Wytyczne lematyzacji tekstowych wykładników słów kluczowych

(frazy pospolite i nazwy własne)

maj 2017

Michał Marcińczuk, Marcin Oleksy
Grupa G4.19
Katedra Inteligencji Obliczeniowej
Politechnika Wrocławska

na bazie Wytyczne do lematyzacji Adam Radziszewski, Marcin Oleksy, Jan Wieczorek [https://clarin-pl.eu/dspace/handle/11321/591]

Wprowadzenie

Lematyzacja fraz polega na przypisaniu wystąpieniom fraz w tekście ich **form podstawowych**. **Formą podstawową** frazy jest fraza, która mogłaby wystąpić jako hasło w słowniku albo na liście słów kluczowych.

Wytyczne dotyczą lematyzacji następujących kategorii fraz:

1. Frazy rzeczownikowe
2. Frazy czasownikowe
3. Nazwy własne
4. Przymiotniki pochodzące od nazw własnych

Ogólne założenia

Ogólne założenia dotyczące wszystkich kategorii:

1. **Wielkość liter** — wielkość liter lematu frazy nie wynika z położenia frazy w tekście (początek/środek zdania) tylko z przyjętego zapisu dla danej frazy. Szczegółowe przypadki zostały opisane w poniższej tabelce.
2. Skróty będące częścią frazy pospolitej należy rozwinąć do pełnej postaci.
3. Skrótów słowa pospolitego będącego częścią nazw własnej nie rozwijamy, np. [Uniwersytetu **im.** Adama Mickiewicza] → [Uniwersytet **im.** Adama Mickiewicza]

4. Skróty nazwy własnej rozwijamy do pełnej postaci, np. [Gorzowie **Wlkp.**] → [Gorzów **Wielkopolski**]
5. W przypadku istnienia kilku możliwych form zlematyzowanych należy wybrać tą, która jest częstsza na liście frekwencyjnej, np. lematem dla frazy “kosztów” może być “koszta” i “koszty”.

Rodzaj frazy	**Zapis lematu**	**Przykładowa fraza**	**Lemat**
Fraza pospolita zapisana kapitalikami	Całość małymi literami	ZNANE POSTACIE HISTORYCZNE	znane postacie historyczne
Skrótowiec będący nazwą własną	Dużymi literami z wyjątkiem spójników, zgodnie z zapisem skrótowców	PKP	PKP
Nazwa własna zapisana dużymi literami, która z reguły pisze się tylko z dużej litery	Zapis zgodny z postacią słownikową	POLSKA	Polska

Kolejne punkty opisują wytyczne lematyzacji dla poszczególnych kategorii fraz. Grupy fraz są rozłączne, co oznacza, że dana fraza należy dokładnie do jednej grupy i zgodnie z zasadami dla danej grupy należy ją zlematyzować.

Nazwy własne

1. Nazw własnych mających postać frazy czasownikowej nie sprowadzamy do gerundium, zachowujemy formę czasownika, np. [Będzie Dobrze] → [Będzie Dobrze]
2. Zachowujemy liczbę i rodzaj składników fraz rzeczownikowych i przymiotnikowych, np. [40 synami i 30 wnukami jeżdżącymi na 70 oślętach] → [40 synów i 30 wnuków jeżdżących na 70 oślętach], [Szybcy i wściekli] → [Szybcy i wściekli], [Dolnośląskiego] → [Dolnośląskie][1], ul. [Długiej] → ul. [Długa]
3. W przypadku rzeczownikowych nazw ulic zachowujemy przypadek, np. ul. [Kochanowskiego] → ul. [Kochanowskiego] (nie *Kochanowski!)

Frazy rzeczownikowe

1. Głowa frazy zostaje sprowadzona do mianownika, pozostałe składniki frazy zachowują składnię rządu lub zgody, np.

[1] Uwaga! http://sjp.pwn.pl/poradnia/haslo/Mazowieckie-mazowieckie;3019.html

    a. składnia rządu
       [czapkę *(kogo? czego?)* Adama] → [czapka *(kogo? czego?)* Adama]
    b. składnia zgody
       [niebieskiego balonika] → [niebieski balonik]
2. Liczba (pojedyncza/mnoga) zostaje utrzymana

Frazy czasownikowe

1. W sytuacjach, w których kandydat na słowo kluczowe jest orzeczeniem lub ośrodkiem równoważnika zdania, zamieniamy formę osobową czasownika (w przypadku orzeczeń) lub imiesłów przysłówkowy, bezokolicznik (w przypadku ośrodków równoważników zdań), np.
   Andrzej [*pisze listy*] → [*pisanie listów*]
   Andrzej narzeka na świat, [*pisząc listy*] → [*pisanie listów*]
2. na gerundium, argument biernikowy przybiera formę dopełniacza, np. [pisze *(kogo? co?)* listy] → [pisanie *(kogo? czego?)* listów]
4. Zachowujemy aspekt czasownika (dokonany lub niedokonany)

Przymiotniki od nazw własnych

1. Rodzaj przymiotnika zostaje sprowadzony do formy męskiej
2. Liczba gramatyczna przymiotnika zostaje sprowadzona do pojedynczej