Wiesław Babik
Uniwersytet Jagielloński w Krakowie
w.babik@uj.edu.pl

KONTROLA SŁÓW KLUCZOWYCH W INDEKSOWANIU I WYSZUKIWANIU INFORMACJI

Wprowadzenie

Języki informacyjno-wyszukiwawcze stanowią kluczowy komponent systemów informacyjno-wyszukiwawczych związanych z gromadzeniem specjalistycznej informacji wysokiej jakości. W erze Google jesteśmy przyzwyczajeni do korzystania z prostych systemów wyszukiwawczych, które wymagają jedynie wprowadzania słów kluczowych. Jest to szukanie przez słowa kluczowe w stylu Google (Myszor 2015). Chociaż tego typu mechanizmy są efektywne, to zawodzą w przypadku wyszukiwania specjalistycznych materiałów. Wyniki dotychczasowych badań pokazują, że konceptualne wyszukiwanie bazujące na słownikach kontrolowanych jest bardziej efektywne niż wyszukiwanie przez swobodne słowa kluczowe. Warto więc szukać sposobów na poprawę skuteczności wyszukiwania za pomocą słów kluczowych. Potrzeba kontroli słownictwa wynika z cech języka naturalnego wykorzystywanego w funkcji wyszukiwawczej. Cechy te to występowanie różnych form słownych na oznaczenie jednego pojęcia (bliskoznaczność i synonimia) oraz różne znaczenia jednej formy (homonimia). Każdy, kto wyszukuje informacje, doskonale wie, że wyszukiwanie swobodne w odróżnieniu od kontrolowanego cechuje się dużym szumem informacyjnym oraz ciszą informacyjną. W związku z tym niezbędna jest kontrola słownictwa, w tym słów kluczowych. O problemach kontroli słownictwa w SIW pisali m.in.: William John Hutchins (1978); Robert Fugmann (1993); Ewa Chmielewska-Gorczyca (1996); Barbara Sosińska-Kalata (Sosińska-Kalata 1999; Sosińska-Kalata, Roszkowski 2016); Heting Chu (2003); Jutta Bertram (2005); Wiesław Babik (2010); Jadwiga Woźniak-Kasperek (2011); Marcin Roszkowski (2016). W praktyce mamy do czynienia ze słowami kluczowymi będącymi elementarnymi jednostkami leksykalnymi typologicznie różnych języków słów kluczowych.

1. Kontrola słownictwa – podstawowa terminologia i problemy

W rozważaniach na podjęty temat istnieje wiele nieporozumień i niejasności. Ich przyczyną są przeważnie różne rozumienia i interpretacje używanych terminów. Stąd rozważania na temat kontroli słów kluczowych rozpocznę od zdefiniowania podstawowych terminów. Słownictwo kontrolowane jest zwykle stawiane w opozycji do słownictwa niekontrolowanego (swobodnego), które jest charakterystyczne dla języka swobodnych słów kluczowych, będącego w istocie wykorzystaniem języka naturalnego w systemach informacyjno-wyszukiwawczych. Stąd wyszukiwanie swobodne jest nazywane wyszukiwaniem w języku naturalnym. Niezależnie od tego niezbędnym jest rozróżnienie między wyszukiwaniem słów kluczowych (wyrazów języka naturalnego) w polach opisu bibliograficznego a wyszukiwaniem pełnotekstowym.

Kontrola słownictwa może występować zarówno w indeksowaniu, jak i w wyszukiwaniu informacji. W obu przypadkach niezbędny jest odpowiedni słownik. Istnieją systemy, które przerzucają ciężar kontroli używanego słownictwa na użytkownika systemu, nie wspomagając go jakimkolwiek systemowym słownikiem. W takiej sytuacji użytkownik jest zdany na samego siebie lub na korzystanie z innych gotowych słowników, na przykład słowników języka naturalnego, słowników terminologicznych, encyklopedii czy tezaurusów innych systemów.

Słownik jako narzędzie kontroli słownictwa może pełnić funkcję słownika języka indeksowania, jak również być elementem interfejsu użytkownika, wspomagając go w formułowaniu instrukcji wyszukiwawczych w danym języku. Jako element interfejsu jest słownikiem wyszukiwawczym. Może być elementem zarówno systemów pełnotekstowych, jak i systemów o indeksowaniu swobodnym.

2. Stopień kontroli słownictwa i jej formy

Kontrola słownictwa może dotyczyć form gramatycznych wyrazów, synonimii, homonimii, relacji hierarchicznych i kojarzeniowych oraz znaczenia i zakresu słów kluczowych, a także zakresu ich użycia. Formę kontroli słownictwa mogą też stanowić przyjęte w danym systemie zasady dotyczące używania nazw własnych, stosowania skrótów lub form rozwiniętych, stopień prekoordynacji (złożoności) słów kluczowych, kolejności elementów w złożonych słowach i frazach kluczowych. Mam tu na myśli zgodność lub niezgodność z szykiem języka naturalnego.

Kontrola form gramatycznych polega na kontroli liczby gramatycznej (pojedynczej lub mnogiej) wyrazów oraz form słownych, na przykład wyrażanie form czasownikowych i przymiotnikowych za pomocą form rzeczownikowych. Rodzajem

tej kontroli jest także kontrola pisowni (gdy istnieją różne warianty). Dotyczy to również słów kluczowych w innych językach niż polski, na przykład języka angielskiego w wersji brytyjskiej czy amerykańskiej. Eliminowanie liczby pojedynczej i mnogiej, a także ujednolicanie różnych form słownych występuje zarówno w indeksowaniu kontrolowanym słownikiem, jak i w indeksowaniu swobodnym. W drugim przypadku jest realizowane za pomocą wyraźnie określonych kryteriów uznawania wyrażeń za słowo kluczowe (na przykład musi to być rzeczownik w pierwszym przypadku liczby pojedynczej lub mnogiej – *pluralia tantum*). W tym celu tworzy się specjalną instrukcję indeksowania (na przykład dla systemu SYNABA). Ten rodzaj kontroli jest możliwy również w systemach pełnotekstowych dzięki tzw. technice maskowania. Niestety ten rodzaj wyszukiwania z reguły cechuje się dużym szumem informacyjnym. W systemach opartych na indeksowaniu kontrolowanym różne formy słowne mogą być powiązane za pomocą odsyłaczy z formą preferowaną.

Eliminacja synonimii słów kluczowych może następować na etapie zarówno indeksowania, jak i wyszukiwania. W indeksowaniu można grupować słowa kluczowe będące synonimami lub wyrażeniami bliskoznacznymi (quasi-synonimia), a następnie odsyłać je do terminu preferowanego. Procedura ta jest najmocniejszym stopniem kontroli (tzw. „twarda kontrola"). Słabszy rodzaj kontroli (tzw. „miękka kontrola") występuje, gdy łączymy synonimy i wyrazy bliskoznaczne relacją ekwiwalencji. Może to występować zarówno w systemach pełnotekstowych, jak i w systemach opartych na indeksowaniu swobodnym (Babik 1999). Polega na dodawaniu wszystkich słów kluczowych grupy równoważnościowej do słowa kluczowego zaproponowanego przez użytkownika (łączenie za pomocą operatora OR/lub).

Eliminowanie homonimii występuje w zasadzie tylko przy słownictwie kontrolowanym. Wieloznaczność jest eliminowana poprzez rozróżnianie poszczególnych znaczeń homonimów (lub wyrażeń homonimicznych) i zaopatrywanie ich w słowniku w dodatkowe wskaźniki cyfrowe lub literowe (ewentualnie dopowiedzenia będące kwalifikatorami precyzującymi ich znaczenie i zakres (w nawiasach), np. zamek (urządzenie), zamek (budowla). Eliminowanie wieloznaczności jest też możliwe w pewnym stopniu w wyszukiwaniu swobodnym, głównie poprzez wskazanie cech kontekstowych słowa kluczowego, na przykład współwystępowanie dwóch słów lub ich odpowiedniej bliskości.

Specyfikowanie relacji hierarchicznych i kojarzeniowych nie jest powszechnie uznawane za element kontroli słownictwa. Jest to raczej element dodatkowego wspomagania użytkownika. Relacje hierarchiczne i kojarzeniowe wyspecyfikowane w słowniku pełnią funkcję kontroli wyboru słów kluczowych na etapie zarówno indeksowania, jak i wyszukiwania. Określają i zapewniają one przyjęty poziom szczegółowości indeksowania i wyszukiwania, gwarantując, że dane słowo kluczowe jest adekwatne (precyzyjne) w istniejącym zestawie dla wyrażenia danego tematu.

Określają i zapewniają też przyjęty poziom szerokości indeksowania i wyszukiwania, gdyż pomagają w wyborze odpowiednich słów kluczowych niezbędnych do przedstawienia zawartości treściowej danego dokumentu lub pytania informacyjnego.

Terminy pokrewne semantycznie powinny być w słowniku powiązane siecią odsyłaczy. Terminy hierarchiczne i kojarzeniowe są narzędziem uściślania znaczenia słowa kluczowego, a w systemach posiadających słownik zintegrowany z bazą danych umożliwiają automatyczne dodawanie słów kluczowych szerszych lub węższych w poszerzaniu strategii wyszukiwawczej.

Określanie znaczenia i zakresu słowa kluczowego, zwłaszcza gdy odbiega ono od zakresu i znaczenia słowa języka naturalnego, występuje tylko w systemach opartych na indeksowaniu i wyszukiwaniu kontrolowanym opartym na słowniku. Wówczas w słowniku umieszcza się odpowiednie uwagi, podaje definicje słowa kluczowego lub w instrukcji indeksowania i/lub wyszukiwania. Uwagi te (tzw. ang. *scope note*) precyzują pożądane znaczenie słowa kluczowego.

3. Narzędzia kontroli słów kluczowych

Narzędziami kontroli słownictwa są przede wszystkim różnego rodzaju słowniki, ale również indeksy, pierścienie synonimów, chmury tagów i ontologie. Szczegółowo przedstawię pierwszą grupę narzędzi.

Najprostszego typu słownika używają klasyczne języki słów kluczowych. Są to słowniki jednojęzyczne. Następuje w nich „naturalizacja" wyrażeń języka naturalnego użytych w funkcji metainformacyjnej. Przykładem systemów wykorzystujących klasyczne języki słów kluczowych mogą być SYNABA czy BAZTECH. Słowniki tych systemów stanowią najprostsze alfabetyczne listy słów kluczowych bez jakichkolwiek informacji dodatkowych.

Słownik słów kluczowych systemu SYNABA podaje słowa kluczowe uporządkowane alfabetycznie i nie zawiera żadnych odsyłaczy. Jest to słownik jawny klasycznego języka słów kluczowych w wydaniu autorskim. Słowami kluczowymi są wyrażenia języka naturalnego ujednolicone stosownie do wymagań języka polskiego. Słownik ten zawiera przeważnie jednowyrazowe słowa kluczowe, ale też wielowyrazowe słowa kluczowe oraz frazy kluczowe. Słowami kluczowymi są też nazwy własne, w tym osobowe oraz akronimy.

Autorskie (albo odautorskie) słowa kluczowe stanowią odmianę języka swobodnych słów kluczowych wyróżnianą ze względu na indeksatora, którym jest autor. W tym wypadku źródłem słów kluczowych jest tytuł pracy, jej abstrakt oraz sam tekst dokumentu, a słownik ma charakter niejawny i opiera się na słowniku mentalnym autora. Jeżeli słowa kluczowe są pobierane z wymienionych źródeł, to sposób indeksowania pokrywa się z indeksowaniem derywacyjnym.

Jak pokazuje praktyka leksykograficzna języków słów kluczowych, słowniki słów kluczowych często zaopatruje się w odsyłacze. Przyczyną tego jest niewystarczalność struktury słowników klasycznych języków słów kluczowych do odwzorowywania rzeczywistości dokumentacyjnej w systemie informacyjno-wyszukiwawczym. Przykładem wzbogacenia list słów kluczowych o system odsyłaczy, który explicite pokazuje elementy struktury paradygmatycznej języka, są słowniki słów kluczowych zbudowane przez OIN PAN w latach 1988–1992. Jednak i w tym przypadku paradygmatyka języka implicite tkwi w kompetencji jego użytkownika w zakresie języka polskiego, a więc interpretacja znaczenia poszczególnych słów kluczowych zależy przede wszystkim od użytkownika w zakresie języka naturalnego (w jego wersji specjalistycznej, czyli terminologii) i jego wiedzy pozajęzykowej.

W omawianej grupie słowników zastosowano sieć relacji asocjacyjnych odwzorowujących powiązania znaczeniowe pomiędzy słowami kluczowymi. Wzbogacenie struktury paradygmatycznej języka słów kluczowych o tego typu zależności semantyczne było rzadko stosowane. W tych słownikach relacje kojarzeniowe nie są specyfikowane: słowom kluczowym przyporządkowane zostały skojarzone terminy. Stosownie do pragmatyki ogólnej języków słów kluczowych układ jest alfabetyczny, słowa kluczowe będące rzeczownikami występują w formie mianownika liczby pojedynczej, w liczbie mnogiej zaś – tylko tzw. *pluralia tantum*. W wielowyrazowych słowach kluczowych zachowano obowiązujący w języku polskim szyk wyrazów[1]. Przy wyborze słów kluczowych kierowano się powszechnie przyjętymi kryteriami, jak bieżące i częste stosowanie, poprawność terminologiczna, wyrazistość strukturalna, rodzimość, zwięzłość.

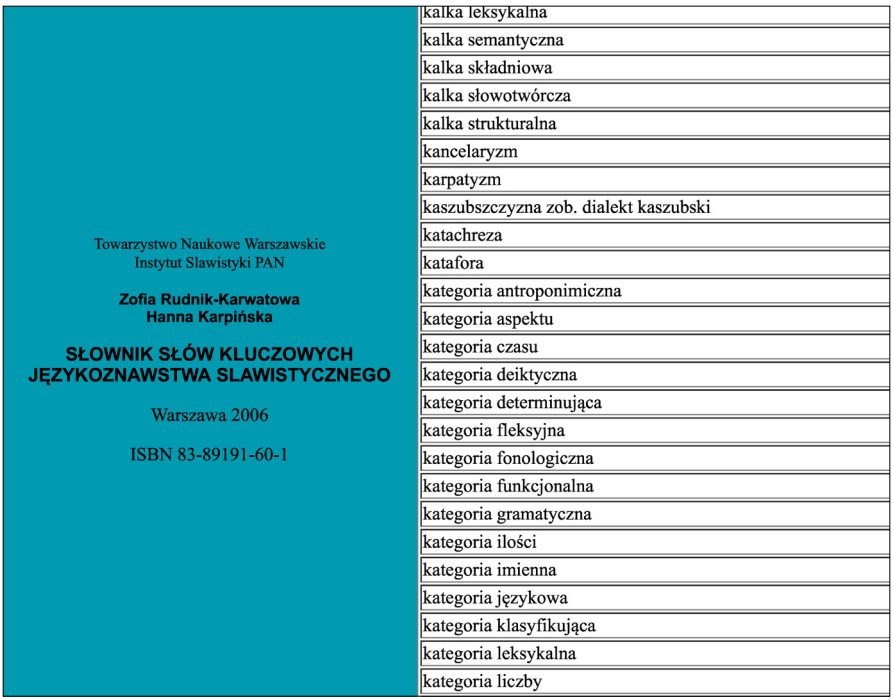
Innym przykładem omawianej grupy słowników słów kluczowych jest *Słownik słów kluczowych językoznawstwa slawistycznego* (Rudnik-Karwatowa, Karpińska 1999), który powstał na potrzeby opisu dokumentów w bazie danych z zakresu językoznawstwa slawistycznego (nowsza wersja opublikowana na CD-ROM: Rudnik-Karwatowa, Karpińska 2006). Oto fragment tego słownika z 1999 roku (Rudnik-Karwatowa, Karpińska 1999: 32):

[…]
derywacja
derywacja afiksalna zob. afiksacja
derywacja alternacyjna
derywacja dezintegralna
derywacja fleksyjna zob. derywacja paradygmatyczna
derywacja morfologiczna
derywacja paradygmatyczna

[1] Stosowanie w obrębie fraz kluczowych metody pozycyjnej w języku polskim sprawia, że najpierw jest wyraz określany, a potem określający.

derywacja prefiksalna zob. prefiksacja
derywacja prefiksalno-sufiksalna
derywacja przedrostkowa zob. prefiksacja
derywacja przyrostkowa zob. sufiksacja
derywacja semantyczna
derywacja słowotwórcza
derywacja sufiksalna zob. sufiksacja
derywacja syntaktyczna
derywacja ujemna zob. derywacja wsteczna
derywacja wymienna
[…]

oraz nowszego wydania tego słownika, opublikowanego w 2006 roku (Rudnik-Karwatowa, Karpińska 2006; rysunek 1).



Rysunek 1. *Słownik słów kluczowych językoznawstwa slawistycznego* (wersja online)

Podane tu informacje odnoszą się do najnowszego wydania tego słownika. Prezentuje on słowa kluczowe językoznawstwa slawistycznego. Jest to zatem dziedzinowy słownik specjalistyczny. We wstępie podano informacje, które przytoczę

poniżej. Leksyka odwzorowuje pole semantyczne dziedziny oraz jej produkcję wydawniczą. Słownik zawiera ponad 3000 słów kluczowych[2]. Jego zadaniem jest pomoc w indeksowaniu i wyszukiwaniu dokumentów, przede wszystkim w tworzonej bazie danych iSybislaw, która stanowi elektroniczną edycję międzynarodowej bibliografii językoznawstwa slawistycznego. W stosunku do wcześniejszej wersji słownik uzupełniono na podstawie reprezentatywnych tekstów bieżącego piśmiennictwa i jego opisów dokumentacyjnych oraz na podstawie źródeł leksykograficznych. Przy gromadzeniu zasobu leksykalnego zastosowano więc metody korpusowe, natomiast sam słownik opracowano metodą indukcyjno-dedukcyjną. Jak piszą autorki we *Wstępie*, słownik ma być traktowany jako kontrolowana lista słów kluczowych. Jest jednojęzyczny, o układzie alfabetycznym. Charakterystyczną cechą języka słów kluczowych w tym słowniku jest oparcie systemu leksykalnego wyłącznie na systemie terminologicznym z zachowaniem zasady *literary warrant*[3]. Do słownika nie wprowadzono nazw języków niesłowiańskich i nazw własnych (poza nazwami szkół językoznawczych). Jedyną relacją syntagmatyczną zachodzącą między słowami kluczowymi jest relacja współwystępowania. Struktura słownika jest płaska, co oznacza, że nie uwzględniono w nim żadnych relacji hierarchicznych. W słowniku zastosowano wskaźnik relacji *zob.*, który odsyła od terminów języka naturalnego do słowa kluczowego będącego terminem preferowanym języka naturalnego. Słowa kluczowe podano czcionką prostą. Terminy niebędące słowami kluczowymi wyróżniono kursywą. Terminy polisemiczne zostały opatrzone cyframi arabskimi oraz uzupełnione (w nawiasie) dopowiedzeniem precyzującym zakres znaczeniowy (cyfry stanowią integralną część słowa kluczowego). Gramatyka tego języka to reguły indeksowania współrzędnego. Sposób prezentacji słownictwa jest typowy dla języków słów kluczowych.

Nowym rozwiązaniem niedawno zastosowanym w słownikach słów kluczowych jest układ gniazdowy słów kluczowych wspomagany systemem terminologicznym. Ten sposób prezentacji języka słów kluczowych dostarcza jeszcze więcej informacji o jego strukturze semantycznej. Zastosowano go w języku słów kluczowych etno-

[2] Wersja słownika z 1999 roku zawierała około 2500 słów kluczowych.

[3] Metoda *literary warrant* została opracowana na początku ubiegłego wieku przez E. Wyndhama Hulme'a i polega na analizie kolekcji pod kątem tematów odwzorowywanych w procesie indeksowania. Uzależnia wybór danego terminu od jego częstego występowania w literaturze. Na podstawie analizy treściowej dokumentów kolekcji bierze się pod uwagę wyłącznie te tematy, które posiadają co najmniej jednego reprezentanta w postaci CHWD. Niedopuszczalne jest tu tworzenie tzw. tematów pustych. W ten sposób generuje się w języku słów kluczowych wykaz wyłącznie autosyntaktycznych jednostek leksykalnych stosowanych do reprezentacji treści dokumentów zbioru informacyjnego. Niektórzy autorzy wyróżniają również tzw. *user warrant* oraz *concept warrant*. Możliwy jest też model hybrydowy.

logii. Język ten był tworzony od roku 1993 w ramach prac zespołu kierowanego przez prof. Czesława Robotyckiego z Instytutu Etnologii i Antropologii Kulturowej Uniwersytetu Jagiellońskiego nad porządkowaniem terminologii tej dyscypliny. Ponieważ uznano, że terminologia nie tylko ułatwia dostęp do wiedzy, lecz także jest źródłem słów kluczowych, umożliwiających odwzorowanie treści dokumentów (źródeł etnograficznych) dla potrzeb systemu informacyjno-wyszukiwawczego, zdecydowano się połączyć te prace z pracami nad słownikiem, a właściwie słownikami słów kluczowych dla poszczególnych kategorii kultury, które to słowniki mają być wykorzystane w budowanym przez Instytut systemie informacji o źródłach archiwalnych dotyczących Karpat polskich o nazwie PROKES. Wcześniej zespół ten opublikował w 1995 roku rezultaty swoich prac jako *Układ słów kluczowych dla bazy danych o źródłach etnograficznych (Kultura ludowa Karpat Polskich)* (Robotycki [red.] 1995).

Listom słów kluczowych towarzyszą spisy terminów w układzie gniazdowym. Pokazują one explicite kategoryzację semantyczną jednostek leksykalnych, a także elementy struktury paradygmatycznej tego języka. Są to w zasadzie pola asocjacyjne łączące terminy podstawowe w etnologii, nazywające główne przedmioty badań z wyrażeniami kojarzącymi się z nimi na zasadzie bliżej nieokreślonych i niespecyfikowanych w słowniku relacji paradygmatycznych. Tego typu układy są jednak bardzo przydatne dla użytkownika. Do *Układu gniazdowego…* dołączono fasety w postaci wykazów świąt, świętych oraz kościołów (wyznań) i roślin leczniczych. Tak prezentowany język nie jest już klasycznym językiem słów kluczowych, którego charakterystyczną cechę stanowi z założenia płaska struktura leksyki, a jego paradygmatyka tkwi implicite w kompetencji jego użytkownika w zakresie odpowiedniego języka naturalnego. Ma to niewątpliwy wpływ na proces indeksowania. Nieukazana explicite paradygmatyka nie może być wykorzystana w wyszukiwaniu.

Omawiany język jest przykładem takiej podwójnej prezentacji systemu leksykalnego języka słów kluczowych. Zilustruję to fragmentem układu gniazdowego terminów wybranych kategorii kultury (Robotycki, Babik [red.]: 30–32).

[…]
SUCHOTY, *astma*, *dera*, *gruźlica*, *nędza*
. . *ETIOLOGIA*
. . . BAWIENIE SIĘ Z KOTEM
. . . CZAR
. . . DZIEDZICZENIE
. . . POŁKNIĘCIE SIERŚCI
. . . POŁKNIĘCIE WŁOSA
. . . POŁKNIĘCIE KOCIEGO WŁOSA
. . . PRZEZIĘBIENIE
. . . TĘSKNOTA ZA PIERSIĄ

. . *RODZAJE*
. . . SUCHOTY GALOPOWE, *suchoty ostre*, *suchoty śmiertelne*
. . . SUCHOTY GARDLANE
. . . SUCHOTY PŁUCNE, *suchoty wewnętrzne*
. . . SUCHOTY ŚWIATOWE
. . . SUCHOTY ŻOŁĄDKOWE

. . *OBJAWY*
. . . BEZSENNOŚĆ
. . . BÓL W PIERSIACH
. . . BRAK APETYTU
. . . CHUDNIĘCIE, *schnięcie*
. . . GORĄCZKA
. . . KASZEL

. . *LEKI I SPOSOBY LECZENIA*
. . . KĄPIEL
. . . . GROCHOWINY
. . . . KAWAŁKI PNIA DO RĄBANIA MIĘSA
. . . . KLUSKI PSZENNE
. . . . ODWAR Z DZIEWANNY
. . . . ODWAR Z GAŁĘZI DĘBU
. . . . ODWAR Z PODRÓŻNIKA
. . . . OSKROBINY [z ziemniaków]
. . . . POMYJE
. . . . Z KOTEM
. . . . Z PSEM
. . . . ZIEMIA Z CMENTARZA
. . . PICIE
. . . . ODWAR Z BAGNA
. . . . ODWAR Z KORZENIA ŻYWOKOSTU
. . . . ODWAR Z LIŚCI PODBIAŁU
. . . . ODWAR Z PĘDÓW SOSNY
. . . SMAROWANIE, NACIERANIE
. . . . ŁÓJ ZE ŚWIECY
. . . . TŁUSZCZ, SADŁO RAKA
. . . . TŁUSZCZ, SADŁO ZAJĘCZE
. . . WKŁADANIE CHOREGO DO PIECA
. . . WKŁADANIE CZASZKI KONIA DO ŁÓŻKA
. . . ZAMAWIANIE
. . . ZAŻEGNYWANIE

. . *CZAS LECZENIA*
. . *MIEJSCE LECZENIA*
. . *OSOBA LECZĄCA*

I odpowiadający mu następujący fragment słownika słów kluczowych:

[…]
SUCHOTY
SUCHOTY GALOPOWE
SUCHOTY GARDLANE
suchoty ostre zob. SUCHOTY GALOPOWE
SUCHOTY PŁUCNE
suchoty śmiertelne zob. SUCHOTY GALOPOWE
SUCHOTY ŚWIATOWE
suchoty wewnętrzne zob. SUCHOTY PŁUCNE
SUCHOTY ŻOŁĄDKOWE
suchoty zob. ASTMA
[…]

Omówione słowniki słów kluczowych to struktury ahierarchiczne, a zatem i równoważnościowe struktury monorelacyjne. Daje to możliwość indeksów rzeczowych, formalnych i mieszanych. Tego typu indeksy mają formę alfabetycznych wykazów jednego rodzaju wyrażeń. Zwykle jest to alfabetyczny wykaz słów kluczowych. Takie indeksy zawierają również jednostki leksykalne w postaci nazw własnych: osobowe, korporatywne oraz geograficzne. Dla tej grupy słów kluczowych zwykle stosuje się odrębną metodę kontroli słownictwa, którą stanowią kartoteki wzorcowe normalizujące formy językowe, na przykład nazw języków etnicznych. Specyfika takich narzędzi kontrolnych zakłada wykorzystanie równoważnościowych struktur organizujących zbiór informacyjny. Tym samym porządek wyrażeń pełniących funkcje wyszukiwawcze ma charakter formalny i wykorzystuje układ alfabetyczny. Wykorzystanie języka słów kluczowych w dostępie do zbioru polega na traktowaniu jego słownika przede wszystkim jako źródła słownictwa w procesie automatycznego wyboru nazw dla punktów dostępu. Tego typu narzędzia opierają się na generowaniu konstrukcji pola semantycznego, która pełni funkcję swoistego rodzaju mapy konceptualnej, dającej użytkownikowi możliwość wglądu w dystrybucję tematów w obiektach informacyjnych kolekcji. Konstrukcje ahierarchiczne (płaskie) są ubogie, gdyż charakterystyka dokumentu jest w nich przyporządkowywana tylko do alfabetycznego indeksu słów kluczowych, w którym porządek słów kluczowych ma formę alfabetyczną, a nie logiczną. Brak podziału logicznego (rozłącznego i adekwatnego) powoduje sytuację, w której na tym samym poziomie wyodrębnia się elementy treści o różnym stopniu szczegółowości, co powoduje pewnego rodzaju niespójność systemu leksykalnego języka słów kluczowych.

Prezentacja słów kluczowych w porządku alfabetycznym wprawdzie nie dostarcza złożonych zależności semantycznych, jednak siłą tego narzędzia jest możliwość dynamicznego modyfikowania zakresu pytania informacyjnego poprzez wykorzystanie do łączenia jednostek leksykalnych algebry Boole'a. Należy tu jeszcze wspomnieć

o możliwości wariantu hybrydowego, w którym nie występują określone zależności strukturalne. Jego podstawową zaletą jest duża elastyczność, która wynika z wielowymiarowej strukturalizacji pola semantycznego języka słów kluczowych. Oparcie punktów dostępu na fasetowym modelu organizacji wiedzy stanowi w tym języku jednak pewnego rodzaju novum.

Wprowadzanie do list słów kluczowych oraz indeksów coraz to nowych słów na podstawie terminów używanych „chwilowo" przez autorów, a nawet w pewnych okresach rozpowszechnionych, wprawdzie ułatwia doraźne poszukiwania, równocześnie jednak staje się źródłem wielu synonimów, jak również homonimów pochodzących stąd, że różni autorzy używają tych samych wyrażeń w różnych znaczeniach. Ważnym problemem jest również uogólnianie bądź wyszczególnianie; słowa wyszczególniające ułatwiają często przydział, utrudniają jednak poszukiwania prac syntetyzujących.

W związku z tym w kontroli słów kluczowych bardzo często wykorzystywane są pierścienie synonimów (ang. *synonym rings*). „Oprócz wiązania relacją ekwiwalencji wyrażeń synonimicznych i bliskoznacznych, łączy się w nich również warianty językowe dla nazw osobowych, zwiększając tym samym kompletność wyszukiwania informacji" (Sosińska-Kalata, Roszkowski 2016: 344).

Od niedawna w technologii Web 2.0 do charakterystyki tradycyjnych i internetowych obiektów cyfrowych stosuje się tzw. folksonomie. Pozwalają one w nowy sposób interpretować relacje między użytkownikiem a zasobami (internetowymi) oraz usługami oferowanymi przez ten system. Użytkownik ma tu większą możliwość wpływania na liczbę i rodzaj udostępnianych mu informacji, stając się nie tylko aktywnym konsumentem, lecz również aktywnym twórcą. Wykorzystując intelektualną aktywność użytkownika, folksonomie angażują go do charakterystyki tych obiektów, tj. tworzenia tekstów w postaci ciągu niekontrolowanych (swobodnych) słów kluczowych reprezentujących treść dokumentu. Folksonomie są wspólnym narzędziem katalogowania/tagowania dokumentów graficznych, dźwiękowych, audiowizualnych, hipertekstowych, a także tradycyjnych. Wykaz użytych słów kluczowych tworzy indeks o strukturze hipertekstowej w postaci tzw. chmury tagów powiązanych z terminem wyszukiwawczym (skojarzeniami, terminami w innych językach). Każde z użytych słów kluczowych otrzymuje status węzła hipertekstowego, którego aktywacja powoduje wyodrębnienie podzbioru charakterystyk obiektów cyfrowych. W ten sposób powstaje jakby równoległa (tym razem z punktu widzenia użytkownika) charakterystyka danego obiektu internetowego, będąca przejawem społecznego klasyfikowania obiektów cyfrowych zasobów WWW, ale obarczona wieloma mankamentami będącymi skutkiem braku kontroli słownictwa.

Słowniki słów kluczowych o strukturze gniazdowej są bliskie internetowym ontologiom, które stają się elementem tzw. semantycznego Webu. Ontologie stanowią rodzaj rozbudowanej sieci semantycznej reprezentującej pojęciową strukturę wiedzy

zawartej w zasobach internetu. W węzłach tej sieci są umieszczane różnojęzyczne wyrażenia (nazwy i terminy). Węzły te są wiązane za pomocą odpowiednich relacji w grupy kategorialne i gniazda semantyczne. W ten sposób słowa kluczowe stają się lingwistycznym narzędziem filtrowania informacji, bez udziału człowieka.

Bez słownika lub innego narzędzia wspomagającego użytkownika systemu informacyjno-wyszukiwawczego w procesie indeksowania i wyszukiwania informacji może on polegać tylko na własnej inwencji, pomysłowości, pamięci i znajomości zagadnienia, co nie zawsze przynosi oczekiwane efekty.

4. Korzyści z kontroli słów kluczowych

Kontrola słów kluczowych podnosi zarówno precyzję, jak i kompletność wyszukiwania. Swobodne wyszukiwanie wymaga od użytkownika słów kluczowych pewnych umiejętności i przygotowania w zakresie stosowania technik maskowania lub operatorów algebry Boole'a (zwłaszcza przy wyszukiwaniu zaawansowanym). Kontrola słów kluczowych może dotyczyć różnych ich aspektów: od kontroli formy gramatycznej słów kluczowych, po łączenie słów kluczowych uznanych za synonimiczne (kontrola synonimii) oraz rozbijanie terminów wieloznacznych (kontrola homonimii).

Wyeliminowanie synonimii w wyszukiwaniu za pomocą słów kluczowych sprawia, że indeksatorzy i użytkownicy używają tych samych słów kluczowych do reprezentowania danego zagadnienia. Użycie jednego terminu, a nie sumy wszystkich form synonimicznych w instrukcji wyszukiwawczej umożliwia wyszukanie wszystkich dokumentów dotyczących danego zagadnienia. Kontrola form gramatycznych i słownych w indeksowaniu umożliwia redukcję liczby słów kluczowych, którym jest łatwiej zarządzać.

Podsumowanie

Na podstawie przeglądu piśmiennictwa po 2000 roku, poświęconego efektywności i zasadności stosowania słów kluczowych Tina Gross i Arlene G. Taylor stwierdziły przydatność wyszukiwania przez słowa kluczowe dla pobieżnych przeszukiwań i jego ograniczenia w przypadku pogłębionych kwerend, w tym prowadzonych w celach naukowych. Dostrzegły też, że w wielu projektach badawczych z różnych dyscyplin uzyskano podobne wyniki, wskazujące że od 1/4 do 1/3 rekordów zostałoby utraconych bez kontrolowanego słownictwa (Gross, Taylor, Jourdrey 2015; Waleszko 2015). Badania te dostarczają mocnych argumentów za kontrolą słownictwa i za używaniem sformalizowanej terminologii.

Słowniki słów kluczowych zintegrowane z bazą danych systemu umożliwiają weryfikowanie poprawności zapisów (eliminacja błędów popełnianych przez indeksatorów i użytkowników przy wprowadzaniu danych), a przy wyszukiwaniu swobodnym umożliwia automatyczne dodawanie form odrzuconych (synonimów). Bardzo podnoszącym wartość słownika słów kluczowych jest dodawanie odpowiedników obcojęzycznych umożliwiających wyszukiwanie obcokrajowcom. Wykorzystywanie w wyszukiwaniu słownika słów kluczowych gwarantuje uzyskanie lepszych wyników wyszukiwawczych niż korzystanie z systemu bez jakiegokolwiek słownika wspierającego użytkownika w wykorzystywaniu języka naturalnego w funkcji wyszukiwawczej.

Omówione wyniki przeprowadzonej analizy języków słów kluczowych potwierdzają szeroki (i znaczenie szerszy niż w innych językach) oraz wzrastający stopień uwzględniania w tych językach relacji kojarzeniowych. Przejawem tej tendencji są występujące w językach słów kluczowych struktury oparte wyłącznie na relacjach kojarzeniowych. Stopień, rzetelność i umiejętność korzystania z tego typu metadanych najczęściej pozostawia wiele do życzenia. Większość systemów indeksujących ignoruje je, poddając analizie statystycznej wyłącznie tekst dokumentu.

BIBLIOGRAFIA

Babik Wiesław, 1999, *Synonimia i homonimia – czy naprawdę niepożądane w nowoczesnych systemach informacyjno-wyszukiwawczych*, „Biuletyn Instytutu Metali Nieżelaznych", numer specjalny, s. 83–87.

Babik Wiesław, 2010, *Słowa kluczowe*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Bertram Jutta, 2005, *Einführung in die inhaltliche Erschliessung. Grundlagen–Methoden–Instrumente*, Language and Communication. Terminology, Language Resources and Semantic Interoperability 2, Würzburg: Ergon Verlag.

Chmielewska-Gorczyca Ewa, 1991, *Język wyszukiwawczy a potrzeby informacyjne użytkowników*, „Zagadnienia Informacji Naukowej", nr 1(58), s. 3–39.

Chmielewska-Gorczyca Ewa, 1996, *Kontrola słownictwa w systemach informacyjno-wyszukiwawczych*, „Українсько-польський науково-практичний журнал «Наука, інновація, інформація»", nr 1, s. 77–83.

Chu Heting, 2003, *Information representation and retrieval in the digital age*, ASIST Monograph Series, Medford: American Society for Information Science and Technology.

Fugmann Robert, 1993, *Subject analysis and indexing. Theoretical foundation and practical advice*, Frankfurt am Main: INDEKS Verlag.

Gross Tina, Taylor Arlene G., Joudrey Daniel N., 2015, *Still a lot to lose. The role of controlled vocabulary in keyword searching*, „Cataloging & Classification Quarterly", t. 53, nr 1–4, s. 1–39, https://doi.org/10.1080/01639374.2014.917447

Hutchins William J., 1978, *Languages of indexing and classification. A linguistic study of structures and functions*, Stevenage: Peter Peregrinus Press.

Myszor Justyna, 2015, *Słowa kluczowe w Google'ach*, praca magisterska napisana pod kierunkiem prof. dr. hab. Wiesława Babika, Kraków: Instytut Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Jagiellońskiego.

Robotycki Czesław (red.), 1995, *Układ słów kluczowych dla bazy danych o źródłach etnograficznych (kultura ludowa Karpat polskich)*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Robotycki Czesław, Babik Wiesław (red.), 2005, *Układ gniazdowy terminów i słownik słów kluczowych wybranych kategorii kultury. Medycyna ludowa*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Rudnik-Karwatowa Zofia, Karpińska Hanna, 1999, *Słownik słów kluczowych językoznawstwa slawistycznego*, Warszawa: Towarzystwo Naukowe Warszawskie, Instytut Slawistyki Polskiej Akademii Nauk.

Rudnik-Karwatowa Zofia, Karpińska Hanna, 2006, *Słownik słów kluczowych językoznawstwa slawistycznego*, wyd. 2, CD-ROM, Warszawa: Towarzystwo Naukowe Warszawskie, Instytut Slawistyki Polskiej Akademii Nauk.

Sosińska-Kalata Barbara, 1999, *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach*, Nauka–Dydaktyka–Praktyka 33, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Sosińska-Kalata Barbara, Roszkowski Marcin, 2016, *Organizacja informacji i wiedzy*, w: *Nauka o informacji*, red. Wiesław Babik, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, s. 305–357.

Waleszko Małgorzata, 2015, *Rola słowników kontrolowanych w wyszukiwaniu przez słowa kluczowe*, http://babin.bn.org.pl/?p=3570 (dostęp 10.09.2018).

Woźniak-Kasperek Jadwiga, 2011, *Wiedza i język informacyjny w paradygmacie sieciowym*, Nauka–Dydaktyka–Praktyka 125, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

BIBLIOGRAPHY (TRANSLITERATION)

Babik, W. (1999). Synonimia i homonimia – czy naprawdę niepożądane w nowoczesnych systemach informacyjno-wyszukiwawczych. *Biuletyn Instytutu Metali Nieżelaznych*, *1999*(numer specjalny), 83–87.

Babik, W. (2010). *Słowa kluczowe*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Bertram, J. (2005). *Einführung in die inhaltliche Erschliessung: Grundlagen–Methoden–Instrumente*. Würzburg: Ergon Verlag.

Chmielewska-Gorczyca, E. (1991). Język wyszukiwawczy a potrzeby informacyjne użytkowników. *Zagadnienia Informacji Naukowej*, *1*(58), 3–39.

Chmielewska-Gorczyca, E. (1996). Kontrola słownictwa w systemach informacyjno-wyszukiwawczych. *Ukraïns'ko-pol's'kyĭ naukovo-praktyvhnyĭ zhurnal "Nauka, innovatsiia, informatsiia"*, 1996(1), 77–83.

Chu, H. (2003). *Information representation and retrieval in the digital age*. Medford, NJ: American Society for Information Science and Technology.

Fugmann, R. (1993). *Subject analysis and indexing: Theoretical foundation and practical advice*. Frankfurt am Main: INDEKS Verlag.

Gross, T., Taylor, A. G., & Joudrey, D. N. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly*, *53*(1–4), 1–39. https://doi.org/10.1080/01639374.2014.917447

Hutchins, W. J. (1978). *Languages of indexing and classification: A linguistic study of structures and functions*. Stevenage: Peter Peregrinus Press.

Myszor, J. (2015). *Słowa kluczowe w Google'ach* (Unpublished MA thesis). Kraków: Instytut Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Jagiellońskiego.

Robotycki, C. (Ed.). (1995). *Układ słów kluczowych dla bazy danych o źródłach etnograficznych (kultura ludowa Karpat polskich)*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Robotycki, C., & Babik, W. (Eds.). (2005). *Układ gniazdowy terminów i słownik słów kluczowych wybranych kategorii kultury: Medycyna ludowa*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Rudnik-Karwatowa, Z., & Karpińska, H. (1999). *Słownik słów kluczowych językoznawstwa slawistycznego*. Warszawa: Towarzystwo Naukowe Warszawskie.

Rudnik-Karwatowa, Z., & Karpińska, H. (2006). *Słownik słów kluczowych językoznawstwa slawistycznego* (2nd ed.) [CD-ROM]. Warszawa: Towarzystwo Naukowe Warszawskie.

Sosińska-Kalata, B. (1999). *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Sosińska-Kalata, B., & Roszkowski, M. (2016). Organizacja informacji i wiedzy. In W. Babik (Ed.), *Nauka o informacji* (pp. 305–357). Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Waleszko, M. (2015). *Rola słowników kontrolowanych w wyszukiwaniu przez słowa kluczowe*. Retrieved September 10, 2018, from http://babin.bn.org.pl/?p=3570

Woźniak-Kasperek, J. (2011). *Wiedza i język informacyjny w paradygmacie sieciowym*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Keywords management in information indexing and retrieval

Abstract

This article concerns selected issues relating to the methods of vocabulary management in the processes of information indexing and retrieval, based on the use of keywords. The aim is to present the results of a review of literature on the subject, published mainly after 2010, and to discuss the state of research and prospects in this field. The study relies on critical analysis of literature on the subject. The obtained picture of the research field in question is an important stage in designing further studies in this area.

Kontrola słów kluczowych w indeksowaniu i wyszukiwaniu informacji

Abstrakt

Przedmiotem artykułu są wybrane problemy dotyczące sposobów kontroli słownictwa w procesach indeksowania i wyszukiwania informacji za pomocą słów kluczowych. Celem publikacji jest prezentacja wyników przeprowadzonego rekonesansu literaturowego publikacji, które ukazały się na ten temat, głównie po 2000 roku, ukazująca stan badań i perspektywy badawcze w tym zakresie. W ustaleniu stanu badań posłużono się metodą analizy i krytyki piśmiennictwa. Uzyskany w wyniku przeprowadzonych badań obraz eksplorowanej problematyki badawczej stanowi ważny etap w projektowaniu dalszych badań w tym zakresie.

Keywords: information retrieval languages; keywords; keywords management; indexing; search for information

Słowa kluczowe: języki informacyjno-wyszukiwawcze; słowa kluczowe; kontrola słów kluczowych; indeksowanie; wyszukiwanie informacji

Wiesław Babik