Options for automatic creation of dictionary definitions from corpora

Marie Stará, Vojtěch Kovář

Faculty of Arts & Natural Language Processing Centre, Faculty of Informatics, Masaryk University

December 3, 2016

Definition

- intensional – traditional definition using genus and differentia or a list of subsets
- extensional – by listing every member of a set or using ostensive definition (defining by pointing)
- synonym or antonym

Data

- nouns “příbor” (cutlery), “pes” (dog) and “bagr” (excavator)
- verbs “trpět” (suffer) and “chytit” (catch)
- adjectives “zádumčivý” (broody), “starý” (old) and “opilý” (drunk)
- conjugations “poněvadž” (because) and “nebo” (or)

Bagr / excavator

1. rýpadlo (digger)
2. plavidlo k bagrování (dredge)

Lemma	Translation	Score	Freq
rypadlo	digger	0.244	3,452
buldozer	bulldozer	0.225	6,743
nákladák	lorry	0.182	23,706
nakladač	traxcavator	0.179	10,753
rýpadlo	digger	0.159	1,263
jeřáb	derrick	0.146	31,472
traktor	tractor	0.141	63,830
kamión	lorry	0.131	11,206
kamion	lorry	0.127	64,591
tahač	tractor unit	0.115	11,959

Table: Thesaurus results for *bagr*

a_modifier			is_subj_of			coord			is_obj4_of		
	2,910	0.20		1,867	0.13		1,098	0.08		1,229	0.09
sací +	360	9.17	zakousnout	35	6.19	buldozer +	133	9.43	zakousnout	5	3.45
sací bagr			zakously bagry			bagry a buldozery			vjet	9	3.26
kráčivý	42	8.71	bagrovat	5	5.88	rypadlo	29	7.51	přijet	98	2.96
korečkový	36	8.35	zakusovat	8	5.81	sbíječka	12	7.25	přijel bagr		
korečkový bagr			vyhloubit	10	5.40	nakladač	62	7.16	čumět	5	2.93
pásový +	150	7.98	hrábnout	7	4.94	bagrů a nakladačů			převážet	5	2.80
pásový bagr			hloubit	7	4.85	nákladák	78	6.78	najet	10	2.67
kráčející	42	7.86	vjet	25	4.71	bagry a nákladáky			povolat	7	2.47
kráčející bagry			vjedou bagry			jeřáb	79	6.24	ukrást	7	1.89
kolový	66	7.67	sesunout	5	4.58	bagry a jeřáby			kreslit	6	1.83
drapákový	19	7.66	zarýt	5	4.41	tatra	9	6.03	řádit	5	1.71
dvoucestný	19	7.02	najet	21	3.73	rýpadlo	5	5.99	nastartovat	6	1.70
dvoucestný bagr			najely bagry			tatrovka	7	5.97	pronajmout	6	1.67
demoliční	31	6.79	vyhrabat	7	3.51	míchačka	8	5.61	míjet	6	1.67
demoliční bagr			přetrhnout	5	3.51	krumpáč	8	5.52	nasadit	12	1.08
mohelnický	9	5.94	nakládat	17	3.36	traktor	79	5.37	pozvat	12	1.07

Figure: Word Sketches for the word *bagr*

Příbor / cutlery

1. souprava náčiní, kterým se jí (lžíce, vidlička, nůž) (utensils used for eating (spoon, fork, knife))
2. souprava jídelního nádobí (set of eating utensils)

Lemma	Translation	Score	Freq
nádobí	utensils, tableware	0.209	74,734
vidlička	fork	0.195	15,428
talířek	dessert plate	0.167	8,831
tácek	coaster	0.159	6,615
tác	tray	0.150	11,233
talíř	plate	0.148	73,763
hrníček	cup	0.147	13,490
hrnek	mug	0.143	31,018
hrneček	cup	0.143	14,385
lžička	teaspoon	0.138	49,997

Table: Thesaurus results for *příbor*

is_obj7_of		
	947	0.06
jezenit	5	6.67
cinkat	16	6.28
jíst +	559	4.44
jíst příborem		
najíst	22	4.03
se najíst příborem		
krájet	5	3.05
praštit	5	2.67
konzumovat	6	1.77
nabírat	5	0.96

prec_včetně		
	28	0.00
nádobí	14	2.58
nádobí včetně příborů		

Figure: Word Sketches for the word *příbor*

Trpět / suffer

1. prožívat, snášet bolest, trýzeň, nepříjemnost (experience, bear pain, suffering, inconvenience
2. být nemocen n. jinak strádat (be ill or suffer)
3. (trpně) snášet (to bear patiently)
4. hovor. mít v oblibě, potrpět si (to like sth)

has_obj7			has_subj			coord		
	110,541	0.36		58,706	0.19		8,920	0.03
deprese +	3,448	9.20	pacient +	1,105	6.13	strádat +	135	7.67
nedostatek +	7,540	8.75	pacient trpí			umírat +	526	6.98
trpí nedostatkem			Kristus +	282	5.74	opominout	46	6.71
porucha +	3,943	8.69	Kristus trpěl			něco konal , opominul nebo trpěl , bude		
nadváha +	1,635	8.57	tiš +	171	5.65	hladovět	40	6.02
trpí nadváhou			tiše trpí .			hladoví a trpí		
bolest +	5,455	8.40	chudák +	153	5.64	sténat	20	5.42
hlad +	1,788	8.33	zvíře +	882	5.60	trpí a sténá		
trpí hladem			dcera +	545	5.60	úpět	16	5.36
nespavost +	1,133	8.25	dcera trpí			krvácet	32	5.12
trpí nespavostí			akné +	111	5.51	zvracet	36	5.11
alergie +	1,705	8.23	trpím akné			mlčet	72	4.93
syndrom +	1,625	8.17	syn +	720	5.49	odpouštět	29	4.91
choroba +	2,588	8.09	syn trpí					

Figure: Word Sketches for the word *trpět*

Lemma	Translation	Score	Freq
projevovat	to show	0.255	221,222
umírat	to be dying	0.254	111,868
zemřít	to die	0.240	397,387
onemocnět	to fell ill	0.238	47,770
trápit	to afflict	0.225	237,852
cítit	to feel	0.219	970,162
žít	to live	0.214	1,333,268
umřít	to die	0.213	115,718
projevit	to show	0.211	326,269
způsobit	to cause	0.206	405,320

Table: Thesaurus results for *trpět*

Poněvadž a nebo / because & or

sp. podř. příčin. (důvod.), protože (subordinating conjunction expressing a cause)

1. vyj. vztah neslučitelnosti, anebo (expression of contradictoriness)
2. vyj. vztah mezi dvěma i více možnostmi, i časovými (relation between two and more options)

coord		
	102	0.00
protože	70	11.29
, poněvadž a protože		
jelikož	17	9.82
jelikož a poněvadž		
páč	3	9.53
nicméně	2	9.19
ježto	1	8.27
přestože	1	7.92
neboť	1	7.90
totiž	1	7.73
tudíž	1	7.67
když	2	4.55
že	1	3.17
pokud	1	2.11
proto	1	1.22

post_inf		
	343,361	0.03
vyvrátit +	1,400	6.90
potvrdit nebo vyvrátit		
počkat +	1,740	6.70
, nebo počkat		
použít +	4,416	6.44
, nebo použít		
brečet +	1,066	6.44
smát nebo brečet .		
využít +	3,856	6.38
, nebo využít		
zkusit +	1,720	6.33
, nebo zkusit		
zrušit +	1,557	6.30
nebo zrušit		
vyměnit +	1,417	6.29
nebo vyměnit		

coord		
	1,579	0.00
buď +	988	13.61
buď a nebo .		
přesto +	148	9.82
. Přesto a nebo právě proto		
i	27	7.73
IS a nebo		
či	8	7.32
a +	103	7.18
a a nebo		
že	15	6.58
nebo nebo že		
když	10	6.26
jestli	12	5.90
nebo nebo jestli		
proto	8	4.08
co	8	3.38

Figure: Word Sketches for the words *poněvadž* and *nebo*

Lemma	Translation	Score	Freq
anžto	because	0.588	2,565
páč	because	0.373	77,458
jelikož	because	0.346	539,891
jenomže	only	0.291	85,636
přestože	although	0.288	391,756

Table: Thesaurus results for *poněvadž*

Lemma	Translation	Score	Freq
či	or	0.908	3,395,720
,		0.884	303,403,489
a	and	0.879	137,863,596
i	also	0.869	25,577,373
-		0.850	18,915,328

Table: Thesaurus results for *nebo*

Conclusions

- Nouns: *hypernym-from-thesaurus*, which *verb-from-word-sketches* (at least sometimes)
- Verbs: using objects
- Adjectives: synonyms, antonyms
- Conjugations: usage patterns?

- Need for different approach. But it could work!