Wytyczne KPWr - słowa kluczowe (1.0)

Marcin Oleksy, Paweł Kędzia, Agnieszka Turek, Jan Wieczorek

Wstęp

Niniejsze wytyczne odnoszą się do opisywania słowami kluczowymi tekstów wchodzących w skład Korpusu Języka Polskiego Politechniki Wrocławskiej. Słowa te stanowią jedną z kategorii metadanych przypisywanych do dokumentów tego korpusu. W pierwszej części przedstawiono rozumienia terminu “słowo kluczowe” w opracowaniach z zakresu teorii informacji oraz w innych zarówno polskich, jaki i zagranicznych korpusów stanowiących punkt odniesienia. W drugiej opisano rozwiązania przyjęte podczas znakowania korpusu KPWr.

Przegląd definicji

Teoria informacji

ISO 5963

Słowo kluczowe to „wyraz lub grupa wyrazów, możliwie leksykograficznie ujednoliconych, wybranych z tytułu lub tekstu dokumentu charakteryzujących jego zawartość i służących do jego wyszukiwania”.

Encyklopedia współczesnego bibliotekarstwa polskiego (Głombiowski i in.)

Element tekstu (wyrażenie, termin, określenie) dokumentu posiadający wartość znaczącą i charakteryzujący zawartość treściową dokumentu lub jego pewnego aspektu. S.k. może być wybrane z tytułu, tekstu podstawowego dokumentu, bądź też z analizy dokumentacyjnej czy adnotacji. S.k. stosowane są do bezpośredniego, swobodnego indeksowania dokumentów, przy czym dokument może być scharakteryzowany za pomocą jednego lub (z reguły) większej liczby s.k. Służą one również jako materiał wyjściowy do ustalania deskryptorów budowy tezaurusów. Przede wszystkim jednak stanowią podstawę do sporządzania indeksów indeksów s.k., najczęściej w postaci indeksów permutowanych, uzyskiwanych w procesie automatycznego przetwarzania informacji.

Słownik terminologiczny informacji naukowej (Dembowska i in.)

Wyraz lub wyrażenie **wybrane z tytułu lub tekstu** dokumentu, charakteryzujące jego treść.

Słownik Encyklopedyczny (Bojar)

Wyrażenie z tekstu dokumentu (często z jego tytułu lub tytułów rozdziałów) lub zapytania informacyjnego charakteryzujące jego treść.

ODLIS Online Dictionary for Library and Information Science

A significant **word or phrase in the title, subject headings (descriptors), contents note, abstract, or text** of a record in an online catalog or bibliographic database that can be used as a search term in a free-text search to retrieve all the records containing it.

Most online catalogs and bibliographic databases include an option that allows the user to type words that describe the research topic (in any order) and retrieve records containing the search terms in the data fields the system is designed to search whenever the keywords option is selected. One disadvantage of a keywords search is that it does not take into account the meaning of the words used as input, so if a term has more than one meaning, irrelevant records (false drops) may be retrieved. Keywords are also used as access points in KWAC, KWIC, and KWOC indexing.

Instrukcja indeksowania za pomocą słów kluczowych - CYTBIN

Opis rzeczowy sporządzany jest w trybie indeksowania wyszczególniającego – należy dobierać słowa kluczowe najbardziej trafnie i adekwatnie charakteryzujące zawartość indeksowanych artykułów (stopień szczegółowości charakterystyki wyszukiwawczej dokumentu powinien być dostosowany do jego treści, bez zbędnych uogólnień).

Indeksowanie odbywa się metodą **indukcyjną** – jako słowa kluczowe wykorzystywane są terminy występujące w tekstach artykułów rejestrowanych w bazie CYTBIN[1]. Dopuszcza się wprowadzenie terminów spoza tekstu indeksowanego dokumentu lub zastępowanie niektórych terminów użytych w tekście bardziej odpowiednimi – w przypadku, gdy wyrażenia użyte przez autora są niewłaściwe z powodu niejasności ich znaczenia, niepoprawności terminologicznej oraz małego rozpowszechnienia w nauce.

Encyklopedia językoznawstwa ogólnego

W informacji naukowej wyraz lub wyrażenie wybrane z tytułu lub tekstu dokumentu w celu najbardziej ogólnego scharakteryzowania jego treści. W wypadku automatycznego indeksowania dokumentu za słowo kluczowe uważa się te słowa, które najczęściej są używane w tekście i nie wchodzą do sporządzonej uprzednio listy słów niekluczowych (obejmującej najczęstsze słowa o znaczeniu gramatycznym)

Podsumowanie albo zamiast podsumowania

Wiesław Babik w swoim obszernym opracowaniu na temat słów kluczowych przytacza wiele definicji tego terminu (2010, 24-32). Warto przywołać w tym miejscu podsumowujące przeprowadzoną przez autora analizę zestawienie powtarzających się we wszystkich definicjach elementów (32):

1. *Słowo kluczowe to wyraz lub połączenie wyrazowe, czyli wyrażenie języka.*
2. *Słowo kluczowe to wyraz lub grupa wyrazów o ustalonym konwencjonalnie znaczeniu, odpowiadającym pojęciu, z czego powinna wynikać ostrość granicy semantycznej, a więc jednoznaczność słowa kluczowego.*
3. *Słowo kluczowe należy do określonej dziedziny wiedzy lub działalności praktycznej. Najczęściej wymienia się naukę i technikę.*
4. *Niektóre definicje podkreślają fakt nadawania wyrazom języka ogólnego specjalnych znaczeń. Nie precyzuje się jednak konstrukcji stanowiących połączenia wyrazowe, które są wyjątkowo częstym sposobem wzbogacania zasobu słów kluczowych.*

[1] Bibliograficzna baza danych Instytutu Bibliotekoznawstwa i Informacji Naukowej Uniwersytetu Śląskiego

5. *W poszczególnych definicjach wymienia się zwykle takie funkcje słów kluczowych, jak funkcja etykietyzacji dokumentu/tekstu, funkcja imperatywna oraz funkcja klucza wyszukiwawczego. Dzięki tym funkcjom stają się one środkiem dostępu do treści/zawartości dokumentu.*

Wielokrotnie pojawia się też sformułowanie, że słowa kluczowe to słowa szczególnie ważne. Można to rozumieć dwojako: albo jako ważność wewnętrzną, "wyznaczoną przez miejsce, jakie pełni dane pojęcie w całokształcie aparatury pojęciowej tekstu" (Babik, 2010, 33), albo jako ważność zewnętrzną - z punktu widzenia odbiorcy tekstu. W tym drugim przypadku ciężko o obiektywizm (wszystko może się okazać ważne), w tym pierwszym zaś powinno się dać odnaleźć pewne wskazówki (np. na podstawie metadanych dokumentu), ale pytanie czy w kontekście ważności zewnętrznej takie poszukiwania mają sens.

Słowa kluczowe w dostępnych korpusach tekstowych

NKJP

Teksty korpusu są również klasyfikowane pod względem tematycznym. W tym wypadku przyjmujemy bez zmian dwie klasyfikacje Biblioteki Narodowej, a mianowicie kod Uniwersalnej Klasyfikacji Dziesiętnej i **system haseł tematycznych**[2]. Służą one do kontroli zróżnicowania tematycznego korpusu, a także w niektórych wypadkach stanowią cenną wskazówkę co do przynależnosci tekstu do danego typu.

Zawartość elementu <profileDesc> różni się natomiast od odpowiadającego mu elementu w nagłówku korpusu, gdyż zawiera wyłacznie jeden element, <textClass>. W elemencie tym znajdują się informacje dotyczące klasyfikacji tekstu według taksonomii przyjętych w NKJP (zob. rozdz. 2[3]). Na przykład zawartość elementu <profileDesc> dla powieści Manueli Gretkowskiej *Namiętnik* wygląda następujaco:

```
<profileDesc>
 <textClass>
  <classCode scheme="#ukd">821.162.1-3</classCode>
  <keywords scheme="#bn">
   <list>
    <item>Opowiadanie polskie -- 20 w.</item>
   </list>
  </keywords>
  <catRef scheme="#taxonomy-NKJP-type"
    target="#typ_lit_proza"/>
  <catRef scheme="#taxonomy-NKJP-channel"
    target="#kanal_ksiazka"/>
 </textClass>
</profileDesc>
```

[2] Por. http://bn.org.pl/dla-bibliotekarzy/jhp-bn/slownik
[3] Tu paragraf wyżej.

Mogą się tu znajdować odnośniki do czterech klasyfikacji: dwóch zewnątrznych w stosunku do NKJP (wspomniana powyżej Uniwersalna Klasyfikacja Dziesiętna oraz klasyfikacja Biblioteki Narodowej; zob. #bn) i dwóch wewnętrznych (typ tekstu: #typ_lit_proza oraz kanał: #kanal_ksiazka).

Inne korpusy

BNC

In the first release of the BNC, most texts were assigned a set of descriptive keywords, tagged as <term> elements within the <keywords> element. These terms were not taken from any particular descriptive thesaurus or closed vocabulary; the words or phrases used are those which seemed useful to the data preparation agency concerned, and **are thus often inconsistent or even misleading.** They have been retained unchanged in the present version of the BNC, pending a more thorough revision. In the World (second) Edition this set of keywords was complemented for most written texts by a second set, also tagged using a <keywords> element, but with a value for its *source* attribute of COPAC, indicating that the terms so tagged are derived from a different source. The source used was a major online library catalogue service (see http://www.copac.ac.uk). Like other public access catalogue systems, COPAC uses a well-defined controlled list of keywords for its subject indexing, details of which are not further given here.

Keyword Search

Use this if you want to search for a list of **written** BNC files which match a particular keyword, or set of keywords, that you have in mind. For example, if you want to create a subcorpus of **published** texts which are to do with the general subject of India, simply type India as your keyword and specify the COPAC[1] library catalogue.

Searching by **COPAC keywords** is the default, and restricts you to published, written texts. If you click on the drop-down menu, you will see that there is another type of *keywords,* labelled "descriptive keywords - BNC1 release". This refers to the original set of keywords entered by the compilers of version 1 of the BNC. In general, these "BNC1 keywords" are less systematic and useful than COPAC keywords because they do not meet professional library cataloguing standards.

Korpusy.net (słowniczek)

Słowo kluczowe - W lingwistyce korpusowej, słowo występujące **w tekście** (lub korpusie) częściej niż sugerowałaby to frekwencja danego wyrazu w danym języku w ogóle.

Komentarz

Oparcie na systemie haseł tematycznych Biblioteki Narodowej, czy na innym tego typu systemie (np. wspomniany COPAC) ma swoje niewątpliwe zalety w postaci kontrolowanego słownika. Ta zaleta może jednak okazać się wadą, np. ze względu na możliwą nieaktualność takiego słownika. Utożsamienie jednak słownika haseł tematycznych ze słowami kluczowymi jest jednak niebezpieczne z jednego jeszcze względu - są to jednostki z innego poziomu. Jak pisze Babik, słowa kluczowe nie są synonimami tematów czy deskryptorów (jednostek niższego

rzędu), choć mogą pełnić ich funkcję, “słowa kluczowe odnoszą się do tekstu; temat tworzy się dla grupy dokumentów” (Babik 2010, 28), a co za tym idzie: ”deskryptor jest słowem kluczowym w dziedzinie, a słowo kluczowe istnieje dzięki tekstowi”. W tym kontekście szczególnie uzasadnione wydaje się ekstrakcyjne podejście do wydobywania słów kluczowych.

Przyjęte rozwiązania w KPWr - wytyczne do znakowania

Podstawowe zasady

1. Interesują nas takie słowa, które w pewien sposób są charakterystyczne dla tego tekstu. Staramy się nie przyjmować perspektywy użytkownika wyszukiwarki, a raczej badacza, który próbuje za pomocą słów kluczy opisać specyfikę danego tekstu.
2. Słowo kluczowe = pojedyncze słowo lub fraza.
3. Frazy nie muszą spełniać wymagań stawianych jednostkom wielowyrazowym (JW) opisanym, np. w tym dokumencie
4. Sprowadzamy słowo lub frazę do formy bazowej zgodnie z wytycznymi do lematyzacji fraz (po uwzględnieniu pozostałych punktów)[4]. Przyjmujemy, że liczba (gramatyczna) po lematyzacji powinna być taka sama jak liczba w tekście, np. *rozbojów* sprowadzamy do *rozboje*, a *rozbojem* do *rozbój*.
5. Bierzemy pod uwagę tylko przymiotniki i rzeczowniki (lub frazy przymiotnikowe i rzeczownikowe)
6. Jeśli uznamy, że kluczowa jest tutaj jakaś sytuacja nazwana za pomocą czasownika, możemy utworzyć od tego czasownika gerundium, jednak umieszczamy je w nawiasie, jeżeli jest ich kilka wtedy każde umieszczamy w osobnym nawiasie oddzielonym przecinkiem.
7. To co uznamy za frazę (poza przedstawionymi poniżej wyjątkami) jest niepodzielne, np. *Uniwersytet Adama Mickiewicza* zostaje jako *Uniwersytet Adama Mickiewicza*, nie powinniśmy wskazywać dodatkowo takich słów kluczowych jak: *Uniwersytet, Adam, Mickiewicz, Adam Mickiewicz*.
8. Możemy w drodze wyjątku uznać za słowa kluczowe składniki fraz (mogą to być zarówno pojedyncze słowa, jak i frazy np. *figurka Matki Boskiej* → *figurka*, *Matka Boska*) po lematyzacji każdego słowa kluczowego. Możemy tak zrobić w przypadku spełnienia jednego z poniższych warunków:
    a. składnik pojawia się w tekście samodzielnie
    b. składnik pojawia się w tekście jako element innych fraz
9. W przypadku synonimów wybieramy ten bardziej popularny lub ten który ma większą frekwencję w tekście.
10. W przypadku tekstu, gdzie same nazwy własne stanowią 10 lub więcej pozycji ograniczając w ten sposób możliwość anotacji innych słów kluczowych, wybieramy imiona i nazwiska najbardziej popularne (np. w sprawozdaniu sportowym).

[4] Używanie liczby mnogiej na oznaczenie kategorii - Le Deuff 2006:
Des règles de bonne indexation par tags ont été édictées:
(...)
employer le pluriel pour définir des catégories. Le pluriel est plus approprié car la catégorie peut contenir différentes variations;

Aspekt techniczny znakowania słowami kluczowymi

1. Dokumenty anotujemy z wykorzystaniem systemu Inforex (nlp.pwr.wroc.pl/inforex)
2. Z tekstu wyróżniamy maksymalnie 7 słów kluczowych
3. Słowa kluczowe wpisujemy w odpowiedniej rubryce “Edit metadata”. Aktualnie nie planuje się wprowadzić oznaczania w tekście.
4. Słowa kluczowe oddzielamy od siebie przecinkiem
5. Wielkie litery stosujemy zgodnie z zapisem w tekście - jak jest błędnie użyta wielka/małą literą, to powtarzamy ten błąd (wielka/mała litera może być ignorowana na etapie postprocessingu)
6. Kolejność słów kluczowych nie ma znaczenia
7. Jeśli wydaje nam się, że słowo, które można uznać za kluczowe, jest niepoprawnie zapisane (błąd gramatyczny, litrówka, źle odmieniony itp.) tak, że może dojść do nieporozumień, wtedy nie włączamy go do słów kluczowych.

Rozwiązania szczegółowe i najczęstsze wątpliwości

Jaki poziom ogólności słów kluczowych jest właściwy?

Stopień szczegółowości charakterystyki wyszukiwawczej dokumentu powinien być dostosowany do jego treści, bez zbędnych uogólnień. Jeśli np. dokument dotyczy biblioteki akademickiej, należy użyć słowa kluczowego „Biblioteka akademicka”, a nie szerszego „Biblioteka”. (CYTBIN)

Staramy się wykorzystywać sformułowania maksymalnie precyzyjne. Wynika to z faktu, że SI jest w stanie z informacji szczegółowej wyciągnąć również informację natury ogólniejszej, w drugą stronę proces jest dużo bardziej zawodny. Zatem jeżeli pojęcia, do których referują leksemy rozważane jako kandydaci na słowa kluczowe, pozostają ze sobą w relacji hipo - hiperonimii, wybieramy kandydata o precyzyjniejszym znaczeniu - hiponim (sytuację tę ilustruje przykład z CYTBIN).

Nazwy własne

Nazwy własne mogą być słowami kluczowymi w takim samym stopniu jak nazwy pospolite - na nasze potrzeby nie nadajemy im żadnego specjalnego statusu. Oznacza to, że wykorzystujemy nazwy własne do opisu słowami kluczami wtedy, kiedy uznamy, że w nazwy te celnie charakteryzują dany tekst - stanowią o jego specyfice i odróżniają go od innych tekstów. Sam fakt, że nazwisko Baracka Obamy pojawiło się w danym artykule to jeszcze za mało, by uznać je za charakterystyczne i nadać mu status słowa klucza. Dopiero gdy z treści wynika, że Barack Obama jest jej istotnym i specyficznym elementem, rozważamy to nazwisko jako słowo kluczowe.

Skróty

Jeżeli w tekście występuje pełna nazwa oraz skrót, wprowadzamy jako słowa kluczowe obie te formy w następującym zapisie: Uniwersytet Jagielloński [UJ], Polskie Koleje Państwowe [PKP], Korpus Politechniki Wrocławskiej [KPWr]. Wpisujemy zatem pełną nazwę a po niej w nawiasie kwadratowym adekwatny skrót. Pełnej nazwy i jej skrótu nie rozdzielamy przecinkiem.

Jeżeli w tekście występuje tylko pełna nazwa instytucji albo tylko skrót tej nazwy, uznajemy za słowo kluczowe adekwatnie albo samą pełną nazwę, albo jej skrót.

Ilość słów kluczowych

Z tekstu wyróżniamy maksymalnie 7 słów kluczowych, co oznacza, że może być ich mniej, ale nie więcej niż 7.

Częstą sytuacją jest konieczność zredukowania ilości potencjalnych kandydatów na słowa kluczowe do siedmiu. Możemy w takim przypadku zastosować kilka procedur:

1. **Zamiast fraz głowa**
   Jeżeli kandydatami są przykładowe frazy: *sąd federalny, sąd stanowy, sąd wyższej instancji* możemy jako słowo kluczowe uznać głowę tych fraz, a więc w tym przypadku jednostkę *sąd.* Jest to wyjątek od reguły mówiącej o tym, że staramy się wyznaczać możliwie najbardziej precyzyjne słowa kluczowe.
2. **Zamiast fraz fraza najczęściej powtarzająca się oraz głowa**
   Jeżeli w tekście pojawia się bardzo często jakaś jednostka jako element jednostki wielowyrazowej lub częstego i stabilnego połączenia wyrazów (kolokacji) (np. 10 wystąpień w tekście frazy *samochód osobowy*) oraz rzadziej jako element innych fraz (np. po trzy wystąpienia fraz: *samochód ciężarowy*, *samochód sportowy*, *samochód opancerzony*), możemy zdecydować się na użycie metody pośredniej. Uznajemy wtedy, że słowami kluczami jest fraza, której ilość wystąpień dominuje w tekście oraz głowa tej frazy, która powtarzana jest również w innych połączeniach wyrazowych. W tym przypadku słowa kluczowe byłyby następujące: *samochód osobowy, samochód*.

Wyliczenia

Jeżeli w analizowanym tekście znajdują się wyliczenia (np. *W naszym gospodarstwie hodowano wiele zwierząt: krów, kur, owiec i świń*) i potencjalny kandydat na słowo kluczowe jest elementem tego wyliczenia, należy sprawdzić, czy jednostka ta pojawia się jeszcze w tekście (*Ale krowy były z nich najważniejsze. Nie tylko dlatego, że ojciec hodował je jako jedyny na wsi. Przede wszystkim krowy pozwoliły rodzinie przeżyć wojnę)*. Jeżeli element wyliczenia powtarza się i spełnia kryterium istotności - specyficzności, możemy go użyć jako słowa kluczowego.

Słowa kluczowe a słowa o wysokiej frekwencji w dokumencie

Nie stawiamy znaku równości pomiędzy słowami kluczowymi a słowami o wysokiej frekwencji w tekście[5]. Wynika to z przesłanek gramatycznych i tekstowych. Po pierwsze bowiem, należy pamiętać, że interesują nas przede wszystkim przymiotniki i rzeczowniki (ew. frazy przymiotnikowe i rzeczownikowe) oraz w wyjątkowych sytuacjach czasowniki sprowadzone do formy gerundialnej. Po drugie zaś, kandydat na słowo kluczowe niezależnie od swojej wysokiej frekwencji musi spełniać kryterium istotności. Przykładowo w tekście biograficznym, którego tematem będą losy dawnego oficera możemy spodziewać się wysokiej frekwencji takich jednostek: *dawny, były* (w znaczeniu *były oficer*), *waleczny, odważny, przeniesiony, słynny, oddział, brygada, dywizja*. Jednostki te są prawdopodobnie ważne z perspektywy wewnętrznej struktury tekstu oraz tematyki, którą porusza. Może się jednak okazać, że nie są to jednak jednostki charakterystyczne, wyróżniające ten tekst od innych (każda notatka biograficzna o byłym oficerze będzie zawierała podobny zestaw wyrazów o wysokiej frekwencji). Może się okazać, że słowo kluczowe to wyraz o jednorazowym wystąpieniu (np. w tym dokumencie słowo kluczowe *zgon* występuje tylko raz, ale ze względu na swą wartość informacyjną zostało uznane za właściwego kandydata. Przeciwna decyzja została podjęta względem słowa *tuba,* które występuje w tekście dwukrotnie, ale nie jest aż tak charakterystyczne dla tekstu. Na jego miejsce wprowadzono bardziej informatywną frazę o jednorazowym wystąpieniu: *aparatura umożliwiająca sztuczne odżywianie*).

Słowo charakteryzujące

Wyrazy charakteryzujące nie są tożsame z jednostkami o najwyższej frekwencji. Po pominięciu wszelkich spójników, partykuł, zaimków i innych wyrazów, które nie mogą być słowami kluczowymi, może okazać się, że jednostki o najwyższej frekwencji będą determinowane przez gatunek i temat tekstu. Oznacza to, że najczęstszymi jednostkami dla artykułów biograficznych będą np.: *urodzić się, umrzeć, studiować, uniwersytet, szkoła, praca, pracować, mąż, żona, dziecko, grób, służba*. W przypadku tekstu o takiej tematyce (biograficznej) powinniśmy szukać wyrazów, które:

- są w tekście
- charakteryzują meritum tekstu
- mają niską frekwencję względną (a więc w odniesieniu do danego gatunku i tematu tekstu)
- wyróżniają ten tekst od innych w podobnym temacie
- wyróżniają ten tekst od tekstów zbliżonych gatunkowo, ujętych w danym podkorpusie

W tekście biograficznym o Józefie Piłsudskim mogą to być zatem jednostki: *Józef Piłsudski, polityk, Naczelnik Państwa, marszałek, Legiony, odzyskanie niepodległości, przewrót*

[5] Innym problemem jest relacja pomiędzy słowami kluczowymi a stosunkiem frekwencji danego słowa w dokumencie (lub wyspecjalizowanym korpusie) do frekwencji tego słowa w korpusie ogólnym. To zagadnienie zostanie omówione w podrozdziale Słowo charakteryzujące.

majowy, dwudziestolecie międzywojenne, sanacja. W haśle wikipedii jednostki *polityk, sanacja czy przewrót majowy* mają zdecydowanie niższą frekwencję niż *rząd*, derywaty od *Warszawy* i *Polski.* Jednostki te odróżniają jednak w większym stopniu treść notki biograficznej nt. Józefa Piłsudskiego od innych notek poświęconych osobistościom znanym w życiu publicznym.

Na gruncie KPWr możemy przeanalizować przykład:
http://nlp.pwr.wroc.pl/inforex/index.php?page=report&corpus=7&subpage=metadata&id=107258

Jest to tekst o wyraźnym temacie ogólnym (polityka) i szczególnym (formowanie się koalicji rządowej PiS z innymi partiami). Proponowane słowa kluczowe to:

- Jarosław Kaczyński (ponieważ jego postać jest wielokrotnie przywoływana oraz streszczane są jego wypowiedzi),
- Prawo i Sprawiedliwość [PiS] (ponieważ to ta partia jest głównym aktantem czynności polegającej na formowaniu się koalicji rządzącej),
- koalicja rządowa (ponieważ to właśnie koalicja jest celem działań opisanych tekście),
- pakt stabilizacyjny (środek do zawiązania koalicji), parlament (ponieważ tłem działań jest parlament),
- Platforma Obywatelska [PO] (ponieważ postawa tej partii w znacznej mierze warunkuje powstawanie koalicji - przynajmniej zdaniem J. Kaczyńskiego),
- wybory (realna alternatywa w przypadku niepowodzenia działań związanych z formowaniem koalicji).

Analizowany tekst jest już trudniejszy niż wcześniej omawiany przykład notki biograficznej. Wynika to z mniej wyraźnych granic gatunkowych, jakim podlega treść wikinewsów oraz wielości tematów poruszanych w Wikinewsach. Trudniej jest zatem ustalić zakres typowego słownictwa dla tego podkorpusu - słownictwa, które nie powinno być uwzględniane podczas wyznaczania słów kluczowych. Sądzimy, że na listę “zakazanych” wyrazów powinny być wciągnięte te jednostki, które jesteśmy skłonni uznać za ogólny (ale już nie szczegółowy) temat tekstu. W przypadku tego tekstu “zakazanymi” jednostkami mogłyby być: *polityka* i *demokracja*.

Stoplista

Co nie może być słowem kluczowym?

Naszym celem jest stworzenie tzw. stoplisty, a więc listy słów uznanych za nieznaczące. Przykład takiej stoplisty znajdziemy pod adresem:
http://www.webpageanalyse.com/dev/stopwords/pl

Słowa trafiają na stoplistę z dwóch głównych powodów: zbyt częste pojawianie się w tekście oraz kategoria gramatyczna lub leksykalna. W tym miejscu skupimy się na tym drugim aspekcie.

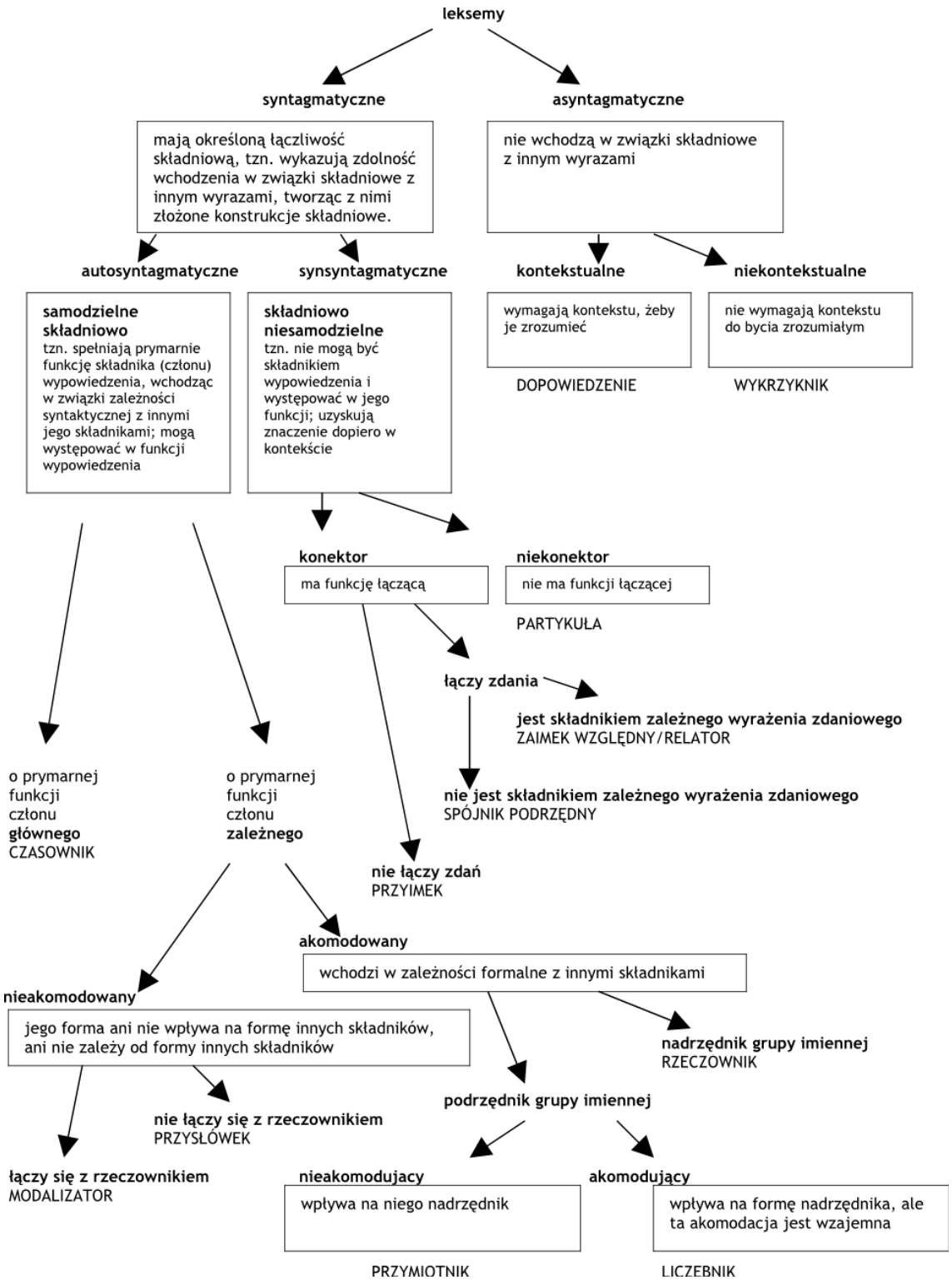
Słowa kluczowe a części mowy

Nie bez znaczenia dla wyznaczania słów kluczowych jest problem klasyfikacji słownictwa na części mowy. Kierunek poniższego przeglądu wyznacza teza, iż jednostkę, która może pełnić funkcję słowa kluczowego, należy traktować jako niezależną od innych leksykalnych jednostek danego języka (Babik 2010, 60)

Pomimo oczywistych mankamentów semantycznej klasyfikacji na części mowy nie do przecenienia jest właściwe temu podziałowi odróżnienie leksemów autosemantycznych od leksemów synsemantycznych, a więc współznaczących (por. Laskowski 1998, 53-54). Do tej drugiej grupy należą przyimek, spójnik i partykuła. Nie mają one pełnego znaczenia i z tego względu byłyby bezwartościowe jako słowa kluczowe[6]. Leksemy niepełnoznaczne klasyfikuje się biorąc pod uwagę zwykle kryteria syntaktyczne. Osobny problem stanowią zaimki, które są elementami opozycyjnymi w stosunku do innych tylko pod względem semantycznym. Są one znakami o charakterze indeksowym, a więc wskazuje na coś, co jest w jego bezpośredniej odległości. Ich sens zależny jest od kontekstu (rozumianego jako sąsiedztwo tekstowe lub sytuacja), więc w oderwaniu od niego zdecydowanie spada ich wartość informacyjna, co z kolei przekłada się na niską użyteczność z punktu widzenia ekstrakcji słów kluczowych. Zaimki stanowią klasę niejednorodną pod względem cech morfologicznych czy składniowych i z tego względu są pomijane lub rozdzielane do innych klas w podziałach opartych na właściwościach innych niż semantyczne.

Poszukiwania kompletnej, nieheterogenicznej i obiektywnej klasyfikacji na części mowy idą jednak w kierunku analizy cech formalnych leksemu. Na takich kryteriach oparte są morfologiczne definicje części mowy, jednak to kryteria syntaktyczne uważane są za jedyne, które pozwalają na skonstruowanie opartej na jednolitej zasadzie i wewnętrznie niesprzecznej klasyfikacji słownictwa na części mowy (Laskowski 1998, 55). Jedną z propozycji podziału prezentuje Laskowski (1988, 59), opierając się na opracowaniu Grochowskiego (1986). Jest to klasyfikacja o charakterze funkcjonalnym - wychodzi z analizy związków syntaktycznych, w jakie poszczególne leksemy mogą wchodzić z innymi elementami wypowiedzenia:

[6] Z wyłączeniem ich interpretacji w planie wyrazowym, czyli w przypadku brania pod uwagę nie wartości semantycznej, lecz funkcji wyrazu jako kategorii formalnej, co jednak wydaje się marginalnym problemem.


leksemy
syntagmatyczne
asyntagmatyczne
mają określoną łączliwość składniową, tzn. wykazują zdolność wchodzenia w związki składniowe z innym wyrazami, tworząc z nimi złożone konstrukcje składniowe.
nie wchodzą w związki składniowe z innym wyrazami
autosyntagmatyczne
synsyntagmatyczne
kontekstualne
niekontekstualne
samodzielne składniowo
tzn. spełniają prymarnie funkcję składnika (członu) wypowiedzenia, wchodząc w związki zależności syntaktycznej z innymi jego składnikami; mogą występować w funkcji wypowiedzenia
składniowo niesamodzielne
tzn. nie mogą być składnikiem wypowiedzenia i występować w jego funkcji; uzyskują znaczenie dopiero w kontekście
wymagają kontekstu, żeby je zrozumieć
DOPOWIEDZENIE
nie wymagają kontekstu do bycia zrozumiałym
WYKRZYKNIK
konektor
ma funkcję łączącą
niekonektor
nie ma funkcji łączącej
PARTYKUŁA
łączy zdania
jest składnikiem zależnego wyrażenia zdaniowego
ZAIMEK WZGLĘDNY/RELATOR
nie jest składnikiem zależnego wyrażenia zdaniowego
SPÓJNIK PODRZĘDNY
o prymarnej funkcji członu głównego
CZASOWNIK
o prymarnej funkcji członu zależnego
nie łączy zdań
PRZYIMEK
akomodowany
wchodzi w zależności formalne z innymi składnikami
nieakomodowany
jego forma ani nie wpływa na formę innych składników, ani nie zależy od formy innych składników
nadrzędnik grupy imiennej
RZECZOWNIK
podrzędnik grupy imiennej
nie łączy się z rzeczownikiem
PRZYSŁÓWEK
łączy się z rzeczownikiem
MODALIZATOR
nieakomodujacy
wpływa na niego nadrzędnik
akomodujący
wpływa na formę nadrzędnika, ale ta akomodacja jest wzajemna
PRZYMIOTNIK
LICZEBNIK

Pewne klasy leksemów są nieprzydatne jako słowa kluczowe ze względu na to, że nie wchodzą w relacje z innymi elementami tekstu (asyntagmatyczne) albo ze względu na to, że nie mogą stanowić samodzielnych składników wypowiedzenia (synsyntagmatyczne). Potwierdza tę tezę charakterystyka języka słów kluczowych. Wiesław Babik pisze, iż "wszystkie słowa kluczowe mają identyczną charakterystykę składniową; są samodzielne składniowo i wyszukiwawczo" (2010, 93).

Odrzucenie leksemów asyntagmatycznych i synsyntagmatycznych nie rozwiąże jednak problemu innej części mowy wyróżnianej ze względów semantycznych, a więc zaimków. We współczesnych klasyfikacjach, które opierają się właśnie na kryteriach morfologicznych i składniowych (i traktują je jako komplementarne), poszczególne klasy zaimków są "wchłaniane" przez inne kategorie (rzeczowniki, przymiotniki i przysłówki) ze względu na identyczne zachowanie w zdaniu. Podobnie rzecz się ma z ich cechami morfologicznymi - zaimki rzeczowne mają fleksję rzeczownikową, przymiotne przymiotnikową itd.

Dla przykładu poniżej przedstawiono klasy leksemów wyróżnione w korpusie IPI PAN (Woliński 2004, 48):

Tabela 1. Klasy leksemów i ich rozbicie na fleksemy. W wypadku leksemów złożonych z tylko jednego fleksemu wspólna nazwa leksemu i fleksemu zajmuje dwie kolumny tabeli.

leksem	fleksem	ozn.
rzeczownik	rzeczownik	subst
	forma deprecjatywna	depr
przymiotnik	przymiotnik	adj
	przymiotnik przyprzymiotnikowy	adja
	przymiotnik poprzyimkowy	adjp
przysłówek odprzymiotnikowy i/lub stopniowalny		adv
liczebnik		num
zaimek nietrzecioosobowy		ppron12
zaimek trzecioosobowy		ppron3
zaimek		siebie
czasownik	forma nieprzeszła	fin
	forma przyszła czasownika	bedzie
	aglutynant czasownika	aglt
	pseudoimiesłów	praet
	rozkaźnik	impt
	bezosobnik	imps
	bezokolicznik	inf
	imiesłów przys. współczesny	pcon
	imiesłów przys. uprzedni	pant
	odsłownik	ger
	imiesłów przym. czynny	pact
	imiesłów przym. bierny	ppas
czasownik typu (forma teraźniejsza)		winien
predykatyw		pred
przyimek		prep
spójnik		conj
kublik (partykuło-przysłówek)		qub
ciało obce nominalne		xxs
ciało obce luźne		xxx

W powyższej klasyfikacji wyróżniono jedynie zaimki zaliczane tradycyjnie do klasy osobowych i zwrotnych. W podziale zaprezentowanym przez Laskowskiego zaś wyróżniono zaimki tradycyjnie nazywane względnymi.

Na problem wartości poszczególnych części mowy dla ekstrakcji słów kluczowych można jednak spojrzeć z innej strony - pozytywnej, a więc spróbować odpowiedzieć nie na pytanie: co nie może być słowem kluczowym, ale na pytanie: co tym słowem być może.

Z tego względu na specyficzną funkcję słów kluczowych (informacyjną, identyfikacyjną, interpretacyjną, wyszukiwawczą, koordynacyjną i organizacyjną) z informacyjnego punktu widzenia powinny one być słowem znaczącym i spełniać określone wymogi formalne. Jak twierdzi Babik, w języku polskim warunek ten spełniają rzeczowniki i przymiotniki w mianowniku liczby pojedynczej lub tzw. pluralia tantum (2010, 33). We współcześnie stosowanych językach słów kluczowych ten zakres został jeszcze bardziej zawężony - funkcję słowa kluczowego może pełnić rzeczownik oraz fraza rzeczownikowa. Taka była i jest konwencja etykietowania tekstu przy użyciu słow kluczowych. Kategorie rzeczownikowe bowiem jako etykiety donotują od razu całą wiązkę cech i silniej wiążą się z wyobraźnią, umożliwiając tworzenie odpowiednich obrazów mentalnych (Babik 2010, 61). Pozostaje pytaniem, czy i w jakim ewentualnie zakresie uznać, że słowa kluczowe mogą być wyrażane przez inne części mowy.

Znaczenie gramatyczne

Ze względu na swój specyficzny charakter z listy słów kluczowych powinny zostać wyłączone słowa o znaczeniu gramatycznym. W języku polskim takie znaczenie mają przede wszystkim morfemy o charakterze fleksyjnym i słowotwórczym. Taki charakter mają również słowa posiłkowe, a więc czasowniki lub wyrażenia czasownikowe, z którymi łączy się określona forma innego (tzw. pełnego) czasownika. Jest to przede wszystkim czasownik *być* (w funkcji tworzenia czasu przyszłego). Można przyjąć, że do grupy słów posiłkowych należą także łączniki (*copula*), a więc czasowniki pełniące funkcję wykładnika predykacji imiennej. W języku polskim oprócz słowa *być* byłyby to: *stawać/stać się, zostać, nazywać się, wydawać się, okazywać/okazać się* oraz *to* i *oto*.

Osobny problem stanowią wyrazy o charakterze modalnym, które łączą się z bezokolicznikiem, tworząc orzeczenia złożone, a więc czasowniki modalne, predykatywy i czasowniki fazowe. Nie można powiedzieć, że ich znaczenie ma tylko charakter gramatyczny. Z drugiej jednak strony należą one do sfery *modus* i nie są samodzielne. Z tego względu należałoby je również zaliczyć do słów niekluczowych.

Kategorie gramatyczne

Na liście powinny się znaleźć wszystkie części mowy poza rzeczownikami i przymiotnikami, a więc:

- przysłówek - adv
- liczebnik - num
- zaimki - ppron12, ppron3, siebie
- czasowniki - fin, bedzie, aglt, praet, impt, imps, inf, pcon, pant, ger, pact, ppas, winien, pred

- przyimki - prep
- spójniki - conj
- kubliki - qub
- ciała obce - xxs, xxx

Z listy słów niekluczowych warunkowo można wyłączyć gerundia (ger) i imiesłowy przymiotnikowe (pact, ppas).

Słowa niekluczowe

Na liście słów niekluczowych znajdują się następujące słowa:

Zaimki
co
cokolwiek
coś
cośkolwiek
cóż
czyj
czyjkolwiek
czyjś
dlaczego
dlaczegoż
dlań
dokąd
dokądkolwiek
dokądś
dokądże
doń
donikąd
dotąd
gdy
gdzie
gdziekolwiek
gdzieniegdzie
gdzieś
gdzież
ich
inaczej
inny
ja
jak
jakby
jaki
jakikolwiek
jakiś
jakiśkolwiek
jakiż
jakkolwiek
jako
jakoś
jakże
jakżeby
jeden

jego
jej
każdy
kędy
kędyś
ki
kiedy
kiedykolwiek
kiedyś
kiedyż
kto
ktokolwiek
którędy
którędykolwiek
którędyż
który
którykolwiek
któryś
któryż
ktoś
ktośkolwiek
któż
mój
my
nań
nasz
nic
niczyj
niejaki
niejeden
niektóry
niektórzy
nigdy
nigdzie
nikt
odeń
odkąd
on
oń
ona
one
oni
ono
ów
owaki
owdzie
ówdzie
pokąd
pokądże
poń
potem
przedtem
przenigdy
przezeń
sam
se
siaki

się
siebie
skąd
skądciś
skądinąd
skądkolwiek
skądś
skądsiś
skądyś
skądże
stąd
stamtąd
swój
tak
taki
takiż
tako
takowy
takoż
tam
tamtędy
tamtejszy
tamten
tamże
tędy
ten
tenże
to (bez względu na część mowy)
toto
toż
tu
tutaj
tuż
twój
ty
wasz
weń
wszelaki
wszelki
wszyscy
wszystek
wszystko
wtedy
wy
żaden
zań
zeń

Pozostałe słowa niekluczowe:
cały
dziś (bez względu na część mowy)
kierunek
lat
pan
pani
rok
dzisiaj (bez względu na część mowy)
jutro (bez względu na część mowy)
ń (ppron3:sg:gen.acc:m1.m2.m3:ter:nakc:praep)
oto
wczoraj (bez względu na część mowy)

Bibliografia

1. Wiesław Babik (2010), *Słowa kluczowe*, Kraków: Wydawnictwo Uniwersytetu Jagielońskiego
2. Winnie Cheng (2012), *Exploring Corpus Linguistics. Language in Action,* London: Routlege (s. 70-72)
3. Olivier Le Deuff (2006), *Folksonomies: les usagers indexent le web*, "Bulletin des Bibliothèques de France", vol. 51, no 4, s. 66-70 [on-line] http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002
4. *CYTBIN. Instrukcja indeksowania za pomocą słów kluczowych* [online] http://ibin.us.edu.pl/cbn/slowa_kluczowe.pdf
5. *Encyklopedia językoznawstwa ogólnego* (1999), red. Kazimierz Polański, Wrocław: ZNiO
6. *Encyklopedia współczesnego bibliotekarstwa polskiego* (1976), red. Karol Głombiowski, Bolesław Świderski, Helena Więckowska, Wrocław: ZNiO
7. Maciej Grochowski (1986),
8. *ISO 5963: 1985, Documentation -- Methods for examining documents, determining their subjects, and selecting indexing terms*
9. Roman Laskowski (1998), *Wyraz*, [w:] *Gramatyka współczesnego języka polskiego. Morfologia*, cz. 1, Grzegorczykowa R., Laskowski R., Wróbel H. (red.), wyd. 2, PWN, Warszawa.
10. *Narodowy Korpus Języka Polskiego* (2012), red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Warszawa PWN
11. *ODLIS Online Dictionary for Library and Information Science* [online] http://www.abc-clio.com/ODLIS/odlis_A.aspx
12. *Polish stopwords (137 words)*, *Webpageanalyse* [online] http://www.webpageanalyse.com/dev/stopwords/pl
13. *Reference Guide for the British National Corpus* (2007), red. Lou Burnard, Oxford University Computing Services, [online] http://www.natcorp.ox.ac.uk/docs/URG/
14. *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych* (2002), red Bożenna Bojar, Warszawa SBP
15. *Słownik terminologiczny informacji naukowej* (1979), red. Maria Dembowska, Wrocław:ZNiO
16. Marcin Woliński (2004), *System znaczników morfosyntaktycznych w korpusie IPI PAN*, "POLONICA XII", s. 39-54
17. *Wordsmih tool manual - Keywords* http://www.lexically.net/downloads/version6/HTML/index.html?keywords_info.htm