powstały na podstawie leksemów zawierałby dopowiedzenie występujące w pierwotnym hasle, to podczas wiązania na podstawie dopowiedzeń dopowiedzenie to nie będzie brane pod uwagę. Na przykład hasło „Aglutynacja (immunologia) – testy” posiada między innymi termin szerszy „Aglutynacja (immunologia)” powstały poprzez wydzielenie tematu z hasła rozwiniętego i zawierający dopowiedzenie „immunologia”. Wynika z tego, że hasła „Aglutynacja (immunologia) – testy” nie powiążemy bezpośrednio z hasłem „Immunologia”. Postępujemy tak, aby nie wprowadzać nadmiarowych relacji. Z powyższego powodu terminy szersze określane na podstawie leksemów powinny być określone przed terminami szerszymi określanymi na podstawie dopowiedzeń.

Jeśli dla pewnego hasła jego dopowiedzenie prowadziło do jego samego (poprzez termin odrzucony), to z oczywistych powodów nie dodajemy takiej relacji. Na przykład w hasle „Żydzi – 598-515 a.C (Niewola Babilońska)” dopowiedzenie „Niewola Babilońska” związanego określnika chronologicznego wskazuje na to samo hasło poprzez jego termin odrzucony „Niewola Babilońska”.

Jeśli dopowiedzenie nie zostanie rozpoznane lub zostanie rozpoznane niejednoznacznie, to generowane jest ostrzeżenie lub informacja o tym. Komunikat „Info” jest generowany dla dopowiedzeń chronologicznych lub dopowiedzeń określników chronologicznych, które nie zostały w ogóle rozpoznane. Dopowiedzenia chronologiczne

¹W rzeczywistości dopowiedzenie „Praga” zostanie rozpoznane niejednoznacznie z poniżej opisanego powodu, a więc nie będziemy mogli powiązać tego dopowiedzenia z żadnym hasłem.

najczęściej nie posiadają haseł w słowniku KABA. W pozostałych przypadkach generowany jest komunikat „Ostrz.” informujący o potencjalnym błędzie w zawartości słownika KABA.

6.3.2 Możliwe udoskonalenia

Inne typy nagłówków oraz terminy odrzucone. Regułę gruntownie przetestowano jedynie dla terminów przyjętych haseł będących nazwami pospolitymi oraz geograficznymi i obecnie tylko na podstawie tych nagłówków tworzone są relacje hierarchiczne. Dla pozostałych nagłówków należy przetestować działanie reguły, aby ewentualnie dostosować jej działanie, w tym wiązanie z terminami szerszymi oraz generowanie ostrzeżeń.

Odmienianie przez liczbę także wyrażen rzeczownikowych. Niektóre dopowiedzenia są wyrażeniami rzeczownikowymi na przykład rzeczownikiem z określającym go przymiotnikiem. Niestety z opisanego wcześniej powodu w słowniku dość często istnieje jedynie hasło w przeciwnej liczbie gramatycznej. Dzieje się tak na przykład dla hasła „Żiguli WAZ-2101 (samochód osobowy)” – w słowniku istnieje jedynie hasło „Samochody osobowe”. Z uwagi na częstość problemu opłacałoby się szukać haseł w liczbie przeciwnej także dla wyrażen rzeczownikowych. W tym celu należałoby odmieniać przez liczbę nie tylko rzeczowniki ale także przymiotniki.

Odmienianie przez liczbę metodą niesłownikową. Aktualnie formę dopowiedzenia w przeciwnej liczbie gramatycznej można uzyskać tylko dla haseł znajdujących się w słowniku języka polskiego. Jednak w słowniku KABA istnieją także hasła nieobecne w używanym słowniku języka polskiego. Na przykład nieobecne jest specjalistyczne hasło „mikroprocesor”. Jeśli chcielibyśmy rozpoznać takie mniej powszechne dopowiedzenia, to moglibyśmy spróbować odmienić je przez liczbę metodą niesłownikową, czyli na podstawie reguł odmiany rzeczowników przez liczbę.

Niejednoznaczność rozpoznania miast. Dopowiedzenia określające nazwę miasta często rozpoznawane są niejednoznacznie. Na przykład w hasle „Jesuitska kolej Klementinum (Praga) (corporate)” dopowiedzenie „Praga” zostanie rozpoznane niejednoznacznie jako miasto „Praga (Czechy)” lub region geograficzny „Praga (Czechy ; region)”. W takim przypadku nie będziemy mogli powiązać hasła korporatywnego z żadnym hasłem. Jednym z rozwiązań mogłoby być wiązanie hasła z hasłem określającym miasto, gdyż wydaje się, że używając w dopowiedzeniu takiej niejednoznacznej nazwy prawie zawsze ma się na myśli nazwę miasta. Co więcej właśnie miasta a nie regiony są lokalizatorami dla nazw geograficznych.

Dopowiedzenia geograficzno-kwalifikujące. Jeśli w danym hasle występuje zarówno dopowiedzenie lokalizujące geograficzne jak i dopowiedzenie kwalifikujące, to nie powinno się wiązać tego hasła do obu haseł osobno, ale do hasła rozwiniętego złożonego z tych dopowiedzeń. Na przykład hasło „Adirondack (Stany Zjednoczone ; góry)” nie powinno być powiązane do haseł „Stany Zjednoczone” i „Góry”, ale do hasła „Góry – Stany Zjednoczone”. Sytuacja ta powtarza się dość często z tego powodu, że prawie wszystkie leksemy geograficzne posiadają zarówno dopowiedzenie geograficzne jak i kwalifikujące. Dopowiedzeń kwalifikujących nie posiadają jedynie miasta. Innymi przykładami poprawnych powiązań byłyby:

Aare (Szwajcaria ; rzeka) → Rzeki – Szwajcaria

Ain (Francja ; departament) → Francja – departamenty

Alabama (Stany Zjednoczone ; stan) → Stany Zjednoczone – stany

6.3.3 Analiza wyników

Jak już wcześniej wspomniano, przeanalizowano jedynie terminy przyjęte nazw pospolitych i geograficznych – łącznie około 67 tysięcy nagłówków. Każde z dopowiedzeń zawarte w tych nagłówkach zostało rozpoznane poprawnie, błędnie, albo nie zostało rozpoznane. Poniżej podano powody nierozpoznania lub błędnych rozpoznań oraz opis poprawnych powiązań.

Dopowiedzenia nie rozpoznane

Niektóre dopowiedzenia nie są reprezentowane przez żadne hasło wzorcowe. Do takich dopowiedzeń należą dopowiedzenia chronologiczne, będące datą lub zakresem dat oraz większość dopowiedzeń określników chronologicznych, będących przeważnie słownymi nazwami zakresu dat, na przykład nazwą epoki lub okresu panowania władcy. Przykładem jest hasło: „Żydzi – 598-515 a.C (Niewola Babilońska)”. Takie dopowiedzenia oczywiście nie mogą być rozpoznane i dlatego nie są dla nich generowane ostrzeżenia o nierozpoznaniu. Dla pozostałych dopowiedzeń które nie zostaną rozpoznane generowane są ostrzeżenia. Łącznie zostało wygenerowanych około 2000 takich ostrzeżeń, z czego znaczna część powtarzała się z powodu powtarzania się tych samych dopowiedzeń w różnych hasłach.

Poniżej przedstawiono powody nierozpoznania dopowiedzeń. Przyczyną może być zarówno błąd w słowniku KABA; niedoskonałości zastosowanego algorytmu, które można naprawić w opisany wcześniej sposób; albo też sytuacje niemożliwe do naprawienia ze względu na za małą liczbę informacji zawartą w słowniku.

- a) dopowiedzenia rozpoznane niejednoznacznie, to jest wskazujące na więcej niż jedno hasło wzorcowe

- i) Część takich sytuacji wynika z istnienia w słowniku KABA haseł różniących się tylko liczbą gramatyczną. Na przykład istnieją dwa hasła: „Statystyka” i „Statystyki”. Pierwsze hasło oznacza naukę o statystyce, drugie oznacza rzeczywiste dane statystyczne, na przykład o przyroście naturalnym Polski. W hasle „Zmienne instrumentalne (statystyka)” dopowiedzenie ma pierwsze znaczenie, jednak ze względu na dowolność stosowania liczby gramatycznej w dopowiedzeniach nie wiemy czy powiązać je z hasłem w liczbie pojedynczej czy w liczbie mnogiej. Sytuacja ta jest trudna lub nawet niemożliwa do naprawienia.
- ii) Innym powodem jest istnienie w słowniku dwóch różnych terminów przyjętych mających ten sam termin odrzucony. Na przykład w hasle „Żydzi – 66-73 (Powstanie)” dopowiedzenie może być rozpoznane jako „Rewolucje” lub „Rewolty”, gdyż oba te hasła mają ten sam wariant „Powstanie”. Często się zdarza, że hasła należące do niejednoznacznego rozpoznania są w relacji hierarchicznej. Na przykład w tym przypadku hasło „Rewolty” jest terminem węższym hasła „Rewolucje”. Częściowym rozwiązaniem w takim przypadku mogłoby być powiązanie dopowiedzenia z terminem szerszym.
- b) dopowiedzenia nie rozpoznane wcale
- i) z powodu błędnej ich odmiany przez liczbę
- nie odmienianie przez liczbę wyrażen rzeczownikowych
 - odmienianie przez liczbę tylko rzeczowników obecnych w używanym słowniku języka polskiego
- ii) z powodu braku w słowniku hasła przez nie reprezentowanego
- brak haseł dla niektórych powiatów. Prawdopodobnie nie ma żadnej książki o tym powiecie, ale mimo to istnieje na przykład hasło dla miasta, które posiada ten powiat jako dopowiedzenie.
 - brak innych haseł, na przykład hasła dla dopowiedzenia w hasle „Źródło D (biblistyka)”
 - literówki w dopowiedzeniach, na przykład w hasle „Zniesławienie (prawo kanoniczne)”
- iii) z tego powodu, że są wyrażeniami zdania naturalnego na przykład: „lud afrykański”, „rodzina mikroprocesorów”, „rasa świń”
- część z nich to dopowiedzenia geograficzne kwalifikujące, które nie mają własnego hasła na przykład: „rejon autonomiczny”
- iv) z tego powodu, że tekst w nawiasach w rzeczywistości nie jest dopowiedzeniem w sensie języka KABA na przykład: „Pismo – studia i nauczanie (podstawowe)”

Widać więc, że dopowiedzenia mogą być nierozpoznane albo z powodu błędnej odmiany przez liczbę albo z powodu użycia niedostatecznie sformalizowanego

wyrażenia, którego komputer nie jest w stanie zrozumieć. Możliwe też, że zamiast używać w dopowiedzeniach haseł nieobecnych w słowniku KABA można by użyć synonimicznych haseł obecnych w słowniku.

Dopowiedzenia błędnie rozpoznane

Czasem zdarza się, że dopowiedzenie zostaje błędnie rozpoznane, ponieważ w słowniku istnieje hasło o tym samym tekście co dopowiedzenie, ale z innej, błędnej dziedziny. Przykładowo w hasle „Żywiec (rodzaj)” rodzaj oznacza rodzaj biologiczny rośliny. Jednak dopowiedzenie to zostanie rozpoznane jako rodzaj gramatyczny: „Rodzaj (językoznawstwo)”, gdyż nie wprowadzono do słownika KABA hasła „Rodzaj (taksonomia)”. Podobnie w hasle „Indianie – Alberta (Kanada ; prowincja)” dopowiedzenie „prowincja” oznacza typ jednostki podziału terytorialnego Kanady. Jednak dopowiedzenie to zostanie powiązane z hasłem „Prowincja” mającym znaczenie „część kraju, nieduże miasto, wieś itp. oddalone od stolicy”. Błędów tych można by uniknąć, gdyby w dopowiedzeniach używało się zawsze tylko terminów przyjętych albo odrzuconych obecnych w słowniku KABA.

Dopowiedzenia rozpoznane poprawnie

Rozpoznano przeszło 16 tysięcy dopowiedzeń, z czego znaczną większość poprawnie. Relacje powstałe na podstawie poprawnie rozpoznanych dopowiedzeń często pokrywają się z relacjami jawnymi. Jednak najczęściej są relacjami, których nie można byłoby uzyskać w inny sposób. Do takich relacji należy między innymi dużo relacji do instancji poszczególnych pojęć, na przykład do nazw poszczególnych metali.

Dla niektórych haseł istnieje bardzo dużo relacji do terminów węższych, które powstały z dopowiedzeń. Może to powodować wydłużenie listy terminów węższych do bardzo dużych rozmiarów. Jest tak na przykład dla hasła „Polska”. Na liście terminów węższych tego hasła znajduje się bardzo dużo miast nie powiązanych z żadną jednostką podziału terytorialnego Polski. Dzieje się tak prawdopodobnie z tego powodu, że nie ma obowiązku podawania jednostki podziału terytorialnego w dopowiedzeniach nazwy geograficznej, gdy nazwa ta jest jednoznaczna na terenie kraju. Długa lista może stanowić problem podczas jej przeglądania przez czytelnika.

6.3.4 Przykłady powiązań

- „Siuksowie (Indianie)” → „Indianie”

Dopowiedzenie kwalifikujące zostało w prosty sposób rozpoznane. Analogiczne wiązania istnieją dla pozostałych plemion indiańskich, żadne z nich nie było powiązane jawnie.

- „Żargon (terminologia)” → „Terminologia (nauka)”
Dopowiedzenie kwalifikujące zostało rozpoznane jako forma bez dopowiedzenia hasła „Terminologia (nauka)”.
- „Ziarnko gorzycy (przypowieść)” → „Przypowieści”
Dopowiedzenie kwalifikujące po zmianie liczby gramatycznej na mnogą zostało rozpoznane jako hasło „Przypowieści”.
- „Popowice (Polska, województwo świętokrzyskie)” (geographic) → „Świętokrzyskie, Województwo (Polska ; 1999-)” (geographic)
Na początku hierarchiczne dopowiedzenie geograficzne „Polska, województwo świętokrzyskie” zostało skompaktowane do formy: „województwo świętokrzyskie (Polska)”. Forma ta została rozpoznana jako termin odrzucony „Województwo świętokrzyskie (Polska)” hasła „Świętokrzyskie, Województwo (Polska ; 1999-)”.

6.4 Wiązanie nazw zależnościowych

W punkcie tym zostanie przedstawiona metoda budowania relacji hierarchicznych z terminami szerszymi na podstawie nazw zależnościowych, a także wyniki jej działania.

6.4.1 Opis działania

Jak napisano w rozdziale czwartym, analizując nazwy zależnościowe wystarczy ograniczyć się do nierozwiniętych nazw pospolitych. Nierozwinięte nazwy pospolite mają prostą budowę, gdyż są zawsze jednoleksmowe, choć leksemy te mogą zawierać dopowiedzenia. Dla prostoty aktualnie zaimplementowano jedynie budowanie hiperonimii dla haseł bez dopowiedzeń.

Tekst nazwy zależnościowej dzielony jest na łańcuchach znaków „ i ”. Jeśli jest mniej lub więcej niż dwie części podziału, to nagłówek nie jest nazwą zależnościową. Natomiast jeśli tekst nagłówka został podzielony na dokładnie dwie części, to obie części próbuje się rozpoznać wśród terminów przyjętych nazw pospolitych (ewentualnie zmieniając liczbę gramatyczną na przeciwną). Operacja rozpoznania przebiega tak samo jak rozpoznawanie dopowiedzeń kwalifikujących, z tym że przeszukiwane są jedynie terminy przyjęte haseł bez usuwania ich dopowiedzeń, a nie wszystkie formy haseł wzorcowych. Jeśli tylko jedna część nazwy zależnościowej została rozpoznana, to dla nierozpoznanej części generowane jest ostrzeżenie o nierozpoznaniu. Jeśli obie części nie zostały rozpoznane, to uznaje się, że nazwa jest nazwą zbiorową i ostrzeżenia o nierozpoznaniu nie są generowane.

6.4.2 Możliwe udoskonalenia

Inne typy nagłówków oraz terminy odrzucone. Możliwe, że niektóre określenia rzeczowe także są nazwami złożonymi. Jeśli tak to można by je także wiązać z ich składowymi.

Prawdopodobnie nie będzie opłacalne analizowanie terminów odrzuconych, gdyż większość nazw zależnościowych posiada warianty zbudowane w ten sam sposób, tyle że z odwrotną kolejnością składników. Na przykład hasło „Ojciec i dziecko” posiada wariant „Dziecko i ojciec”. Choć możliwe jest, że istnieją zależnościowe terminy odrzucone o innej budowie.

Nazwy zależnościowe z dopowiedzeniami oraz szukanie wśród nazw z dopowiedzeniami. Istnieje około dwudziestu nazw zależnościowych z dopowiedzeniami. Aktualnie nie są one analizowane, ale je także można by było łączyć z ich składowymi.

Aktualnie składowa nie zostanie rozpoznana, jeśli oryginalne hasło ją reprezentujące posiada dopowiedzenie. Możliwe, że podczas rozpoznawania składowych powinniśmy ich szukać także wśród haseł bez dopowiedzeń.

Odmienianie przez liczbę. Możliwe, że szukanie haseł także w przeciwnej liczbie gramatycznej przynosi więcej szkody niż pożytku. Szukanie w przeciwnej liczbie gramatycznej może powodować rozpoznanie niejednoznaczne. Jeśli w praktyce składowe miałyby prawie zawsze taką samą liczbę gramatyczną co pojedyncze hasła je reprezentujące, to lepiej nie szukać formy w przeciwnej liczbie gramatycznej, aby nie powodować niejednoznaczności. Ewentualnie można by szukać formy liczby przeciwnej tylko wtedy, gdy forma hasła w tej samej liczbie nie została odnaleziona.

6.4.3 Analiza wyników

W słowniku KABA znajduje się około 1200 nazw zależnościowych i zbiorowych. Dla tych nazw zostało wygenerowanych około 300 ostrzeżeń, a więc prawdopodobnie tyle składowych nazw zależnościowych nie zostało rozpoznanych. Błędne rozpoznania zdarzają się bardzo rzadko. Rozpoznanie niejednoznaczne często wynika z szukania oryginalnego hasła także w liczbie przeciwnej. Jak już wcześniej wspomniano, należy zastanowić się czy szukanie także formy liczby przeciwnej jest opłacalne i dopiero wtedy można przeprowadzić dalszą analizę innych powodów nierozpoznania składowych. Około 1200 składowych zostało poprawnie rozpoznanych. Często relacje powstałe na ich podstawie nie mogłyby być utworzone w inny sposób.

6.5 Wiązanie hasła z hasłami zaczynającymi się nim

Tematy niektórych haseł składają się z więcej niż jednego wyrazu. Ze względu na sformalizowanie słownictwa oraz dogodność wyszukiwania w indeksie alfabetycznym część haseł ma następującą budowę:

<hasło podstawowe> <jedno- lub kilkuwyrazowy tekst wyszczególniający>

Oznacza to, że w słowniku KABA istnieje inne hasło, którym rozpatrywane hasło się zaczyna. Poniżej podano przykładową nazwę pospolitą, geograficzną i określnik formy oraz hasła zaczynające się tymi nazwami:

Aerodynamika

Aerodynamika konstrukcji

Aerodynamika przepływów naddźwiękowych

Aerodynamika przepływów hipersonicznych

Aerodynamika przepływów przydźwiękowych

Afryka (geographic)

Afryka anglojęzyczna (geographic)

Afryka francuskojęzyczna (geographic)

Afryka luzofońska (geographic)

Afryka Czarna (geographic)

Afryka Północna (geographic)

Afryka Północno-Wschodnia (geographic)

Afryka Wschodnia (geographic)

Afryka Wschodnia anglojęzyczna (geographic)

Afryka Wschodnia niemiecka (geographic)

Afryka Południowa (geographic)

Afryka Zachodnia (geographic)

Afryka Zachodnia francuskojęzyczna (geographic)

Afryka Środkowa (geographic)

- adaptacje (form subdiv)
- adaptacje filmowe i telewizyjne (form subdiv)
- adaptacje muzyczne (form subdiv)
- adaptacje radiowe (form subdiv)

Jeśli dla nazwy pospolitej, geograficznej lub określnika formy znajdziemy hasło podstawowe którym się ona zaczyna, to prawie zawsze hasło podstawowe można uznać za termin szerszy hasła złożonego. Z oczywistych względów hasła te powinny mieć ten sam typ. Z przeprowadzonej analizy wynika, że powyższa zasada przeważnie

nie jest spełniona dla określników rzeczowych.

Jeśli pewne hasło podstawowe ma niewiele takich terminów węższych, to relacje do nich zapisywane są w słowniku w sposób jawny, choć mimo to wielu takich wpisów brakuje. Jeśli takich terminów węższych jest więcej, to w hasle podstawowym umieszczany jest odsyłacz orientacyjny uzupełniający wskazujący na te hasła, choć znowu wielu takich odsyłaczy brakuje. Z drugiej strony niektóre hasła wskazywane przez taki odsyłacz nie muszą być terminami węższymi. Oznacza to, że informacja zawarta w odsyłaczu nie jest ani informacją pewną ani pełną. Dlatego nie opłaca się opierać na informacji zawartej w odsyłaczach. O wiele pewniejszym i prostszym sposobem utworzenia hiperonimii jest przeanalizowanie tekstów haseł i znalezienie tych zaczynających się wyrazami hasła podstawowego.

6.5.1 Odmiana przez liczbę hasła podstawowego w hasle złożonym

Dla niektórych haseł podstawowych istnieją hasła złożone, w których hasło podstawowe występuje w przeciwnej liczbie gramatycznej. Dzieje się tak wtedy, gdy znaczenie hasła złożonego lepiej jest określone przez inną liczbę gramatyczną. Przykładowo hasło podstawowe „Aloesy” jest w liczbie mnogiej, gdyż jest wiele gatunków aloesów; natomiast hasło złożone „Aloes zwyczajny” jest w liczbie pojedynczej, gdyż określa pojedynczy gatunek. Z podobnego powodu różnią się liczbą hasła „Azotany” i „Azotan amonu”.

Przeważnie dla ustalonego hasła podstawowego jest mało haseł złożonych w innej liczbie i są one wiązane jawnie z hasłem podstawowym. Jak nietrudno się domyślić, wiązanie jawne jest w tym przypadku szczególnie ważne podczas wyszukiwania haseł w indeksie alfabetycznym, a także podczas automatycznego tworzenia relacji hierarchicznych, gdyż jak później zobaczymy automatyczna zmiana liczby jest złym rozwiązaniem. Jeśli haseł złożonych w innej liczbie jest więcej, to dodawany jest odsyłacz orientacyjny uzupełniający. Przykładowe postacie takich odsyłaczy to:

- „Zobacz też hasła zaczynające się od wyrazów Bomba/y”
- „Zobacz też hasła zaczynające się od wyrazów Chlorki/Chlorek”
- „Zobacz też hasła zaczynające się od wyrazów Interferon/y”
- „Zobacz też hasła zaczynające się od wyrazów Sport ; Sporty”

Jak widać format tych odsyłaczy nie jest do końca sformalizowany, a więc dopóki nie zostanie sformalizowany, ich komputerowe przetwarzanie byłoby skomplikowane. Stanowi to pewien problem, gdyż jak już wcześniej wspomniano automatyczna zmiana liczby jest złym rozwiązaniem. Oznacza to, że jeśli relacja do hasła złożonego *w innej liczbie* nie jest zapisana jawnie, to obecnie nie zostanie ona wykryta automatycznie.

Dlaczego automatyczna zmiana liczby nie opłaca się

Podczas analizy wyszczególniono dwa przypadki: hasła podstawowe dla których istnieje inne hasło podstawowe o takim samym tekście, tyle że w innej liczbie gramatycznej oraz hasła podstawowe dla których nie istnieją takie hasła.

Hasła podstawowe pierwszego rodzaju powinny być łączone prawie zawsze jedynie z hasłami złożonymi w tej samej liczbie gramatycznej co hasło podstawowe. Połączenie z hasłem w innej liczbie prawdopodobnie byłoby połączeniem błędnym, to znaczy nie byłoby połączeniem termin szerszy – termin węższy.

Dla haseł podstawowych drugiego rodzaju hasła złożone w przeciwnej liczbie powiązane są najczęściej jawnie. Natomiast łączenie automatyczne haseł podstawowych z hasłami złożonymi w przeciwnej liczbie prowadziłyby najczęściej do połączeń błędnych.

Z powyższych faktów wynika, że automatyczna zmiana liczby prowadziłyby częściej do błędnych powiązań niż do poprawnych, co jest całkowicie nieopłacalne.

6.5.2 Niepewność relacji

W przeciwieństwie do poprzednich reguł hasła złożone analizujemy w sposób bardziej wyrazowy niż pojęciowy. W poprzednich regułach rozpoznawaliśmy jednostki nagłówków haseł wzorcowych: leksemy, dopowiedzenia i składowe nazw zależnościowych. W obecnej regule poszliśmy o krok dalej i zaczęliśmy analizować wewnątrz jednostek języka hasła przedmiotowych. Poprzednio rozpoznawaliśmy jednostki, które powinny być terminami przyjętymi lub odrzuconymi zarejestrowanymi w słowniku. Dzięki ich sformalizowaniu, jeśli je rozpoznaliśmy, mieliśmy większą szansę na to, że rozpoznanie było poprawne. Obecnie rozpoznajemy wyrazy, które nie są jednostkami języka haseł przedmiotowych. Ich rozpoznawanie częściej prowadzi do błędnych relacji, choć nadal ze względu na sformalizowanie języka haseł przedmiotowych nie jest to bardzo częste. Przykładowo w słowniku KABA istnieje hasło „Akta prawne” oraz jego termin *węższy* „Akta” powiązany z pierwszym relacją jawną. Jednak posługując się przedstawioną regułą utworzymy relację mówiącą, że hasło „Akta prawne” jest terminem węższym hasła „Akta”. Co więcej, do tego samego wniosku doszlibyśmy posługując się odsyłaczem orientacyjnym uzupełniającym hasła „Akta” wskazującym na wszystkie hasła zaczynające się wyrazem „Akta”. Widać więc, że próba zrozumienia wyrazowego znaczenia leksemów nie zawsze prowadzi do poprawnych wyników. Innymi przykładami par haseł które prowadzą do błędnych relacji są:

- hasło „Drogi” (z dziedziny transportu) oraz hasło „Drogi pokarmowe” (anatomia),
- hasło „Architektura” (z dziedziny budownictwa) oraz hasło „Architektura

komputerów”,

- hasło „Agrest” (owoc, jadalna część rośliny) oraz „Agrest zwyczajny” (nazwa biologiczna gatunku tejże rośliny) – oczywiście cała roślina jest terminem szerszym od jej owoców.

Niektóre z błędnych relacji można by usunąć automatycznie. Na przykład można usunąć sprzeczną relację między hasłem „Akta prawne” i „Akta”, gdyż między tymi hasłami istnieje już jawna relacja skierowana w drugą stronę.

Niestety błędne relacje nie prowadzące do powstania cykli w grafie powiązań haseł nie zostaną wykryte przez komputer i pozostaną w hierarchii. Nie powinno to jednak stanowić większego problemu z dwóch powodów. Po pierwsze błędów takich jest stosunkowo mało. Po drugie w dwóch zastosowaniach: przeglądarce haseł oraz wyszukiwarce książek błędy mogą być w łatwy sposób zauważone przez czytelnika. W przeglądarce wystarczy nie podążać ich tropem, natomiast w wyszukiwarce wystarczy pominąć na liście wyników książki nie na temat. Dodatkowo w obu zastosowaniach czytelnik może przekazać do NUKAT-u informację o tym, że relacja jest błędna lub książka jest nie na temat. Takie zgłoszone błędne relacje można by było wprowadzić na listę wyjątków używaną przez omawianą regułę. Podczas wykorzystywania hierarchii powiązań do obliczania statystyk błędy w hierarchii zaburzałyby wartości obliczonych statystyk w sposób trudny do zauważenia. Jednak ważniejsze błędy prawdopodobnie byłyby już poprawione dzięki interakcji z czytelnikami.

6.5.3 Opis działania

Jak już wcześniej wspomniano analizowane są tylko nazwy pospolite, geograficzne oraz określniki formy. Z tego samego powodu co dla nazw zależnościowych wystarczy ograniczyć się do nagłówków jednoleksemowych. Podobnie jak dla nazw zależnościowych dla prostoty ograniczamy się obecnie do nagłówków bez dopowiedzeń. Dla każdego nagłówka spełniającego powyższe założenia szukamy w indeksie terminów przyjętych haseł zaczynających się tekstem nagłówka podstawowego oraz zawierających po tym tekście spację. Wymagamy spacji, aby znaleźć hasła zaczynające się pewnym wyrazem (lub wyrazami), a nie tylko przedrostkiem wyrazu. Podczas wyszukiwania wykorzystujemy indeks początków wyrazów zwracający wynik w czasie niezależnym od ilości zwróconych haseł. Zwrócone hasła należy przefiltrować, gdyż hasła złożone powinny:

- mieć ten sam typ co hasło podstawowe,
- być jednoleksemowe,
- nie mieć dopowiedzeń.

Dodatkowo nie są zwracane hasła złożone, które można otrzymać przez przechodność omawianej relacji a także te, które można otrzymać z reguły dla nazw

zależnościowych. Wynika z tego, że dla hasła „Afryka” (geographic) otrzymamy następujące, pisane zwykłą czcionką terminy węższe:

- Afryka anglojęzyczna (geographic)
- Afryka francuskojęzyczna (geographic)
- Afryka luzofońska (geographic)
- Afryka Czarna (geographic)
- Afryka Północna (geographic)
- Afryka Północno-Wschodnia (geographic)
- Afryka Wschodnia (geographic)
- Afryka Wschodnia anglojęzyczna (geographic)*
- Afryka Wschodnia niemiecka (geographic)*
- Afryka Południowa (geographic)
- Afryka Zachodnia (geographic)
- Afryka Zachodnia francuskojęzyczna (geographic)*
- Afryka Środkowa (geographic)

gdyż terminy pisane kursywą można otrzymać korzystając z przechodniości relacji. Natomiast dla hasła „Archeologia” otrzymamy następujące terminy węższe:

- Archeologia prehistoryczna
- Archeologia nowożytna
- Archeologia podwodna
- Archeologia ratunkowa

ale nie otrzymamy nazwy zależnościowej „Archeologia i religia”.

Warto dodatkowo zwrócić uwagę, że omawiana reguła nie wykryje takich zależności jak zależność między hasłami „Afryka anglojęzyczna” (geographic) oraz „Afryka Wschodnia anglojęzyczna” (geographic). Jednak na szczęście zależności tego typu prawie zawsze zapisywane są w słowniku jawnie.

6.5.4 Możliwe udoskonalenia

Odsyłacze orientacyjne uzupełniające wskazujące na przeciwną liczbę.

Jak już wcześniej wspomniano, dla niektórych haseł podstawowych istnieją hasła złożone w przeciwnej liczbie gramatycznej, do których nie prowadzi relacja jawna, ale za to prowadzi odsyłacz orientacyjny uzupełniający wskazujący na tą liczbę. Tak jest na przykład dla hasła „Antygeny” zawierającego odsyłacz o postaci „Antygen/y”. Wskazuje on między innymi na hasło „Antygen karcynoembrionalny”.

Terminy odrzucone. Możliwe, że analizując terminy odrzucone otrzymałoby się wiele dodatkowych relacji, ale możliwe też, że wiele z tych relacji byłoby błędnych.

Nazwy złożone z dopowiedzeniami oraz szukanie wśród nazw z dopowiedzeniami. Podobnie jak dla nazw zależnościowych można by analizować także takie hasła.

6.5.5 Analiza wyników

Dzięki opisanej regule otrzymano przeszło 13 tysięcy relacji. Część z nich była już jawnie podana w słowniku KABA. Mimo tego, że reguła nie daje pewnych wyników, to dla 16 sprawdzonych haseł podstawowych nie znaleziono żadnej błędnej relacji prowadzącej do hasła złożonego. Przeglądając powiązania znaleziono jedynie trzy błędne relacje, które podano i omówiono w punkcie 6.5.2.

6.5.6 Przykłady powiązań

- Dla hasła „Aeronautyka” zostały odnalezione następujące terminy węższe:
 - Aeronautyka w leśnictwie
 - Aeronautyka w meteorologii
 - Aeronautyka w rolnictwieprowadzące do różnych zastosowań aeronautyki.
- Dla hasła „Aerodynamika” zostały odnalezione następujące terminy węższe:
 - Aerodynamika przepływów przydźwiękowych
 - Aerodynamika przepływów naddźwiękowych
 - Aerodynamika przepływów hipersonicznychprowadzące do aerodynamiki wielkich prędkości. Należy zwrócić uwagę, że jawna relacja istniała jedynie między hasłami: „Aerodynamika przepływów naddźwiękowych” oraz „Aerodynamika przepływów hipersonicznych”.
- Dla hasła „Adwentyści” został odnaleziony następujący termin węższy:
 - Adwentyści Dnia SiódmegoAdwentyści Dnia Siódmego to największy kościół adwentystyczny.
- Dla hasła „Abrewiacje” zostały odnalezione następujące terminy węższe:
 - Abrewiacje angielskie
 - Abrewiacje francuskie
 - Abrewiacje greckie
 - Abrewiacje łacińskie
 - Abrewiacje niemieckie
 - Abrewiacje polskie
 - Abrewiacje rosyjskieHasło „Abrewiacje” stosuje się do skrótów, a skróty w poszczególnych językach mają osobne hasła.

6.6 Propozycje pozostałych reguł wiązania haseł

Poniżej zostaną przedstawione trzy nowe reguły tworzenia hiperonimii.

6.6.1 Wiązanie określników z tematami

Jeśli dwa hasła – jedno tematowe a drugie określnikowe (rozwinęte lub proste) mają taką samą treść, to można je wiązać ze sobą uważając hasło określnikowe za termin węższy (aspekt) hasła tematowego. Należałoby zachować zgodność typów wiążąc nazwy pospolite tylko z określnikami rzeczowymi, chronologicznymi i formy, a nazwy geograficzne z określnikami geograficznymi. Przy czym wiązanie nazw geograficznych w słowniku KABA nie zachodziłoby ze względu na fakt, że nie ma w nim samodzielnych określników geograficznych.

Przykładami powiązań byłyby:

Oprogramowanie ← – oprogramowanie (general subdiv)

Egzaminy – poradniki dla studentów ← – egzaminy – poradniki dla studentów (form subdiv)

Jeśli książka jest opisana hasłem „Oprogramowanie”, to jest to jeden z jej tematów. Natomiast jeśli jest opisana hasłem rozwiniętym zawierającym określnik „– oprogramowanie”, to oprogramowanie jest tylko aspektem jednego z tematów książki.

Autorowi nie jest znany powód istnienia w słowniku KABA haseł tematowych i określnikowych o tym samym tekście. Wydaje się, że zamiast dwóch oddzielnych haseł można by oznaczać temat jako mogący pełnić funkcję określnika, chyba że definicje obu haseł są nieco odmienne.

6.6.2 Wiązanie części obiektu geograficznego z jego całością

W słowniku KABA przyjęto, że jeśli pewien obiekt geograficzny położony jest na terenie kilku państw, to oprócz hasła reprezentującego cały obiekt można także utworzyć hasła reprezentujące części obiektu leżące na terenie poszczególnych państw. Przykładowo rzeka Ren posiada następujące hasła:

Ren (rzeka)

Ren (Niemcy ; rzeka)

Drugie hasło dotyczy części niemieckiej Renu.

Hasła takie powinny być połączone hiperonimią. Dokładniej mówiąc, można łączyć dwa hasła będące tematowymi leksemami geograficznymi o tym samym tekście i o tym samym dopowiedzeniu kwalifikującym, w przypadku gdy jedno hasło nie ma dopowiedzenia lokalizującego, a drugie ma niehierarchiczne dopowiedzenie lokalizujące.

6.6.3 Zastępowanie leksemu przez leksem szerszy

Wiązanie hasła na podstawie jego leksemów następowało poprzez usuwanie leksemów. Mniej radykalną wersją tej reguły byłoby zastępowanie jednego z leksemów jego hiperonimem zamiast całkowitego jego usuwania. Na przykład w hasle „Nauka – abrewiacje” określnik „– abrewiacje” można by zastąpić jego hiperonimem „– notacja” i dzięki temu powiązać je z hasłem „Nauka – notacja”. Leksemy można zastępować dalej, tworząc kolejne rozwinięte terminy szersze. Ostateczne usunięcie leksemu można traktować jako jego zastąpienie przez leksem „ANY SUBJECT”. Zastępowanie takie byłoby szczególnie przydatne dla tematowych nazw pospolitych i geograficznych oraz określników geograficznych. Wspomnianą wcześniej możliwość zastępowania określników chronologicznych określnikiem „– historia” także można zaliczyć do przedstawianej reguły. Poniżej znajdują się trzy inne przykłady zastosowania tej reguły:

Literatura polska – 1945-1990 → Literatura polska – 20 w → Literatura – 20 w
 Żydzi – Bawaria (Niemcy) → Żydzi – Niemcy → Żydzi – Europa Środkowa
 Niemcy – historia → Europa Środkowa – historia

Należy zwrócić uwagę na fakt, że jeśli w słowniku hipotetycznie nie występowałyby hasła „Literatura polska – 20 w”, to aby powiązać hasło „Literatura polska – 1945-1990” z hasłem „Literatura – 20 w” należałoby zastąpić jednocześnie oba leksemy ich hiperonimami. Dodatkowo jeśli nie znajdziemy hasła utworzonego przez zastąpienie leksemu jego bezpośrednim hiperonimem, to powinniśmy próbować zastąpić go hiperonimami pośrednimi. Wynika z tego, że omawiana reguła jest złożona obliczeniowo.

Rozpatrzmy jeszcze jeden problem. Przypuśćmy, że w słowniku istnieją hasła: „Żydzi – Bawaria (Niemcy) – historia” i „Semici – Bawaria (Niemcy)”, jednak nie istnieją następujące hasła: „Żydzi – Bawaria (Niemcy)” i „Semici – Bawaria (Niemcy) – historia”. Pierwsze hasło w rzeczywistości jest hiponimem drugiego, ale czy dzięki przedstawionym regułom uda nam się tą zależność odkryć? Nie uda to się, jeśli wykorzystamy sekwencyjnie regułę wiązania na podstawie leksemów oraz regułę aktualnie omawianą. Należy je zastosować jednocześnie, to znaczy jednocześnie usuwać różne leksemy oraz zastępować je hiperonimami. W przedstawionym przypadku należy jednocześnie zastąpić pierwszy leksem jego hiperonimem oraz usunąć ostatni leksem.

6.7 Pozostałe czynności do wykonania

Wykonano najtrudniejsze elementy konwersji słownictwa kontrolowanego do tezaury. Jednocześnie pozostawiono do wykonania czynności prostsze i bardziej zorientowane na analizę jakości powiązań hierarchicznych tezaury. W celu dokoń-

czenia budowy tezaurusa należy wykonać następujące czynności:

- a) Zbudowanie brakujących hiperonimii:
 - i) Analizowanie terminów odrzuconych. Do tej pory we wszystkich regułach analizowano jedynie terminy przyjęte haseł. Dzięki analizie terminów odrzuconych można uzyskać dużo nowych hiperonimii.
 - ii) Rozszerzenie zmiany liczby gramatycznej na pozostałe rzeczowniki i ich wyrażenia. Można tego dokonać stosując reguły przetwarzania języka naturalnego.
 - iii) Zastosowanie usprawnień zaimplementowanych reguł budowy hiperonimii oraz zaimplementowanie reguł opisanych w poprzednim punkcie.
 - iv) Sprawdzenie czy w słowniku KABA pozostały jeszcze jakieś niejawne hiperonimie. Jeśli tak to zastanowienie się nad możliwością ich utworzenia.
- b) Usunięcie relacji sprzecznych. W zbudowanej hierarchii tezaurusa niektóre hiperonimie będą tworzyły skierowany cykl. Oczywiście hierarchia powinna być ich pozbawiona. Dlatego trzeba będzie je zlokalizować traktując hierarchię jako graf skierowany i szukając w nim cykli. Znalezione cykle powinny być przeanalizowane, a powód ich powstania usunięty na drodze programistycznej lub przez edycję słownika KABA. Najprostszym cyklem jest sytuacja, w której dwa hasła wskazują wzajemnie na siebie w ten sposób, że między nimi istnieją dwie sprzeczne hiperonimie. Należy zaznaczyć, że fakt ten bardzo często wynika z błędnie podanych relacji jawnych lub innych informacji w słowniku KABA, a nie z błędów w programie.
- c) Oznaczenie relacji nadmiarowych, to jest takich które wynikają z przechodniości hiperonimii. Relacji takich można by nie wyświetlać w przeglądarce tezaurusa w celu ograniczenia liczby bezpośrednich hiponimów.
- d) Ograniczenie liczby bezpośrednich hiponimów. Duża liczba bezpośrednich hiponimów utrudnia przeglądanie tezaurusa. Dlatego też jeśli pewne hasło ma ich dużo, należy zastanowić się z czego to wynika. Sytuacja taka powinna być poprawiona poprzez reorganizację sieci połączeń. Można tego dokonać grupując hiponimy i przypisując tym grupom hasła już istniejące w słowniku albo nowe. W obecnym stanie tezaurusa niektóre hasła na przykład „Polska (geographic)” posiadają bardzo dużo hiponimów. Często wynika to z faktu lokalizacji nazw geograficznych jedynie nazwą kraju w przypadku, gdy jest to wystarczające. Część takich hiponimów zostanie oznaczona jako nadmiarowe, gdyż wynikają one z przechodniości hiperonimii.
- e) Analiza haseł nie posiadających hiperonimów. Wszystkie hasła z wyjątkiem najwyżej kilku powinny posiadać hiperonimy. Dzięki temu będzie można je wybrać w przeglądarce tezaurusa przechodząc go od góry do dołu. Obecnie 1996 z 48688 nazw pospolitych nie posiada relacji do terminu szerszego. W rzeczywistości prawie wszystkie z nich mają terminy szersze łatwe do wyznaczenia dla człowieka.

- f) Sortowanie list terminów węższych i szerszych. Kolejność wyświetlania hiponimów i hiperonimów jest bardzo ważna w przeglądarce tezaursusa. Należy podjąć decyzję, jakie sortowanie będzie najlepsze.
- g) Analiza funkcjonalności tezaursusa podczas stosowania go w przeglądarce tezaursusa. Czy tezaurus umożliwi w szybki sposób zlokalizowanie wszystkich interesujących nas tematów, czy też należy jeszcze coś poprawić?

6.8 Wnioski

W słowniku KABA istnieje około 70 tysięcy hiperonimii jawnie podanych w rekordach haseł. Przy pomocy opisanych reguł utworzono następnych 70 tysięcy hiperonimii nie podanych jawnie. Część z tych haseł będzie prawdopodobnie wynikała z przechodniości innych haseł – takie relacje nie powinny być dodawane do słownika. Jednak utworzenie dodatkowych hiperonimii opisanych w poprzednim punkcie spowoduje, że ostateczna liczba hiperonimii prawdopodobnie będzie większa niż 140 tysięcy. Utworzone hiperonimie są prawie zawsze poprawne, a dzięki interakcji z czytelnikami ten stan będzie można jeszcze poprawić.

Tezaurus posiada mankamenty opisane w poprzednim punkcie. Jednak ich naprawy najlepiej będzie się podjąć po ocenie dotychczasowych efektów przez środowisko bibliotekarzy. Wydaje się, że naprawa mankamentów tezaursusa jest możliwa do zrealizowania akceptowalnym nakładem pracy Centrum NUKAT. Natomiast od strony informatycznej prawie wszystkie problemy zostały rozwiązane i zaimplementowane.

Aktualny stan tezaursusa można ocenić na pokazywanym wcześniej rysunku 3.1 oraz w pliku tekstowym, w którym została zapisana jego zawartość. Można także wykonywać analizy programistyczne jego zawartości dzięki opisanemu w pracy pakietowi `dictionary`.

Rozdział 7

Podsumowanie

W ramach pracy magisterskiej zaprojektowano i zaimplementowano bibliotekę programistyczną umożliwiającą przeprowadzanie programistycznych analiz zawartości słownika haseł przedmiotowych. Dzięki temu znaleziono dużą liczbę błędów w słowniku KABA i przekazano je do poprawienia Centrum NUKAT. Pokazano, że większość niejawnych hiperonimii znajdujących się w słowniku KABA może być utworzona automatycznie przez komputer, wykorzystując w tym celu narzędzia przetwarzania języka naturalnego. Udało się wykonać prawie wszystkie prace programistyczne potrzebne do utworzenia takich relacji, a wiele z tych relacji zostało już utworzonych. Pełna konwersja do tezaury wymaga wykonania pozostałych, w większości nieinformatycznych prac, najlepiej przy współudziale środowiska bibliotekarzy.

Zaproponowane zastosowania utworzonego tezaury nie ograniczają się tylko do analiz statystycznych. Pokazano, że tezaurus może być zastosowany także w wyszukiwaniu publikacji oraz w wyborze interesujących czytelnika tematów.

Wyniki pracy można bezpośrednio zastosować nie tylko w bibliotekach cyfrowych opartych na platformie dLibra, ale w większości polskich bibliotek naukowych. Podobne prace można także wykonać dla wszystkich bibliotek używających katalogu przedmiotowego opartego na słowniku kompatybilnym ze słownikiem LCSH.

Praca pokazuje w jaki sposób Semantyczny Internet i biblioteki mogą wzajemnie wpłynąć na swój rozwój. W obu dziedzinach można znacznie zwiększyć jakość wyszukiwania – w bibliotekach dzięki wykorzystaniu ontologii w katalogach przedmiotowych, a w Semantycznym Internecie dzięki wykorzystaniu słownictwa kontrolowanego do opisu treści stron WWW. Informatyzacja bibliotek prawie w ogóle nie zmieniła zasad działania katalogów przedmiotowych. Jednak dzięki wykorzystaniu najnowszych osiągnięć informatyki, szczególnie ontologii, jest szansa na zmianę paradygmatu organizacji wiedzy panującego od wielu lat w wyszukiwaniu tematycznym.

Praca magisterska sprawiła dużo trudności wynikających z wielkości słownika

KABA, konieczności oczyszczenia go z błędów oraz tego, że część reguł języka haseł przedmiotowych należało odkryć metodą indukcyjną analizując hasła słownika. Z tego powodu zabrakło czasu na zaimplementowanie zastosowań tezaury w tym analiz statystycznych będących jednym z pierwotnych zadań pracy magisterskiej. Jednak równocześnie praca magisterska sprawiła, że wykonanie tego zadania stało się możliwe.

Na koniec chciałbym podziękować wszystkim osobom, które pomagały w napisaniu tej pracy. Po pierwsze chciałem podziękować promotor – profesor Joannie Józefowskiej za wybór ciekawego tematu pracy magisterskiej oraz konsultacje podczas jej trwania. W konsultacjach uczestniczyli także magister Agnieszka Ławrynowicz oraz doktor Tomasz Łukaszewski. Dziękuję doktorowi Jackowi Martinkowi za pomoc w dziedzinie przetwarzania języka naturalnego zarówno podczas studiów jak i podczas pisania pracy magisterskiej. Profesor Józefowskiej oraz doktorowi Martinkowi pragnę także podziękować za poprowadzenie wykładu ze wstępu do sztucznej inteligencji, bez którego powstanie tej pracy magisterskiej nie byłoby możliwe.

Pragnę podziękować osobom związanym z bibliotekami. Marcin Werla z Poznańskiego Centrum Superkomputerowo-Sieciowego przybliżył mi architekturę systemu dLibra. Pani Ewa Kołodzińska z Biblioteki Politechniki Poznańskiej udzielała w początkowym okresie wskazówek na temat kartoteki KABA.

Szczególne podziękowania składam Centrum NUKAT Biblioteki Uniwersytetu Warszawskiego za udostępnienie słownika KABA oraz wyjaśnianie jego niuansów. W szczególności dziękuję Paniom Marii Burchard, Annie Hallay i Marii Nasiłowskiej oraz Panu Szymonowi Siemianowskiemu.

Dziękuję także profesorowi Wiesławowi Babikowi z Uniwersytetu Jagiellońskiego za udzielenie wskazówek bibliograficznych z dziedziny bibliotekoznawstwa.

Literatura

- [1] *Analiza dokumentu i jego opis przedmiotowy*, red. Głowacka T., seria: *Katalogowanie w języku haseł przedmiotowych KABA*, red. Woźniak J., część 1., Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 2003.
- [2] dLibra, Digital Library Framework – <http://dlibra.psnc.pl/>.
- [3] *Język haseł przedmiotowych KABA: zasady tworzenia słownictwa*, red. Głowacka T., Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 2000.
- [4] Library of Congress Subject Headings – <http://authorities.loc.gov>.
- [5] *MARC 21 Concise Format for Authority Data, Update No. 6*, Library of Congress, 2005. Dostępny pod adresem: www.loc.gov/marc/authority/.
- [6] MARC4J, API w Javie umożliwiające pracę z rekordami w formacie MARC i MARC XML – <http://marc4j.tigris.org/>.
- [7] MARC XML, rekordy MARC zapisane w XML-u – www.loc.gov/standards/marcxml/.
- [8] Narodowy Uniwersalny Katalog Centralny NUKAT – www.nukat.edu.pl.
- [9] *Nowa encyklopedia powszechna PWN*, PWN, Warszawa 2001.
- [10] Paluszkiwicz A.: *Format USMARC rekordu kartoteki haseł wzorcowych: zastosowanie w Centralnej Kartotece Haseł Wzorcowych*, Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, Warszawa 1999.
- [11] Swędrzyński A. i inni: *Zastosowanie oprogramowania dLibra do budowy Wielkopolskiej Biblioteki Cyfrowej*, Konferencja *Internet w bibliotekach II. Łączność, współpraca, digitalizacja*, Wrocław 23–26 września 2003. <http://ebib.oss.wroc.pl/matkonf/iwb2/dlibra.php>.
- [12] Szafran K.: *Analizator morfologiczny SAM-95. Opis użytkowy*, Instytut Informatyki UW, 1996. Dostępny pod adresem: www.mimuw.edu.pl/~kszafran/SAM-dists/tr226.ps.
- [13] Tokarski J.: *Schematyczny indeks a tergo polskich form wyrazowych*, PWN, Warszawa 1993.
- [14] Wielkopolska Biblioteka Cyfrowa – www.wbc.poznan.pl.

- [15] Wikipedia: Controlled vocabulary – http://en.wikipedia.org/wiki/Controlled_vocabulary.
- [16] Wikipedia: Universal Decimal Classification – http://en.wikipedia.org/wiki/Universal_Decimal_Classification.

Dodatek A

Zawartość płyty CD

Płyta CD dołączona do pracy magisterskiej zawiera między innymi tekst pracy magisterskiej w postaci elektronicznej, kod źródłowy zbudowanego systemu, a także pliki z danymi importowanymi oraz generowanymi przez system.

W celu dokładnego zapoznania się z zawartością płyty można skorzystać z plików „Readme.txt” znajdujących się w poszczególnych katalogach. Poniżej znajduje się pobeżne omówienie zawartości. Płyta CD składa się z następujących katalogów:

a) **Dokumenty** – katalog zawiera:

- elektroniczną wersję tekstu pracy magisterskiej,
- artykuły streszczające pracę magisterską w polskiej i angielskiej wersji językowej,
- prezentację pracy magisterskiej,
- dwie pozycje literaturowe dostępne w zwartej wersji elektronicznej.

b) **Implementacja** – katalog zawiera materiały powstałe podczas implementacji programu pracy magisterskiej w tym:

- materiały powstałe podczas implementacji modułu do przechowywania listy publikacji oraz modułu do odmiany rzeczowników przez liczbę,
- kompletny program zbudowanego systemu, razem z kodem źródłowym,
- notatki programistyczne powstałe podczas implementacji, a także kody źródłowe wykonujące kilkanaście analiz słownika KABA,
- specjalnie utworzone kilkurekordowe słowniki KABA, które były wykorzystane podczas testów niektórych funkcji.

c) **Dane** – katalog zawiera pliki będące zawartością poszczególnych przetwarzanych struktur w tym:

- pliki z podstawowym opisem publikacji Wielkopolskiej Biblioteki Cyfrowej importowane przez zbudowany system,
- pliki wejściowe i wyjściowe analizatora SAM, a także wynikową tabelę odmian rzeczowników przez liczbę,

- oryginalny słownik KABA oraz wynikowy tezaurus KABA w postaci tekstowej. Elementy te są dostępne tylko w Centrum NUKAT. Dane nie mogły być zawarte na tej płycie, ponieważ słownik KABA nie jest udostępniany publicznie.