

Słownosieć 3.2 - leksykalna sieć semantyczna języka polskiego i jej zastosowania

Poziom rozszerzony

Agnieszka Dziob



Politechnika Wroclawska

Katedra Inteligencji Obliczeniowej

Grupa Naukowa G4.19

agnieszka.dziob@pwr.edu.pl

Relacje synsetów (przykłady)



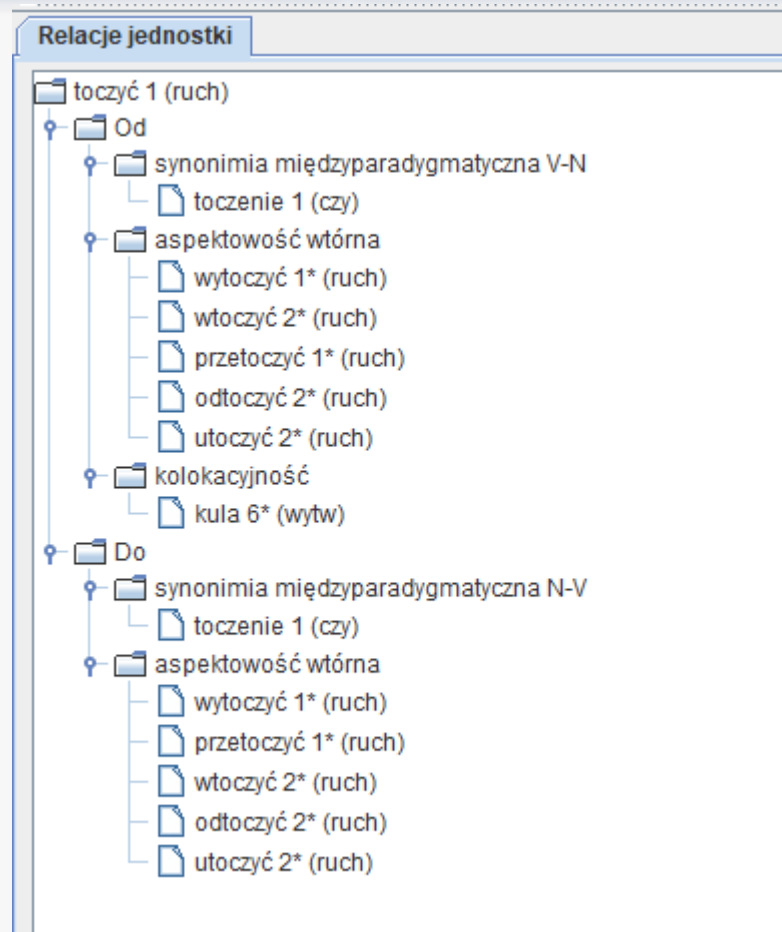
- Podstawowa: hiperonimia (odwrotna hiponimia) – wiąże z kategorią nadrzędną
- Relacja cząstkowości-całościowości: meronimia-holonimia dla rzeczowników
- Relacja mero-/holonimii taksonomicznej dla rzeczownika – wiąże drzewo taksonomii naukowej z drzewem wiedzy potocznej
- Relacja dla nazw własnych: typ-egzemplarz (oznaczenie nazw własnych: NP w komentarzu)
- Relacje opisu związków przyczynowo-skutkowych dla czasowników:
 - Wcześniej (przyczyna): uprzedniość i presupozycja
 - Jednocześnie: mero-/ i holonimia czasownikowa
 - Później (skutek): kauzacja (inchoatywność) i procesywność
- Relacje wielokrotności dla czasowników: dystrybutywność i iteratywność

- Wybrane relacje derywacyjne:
 - dla rzeczownika:
 - noun-noun: żeńskość, istota młoda, nacechowanie (ekspresywność, deminutywność)
 - noun-verb: rola (np. narzędzie drukarka „narzędzie, za pomocą którego ktoś drukuje”)
 - dla przymiotnika:
 - adj-verb: rola (np. wędrowny „taki, który wędruje”), predyspozycyjność (np. mieszkalny „taki, w którym można mieszkać”)
 - adj-noun: rola-materiał (dla przymiotników materiałowych, np. drewniany „taki, który jest zrobiony z drewna”)
 - stopniowanie – szczątkowy opis podyktowany względami praktycznymi (bardzo mało instancji)
 - dla przysłówka: stopniowanie, szczątkowy opis
 - dla czasownika:
 - verb-noun: zawieranie roli (np. nocować to „spędzać noc”)

Relacje jednostek



- Wybrane relacje derywacyjne:
 - aspektowość (dla czasownika; verb-verb):
 - czysta (pomiędzy perf i imperf nie ma zmiany znaczenia)
 - wtórna (głównie czasowniki tworzone od prefiksów znaczących)
 - dla wszystkich części mowy: derywacyjność (obrazuje mniej regularne procesy derywacyjne)



- Relacje niederywacyjne:
 - kolokacyjność: ok. 16 tys. kolokacji przypisanych do wszystkich części mowy; relacja kolokacyjności łączy konkretne znaczenia, nie słowa
 - relacje przeciwieństwa: antonimia (właściwa i komplementarna, konwersja)
 - antonimia komplementarna: X jest przeciwieństwem Y i żadne Z nie jest możliwe (np. starczyć i braknąć)
 - antonimia właściwa: X jest przeciwieństwem Y, ale jest możliwe jakieś Z (np. włączyć i wyłączyć; Z oznacza nie wykonywać żadnego działania); w tej grupie mieszczą się antonimy kulturowe (skonwencjonalizowane) a także znaczna część czasowników prefiksowanych
 - konwersja: X do Z ma się tak samo jak Y do Z, X i Y oznaczają zmianę perspektywy (np. kupić i sprzedać)


Powiązanie z innymi zasobami

MWELexicon, Word2Vec



- Słownik jednostek wielowyrazowych: MWELexicon
 - <https://clarin-pl.eu/dspace/handle/11321/274>
 - opis składniowy (WCCL) jednostek wyrazowych opisanych semantycznie w Słowsieci
 - przykład operatora WCCL:

```
<mwegroup class="adv" id="PrepAdjpFixAdv;s:B1,s:B2" name="PrepAdjp" type="fix">
  <condition>
    //z cicha
    and(
      inter(base[0], $s:B1), // z
      inter(orth[1], $s:B2), // cicha
      inter(class[0], {prep}),
      inter(class[1], {adj, adjp, subst}),
      setvar($Pos1, 0), setvar($Head, 1),
      setvar($Pos2, 1)
    )
  </condition>
```



```
<MWE base="po polsku">
  <var name="B1">po</var>
  <var name="B2">polski</var>
</MWE>
```

Powiązanie z innymi zasobami

MWELexicon, Word2Vec



- Opis semantyczny jednostek wielowyrazowych w Słownosieci
 - jednostki niekompozycyjne semantycznie, np. czerwona kartka (w Słownosieci oznaczone jako <##DD> w komentarzu)
 - jednostki niekompozycyjne składniowo, tzn. takie, których frekwencja w korpusie wskazuje na nieseparowalność i szyk ustalony, np. płytka posadzkowa (w Słownosieci oznaczane jako <##as1DD>, <##as2DD>, <##as3DD>)
 - terminy, np. prądnica tachometryczna (oznaczone jako <##DD> w komentarzu, znaczna część z nich zawiera podane źródła <##REF>)
 - niewielką ilość jednostek <##DS> uznanych mocą autorytetu koordynatora (funkcja pragmatyczna)
 - <##nDD> - połączenia kompozycyjne, zostaną usunięte ze Słownosieci (zebrane w oddzielnym zasobie częstych kolokacji),
 - bi- i trigrammy (niewielka ilość tetra- i pentagramów), części mowy: rzeczownik, przymiotnik, przysłówki, dla czasownika tylko bigrammy z „się”
- wytyczne dla lingwistów: http://clarin-pl.eu/wp-content/uploads/2015/05/Jednostki-wielowyrazowe_Procedura-sprawdzania-leksykalno%C5%9Bci-po%C5%82%C4%85cze%C5%84-wielowyrazowych.pdf)

Powiązanie z innymi zasobami

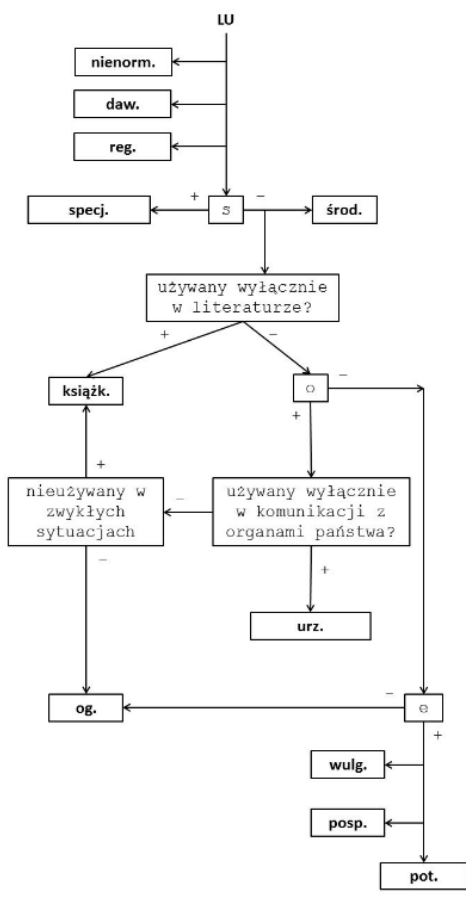
MWELexicon, Word2Vec



- Usługa sieciowa wykorzystująca powiązania dystrybucyjne pomiędzy jednostkami Słownosieci (metoda wektorowa):
 - <http://ws.clarin-pl.eu/w2vdemo.shtml>
- Lista wektorowa dla czasownika palić (nie ma usługi, proszę o kontakt)

	A	B	C	D
1	się	<u>qub</u>	<u>refl</u>	16541
2	.	<u>interp</u>	<u>punct</u>	7956
3	i	<u>conj</u>	<u>conjunct</u>	7419
4	,	<u>interp</u>	<u>punct</u>	6906
5	nie	<u>qub</u>	<u>neg</u>	5063
6	papieros	<u>subst</u>	<u>obj</u>	3217
7	,	<u>interp</u>	<u>conjunct</u>	2591
8	a	<u>conj</u>	<u>conjunct</u>	2307
9	że	<u>comp</u>	<u>comp_fin</u>	2011
10	światło	<u>subst</u>	<u>subj</u>	1537
11	on	<u>ppron3</u>	<u>obj</u>	1479
12	ale	<u>conj</u>	<u>conjunct</u>	1275
13	być	<u>aglt</u>	<u>aglt</u>	1030
14	który	<u>adj</u>	<u>subj</u>	980
15	zacząć	<u>praet</u>	<u>comp_inf</u>	693
16	-	<u>interp</u>	<u>punct</u>	689
17	jak	<u>adv</u>	<u>adjunct</u>	666
18	fajka	<u>subst</u>	<u>obj</u>	648
19	cygaro	<u>subst</u>	<u>obj</u>	588
20	i	<u>conj</u>	<u>obj</u>	575
21	ogień	<u>subst</u>	<u>subj</u>	574
22	on	<u>ppron3</u>	<u>obj_th</u>	550
23	już	<u>qub</u>	<u>adjunct</u>	545
24	być	<u>bedzie</u>	<u>aux</u>	539
25	lampa	<u>subst</u>	<u>subj</u>	450

- Wytyczne dla lingwistów Słowosieci: <http://clarin-pl.eu/wp-content/uploads/2017/06/Kwalifikator%C3%B3w-system.pdf>



- Szczegółowe zasady:

- synchroniczny opis języka (słownictwo daw. jest rzadkie, archaizmów nie ma)
- Nienorm. – tylko jeśli ma wysoką frekwencję w korpusie
- Środ. – ograniczone do języka młodzieżowego, żargonu grup zamkniętych (np. więziennej) i zawodowych, niezrozumiałego poza tymi grupami
- Reg. – tylko jeśli ma wysoką frekwencję w korpusie (Słowosieć nie opisuje systemowo odmian regionalnych)

Wykorzystanie WNLooma

Lexical units Synsets Candidates

Search type:
Lemma

Search:
[]

Lexicon:
All

Parts of speech:
All

Grammatical gender:
All

Style:
All

Domain:
All

Semantic Field:
All

S. Field Modifier:
All

Status:
All

Age:
All

Source:
All

Lexical Characteristic:
All

Etymology:
[]

Lexical units Synsets Candidates

Search:
[]

Lexicon:
All

Parts of speech:
All

Domain:
All

Relations:
All

Definition:
[]

Comment:
[]

Synset type

- All
- Normal
- Artificial

Limit to 500 elements

Długotrwały kontakt językowy i jego skutki leksykalne oraz semantyczne na przykładzie polskich zapożyczeń leksykalnych w jidysz - grant NCN (OPUS 11); kierownik prof. Ewa Geller (UW)



Dziękuję bardzo za uwagę

Zapraszam do kontaktu:
agnieszka.dziob@pwr.edu.pl