Options for Automatic Creation of Dictionary Definitions from Corpora

Marie Stará, Vojtěch Kovář

Faculty of Arts
and
Natural Language Processing Centre, Faculty of Informatics

Masaryk University
Brno, Czech Republic
413827@mail.muni.cz, xkovar3@fi.muni.cz

Abstract. This paper maps the possibilities of using existing corpus tools to acquire definitions for Czech in an automatic way. It compares definitions from Dictionary of contemporary Czech (Slovník současné češtiny pro školu a veřejnost) and data acquired using Thesaurus and Word sketch in corpus czTenTen12.

Key words: dictionary definition, corpora, word sketch

1 Introduction

Creation of definitions is one of the key steps in compilation of a monolingual learner's dictionary.

There is a long tradition in creating learner's dictionaries with heavy support of corpora and related tools (corpora were first used in lexicography in Collins COBUILD English Language Dictionary published in 1987 [1]). However, so far there is no good automatic method of supporting lexicographers in creating definitions.

There have been attempts at automatic extraction of definitions from corpora [2,3,4,5,6], but it seems that there is simply not enough definitions in general language corpora. However, there may be enough information to *create* a definition, i.e. to aggregate the available information and build a new definition.

The purpose of this paper is to make a survey over the corpora data for Czech to find out, to what extent it contains information suitable for such automatic definition building.

2 Method

To create a definition we first need to know what a definition should contain. According to Manuál lexikografie (Manual of Lexicography) [7] the basic types of definitions are:

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 111–124, 2016.

Table 1: Sketch Engine thesaurus for *příbor*

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

Fig. 1: Word sketch for *příbor*

- 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)

Another possibility is to use a synonym or antonym.

Next, we need to find out if there is a way how to get such definitions (or something close to them) using current corpus tools. In the next sections, we show definitions from "Slovník spisovné češtiny pro školu a veřejnost" (Dictionary of contemporary Czech, further referred to as SSČ) [8], the latest Czech monolingual learner's dictionary, and the data acquired from the 4-billion Czech web corpus `czTenTen12`, for a set of 10 words: 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), as an example of synsemantics.

This comparison forms the core of this paper and should show us if there is a potential for automatic creation of definitions from corpora.

3 Nouns

3.1 Příbor (cutlery)

According to SSČ, *příbor* has two meanings:

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

These two senses are practically identical and any automatic disambiguation would probably not be able to distinguish between them. However, according to Kilgarriff [9],

> if the instance exemplifies a pattern of use which is sufficiently frequent, and is insufficiently predictable from other meanings or uses of word, then the pattern qualifies for treatment as a dictionary sense.

SSČ does not abide this rule and lists results predictable from other meanings; so it is rather a problem in SSČ.

Table 1 shows that the word *nádobí* (utensils) is most similar to *příbor*. Word sketch shows that *příbor* is used together with *jíst* (to eat). Also from category *prec_včetně* (including), it is apparent that the word *nádobí* is a hypernym of *příbor*.

3.2 Pes (dog)

Pes has three meanings defined in SSČ:

1. šelma ochočená k hlídání, lovu ap. (domesticated carnivore for guarding, hunting etc.)
2. samec psovité šelmy (male canine)
3. expr. bezohledný, krutý člověk (expressively cruel person)

Table 2 shows that the word *zvíře* (animal) is most similar to *pes*. It is also a hypernym, however, this information is not present in the corpus results. Word sketch shows strong collocation between noun *pes* and verb *štěkat* (bark) – see Figure 2.

3.3 Bagr (excavator)

There are two meanings for the word *bagr* in SSČ:

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

Table 3 shows that the word *bagr* is most similar to *rypadlo*. This is not a hypernym as in previous cases, but a synonym. The importance of this interrelation is described bellow. Word sketches provide little relevant information. There is *rypadlo* in category *coord* but it has low frequency and also a relatively low score. The other categories shown in Figure 3 are also not very useful. The verb with highest score is *zakousnout* (have a bite) which is far from perfect for dictionary definition (but still, it is relevant, as the other verbs in the lists).

Table 2: Thesaurus results for *pes*

Lemma	Translation	Score	Freq
zvíře	animal	0.468	564,849
kočka	cat	0.447	294,032
dítě	child	0.428	4,455,634
pejsek	doggy	0.423	174,852
kůň	horse	0.402	485,617
muž	man	0.380	1,616,460
člověk	human	0.378	8,036,909
žena	woman	0.370	1,899,472
kluk	boy	0.348	671,754
rodič	parent	0.324	949,735

is_subj_of			post_verb			post_inf			prec_verb		
	126,125	0.12		90,507	0.09		24,820	0.02		48,529	0.05
štěkat +	2,377	9.04	štěkat +	1,229	8.49	bít +	1,357	8.89	štěknout +	604	8.48
štěknout +	2,105	9.01	pes štěká			chce psa bít , hůl si			neštěkne ani pes .		
ani pes neštěkne			štěknout +	918	8.27	venčit +	135	6.81	venčit +	608	8.30
chcípnout +	962	7.84	ani pes neštěkne .			psa venčit			venčí psa		
chcípl pes			žrát +	443	6.50	štěkat +	143	6.66	štěkat +	470	7.78
pokousat +	840	7.69	pes žere			psa štěkat			štěká pes		
pokousal pes			kousat +	236	6.00	odnaučit	91	6.49	srát +	483	7.67
pobíhat +	731	7.18	pes kouše			psa odnaučit			sere pes .		
srát +	554	6.87	vrčet +	182	5.87	vycvičit	90	6.37	pobíhají psi		
sere pes			pes vrčí			psa vycvičit			pobíhají psi		
skákat +	764	6.84	výt +	162	5.74	vykastrovat	66	6.15	vyvenčit +	160	6.52
Skákal pes			pes vyje			Než necháte svého psa vykastrovat ... "			vyvenčím psa		
žrát +	659	6.78	krmit +	262	5.60	vyvenčit	70	6.11	pořizovat +	333	6.41
pes žere			venčit +	156	5.60	psa vyvenčit ,			pořizuje psa		
běhat +	907	6.73	běhat +	352	5.56	vykoupat	80	5.79	cizit +	378	6.29
sežrat +	546	6.69	pes běhá			psa vykoupat			vzteklit +	107	6.16

Fig. 2: Word sketches for the word *pes*

3.4 Nouns: Summary

Some nouns can be defined as "*hypernym-from-thesaurus* that *verb-from-word-sketches*":

- příbor: nádobí, kterým se jí (cutlery: utensils that are used for eating)
- pes: zvíře, které štěká (dog: animal that barks)

However, this definition includes only the primary meaning. And it is not applicable for every noun, for example *bagr* is not really *rypadlo, které zakusuje* (*excavator* is not really *a digger that bites*). Also, it might be useful to distinguish when such hypernym is subject and when it is object (here distinguished by using passive voice) – current word sketch relations for Czech are not really good in this aspect.

Table 3: Thesaurus results for *bagr*

Lemma	Translation	Score	Freq
rypadlo	digger	0.244	3,452
buldozer	bulldozer	0.225	6,743
náklaďák	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

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	náklaďák	78	6.78	najet	10	2.67
kráčející	42	7.86	vjet	25	4.71	bagry a náklaďáky			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	mít	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

Fig. 3: Word sketches for the word *bagr*

4 Verbs

4.1 Trpět (to suffer)

Trpět has four meanings defined in SSČ:

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)

Table 4 shows that the word *trpět* is most similar to *projevovat* (to show) and *umírat* (to die). However, these are not hypernyms nor synonyms of *trpět*, although they are somehow semantically similar. Apparently we cannot define verbs in the same way as we outlined for nouns. Word sketch category *coord*

Table 4: Thesaurus results for *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

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					

Fig. 4: Word sketches for the word *trpět*

shown in Figure 4 yields *strádat* (suffer), a synonym used in meaning 2 in SSČ. Categories *has_subj* and *has_obj7* show very relevant pattern of usage, e.g. *pacient trpí depresemi* (patient suffers by depression).

4.2 Chytit (to catch)

Chytit has, according to SSČ, seven meanings:

1. rukou n. rukama uchopit a podržet, prudce vzít (to grab sth by hand)
2. zmocnit se lovem ap. (to hunt down)
3. rychlým pohybem dostihnout (to catch)
4. hovor. dostat, získat (to gain)
5. zachytit se, přilnout (to hold on sth)
6. hovor. i chytit se, zachvátit, zmocnit se (to capture)
7. začít hořet, vzplanout (to catch fire)

Table 5 shows that the word *chytit* is most similar to *chytnout* and *chytat*. These verbs are not suitable for definition: *chytit* and *chytnout* are semantically identical and only their written form differs, *chytit* and *chytat* differ only in verbal aspect.

Table 5: Thesaurus results for *chytit*

Lemma	Translation	Score	Freq
chytnout	to catch	0.517	126,197
chytat	to catch	0.389	111,670
vytáhnout	to pull up	0.289	243,591
popadnout	to grab	0.278	43,277
pustit	to drop, to let go	0.276	490,512
vzít	to take	0.265	1,263,663
držet	to hold on	0.262	959,864
uchopit	to catch	0.251	44,547
zabít	to kill	0.246	339,422
uvidět	to see	0.235	707,341

has_obj4			coord			post_na		
	29,937	0.17		11,054	0.06		1,815	0.01
zloděj +	1,923	8.95	pustit +	1,022	6.39	udička	53	9.29
chytte zloděje			chyť a pusť			vějička	24	7.85
penalta +	361	7.49	uvěznit	76	6.34	udice	42	7.70
chytil penaltu			chycen a uvězněn			chytit na udici		
dech +	806	7.31	usvědčit	44	6.12	lep	21	7.20
chytil druhý dech			chytit a usvědčit			chytit na lep		
kapr +	342	7.06	přiskočit	32	6.03			
míza +	144	7.04	přiskočil a chytil					
chytil druhou mízu			odvléct	30	5.80	špek	21	5.83
zlatonka +	112	6.89	chytili a odvlekli			boilies	10	5.78
chytil zlatonku			popravit	45	5.69	třpytka	7	5.54
slina +	178	6.80	chycen a popraven			flop	13	5.48
chytil slinu			dohonit	24	5.31	chytil na flopu		
rybka +	200	6.72	dohonit a chytit			boilie	8	5.40
chytí zlatou rybku			okroužkovat	14	5.31	háček	53	5.36
štika +	115	6.51	chytit a okroužkovat			švestka	21	5.18
chytil štiku			odtáhnout	38	5.21	chytit na švestkách		

Fig. 5: Word sketches for the word *chytit*

As for word sketches (Figure 5), the *coord* category has only one potentially applicable item with high score, antonym *pustit* (to drop, to let go). Other items (thief, breath, etc.) can be interesting but it is not clear how to use them directly.

4.3 Verbs: Summary

It is obvious that verbs require a different approach than nouns. Current corpus tools do not in fact offer any efficient way how to create definitions similar to those used in SSČ; however the word sketch results show objects that could help to describe meaning.

5 Adjectives

5.1 Zádumčivý (broody)

According to SSČ, *zádumčivý* has two meanings:

Table 6: Thesaurus results for *zádumčivý*

Lemma	Translation	Score	Freq
zadumaný	pensive	0.369	1,924
zasmušilý	melancholic	0.257	1,982
tklivý	touching	0.235	2,441
snivý	dreamful	0.233	1,432
zamyšlený	wistful	0.227	5,183
posmutnělý	unhappy	0.227	2,504
zasněný	wistful	0.215	5,095
teskný	sorrowful	0.205	2,074
introvertní	introvert	0.189	3,388
zamlklý	taciturn	0.181	5,388

Fig. 6: Word sketches for the word *zádumčivý*

1. zasmušilý (melancholic)
2. působcí takovým dojmem, smutný (looking broody; sad)

Table 6 shows that it is most similar to words *zadumaný* (pensive) and *zasmušilý* (melancholic). Word sketch relation *coord* (see Figure 6) displays quite similar results. Basically, all the results present there are synonyms.

5.2 Starý (old)

SSČ provides eleven meanings of *starý*:

1. jsoucí v závěrečném období života, vysokého věku (nearing the end of life, of advanced age)
2. (o člověku) jsoucí urč. věku, vytvořený před urč. dobou (being of certain age, created certain time ago)
3. v stáří obvyklý, stáří vlastní (typical to old age)
4. vytvořený před delší dobou, dlouhým užíváním opotřebovaný, bezcenný (created long time ago, timeworn, obsolete)
5. jsoucí dávného původu (being old, ancient)

Table 7: Thesaurus results for *starý*

Lemma	Translation	Score	Freq
nový	new	0.494	6,374,792
známý	known	0.433	1,412,791
původní	original	0.415	839,282
samotný	alone	0.407	932,916
velký	big	0.406	7,849,239
mladý	young	0.404	1,632,302
vlastní	one's own	0.399	1,877,789
krásný	beautiful	0.398	1,301,992
jediný	only	0.394	1,678,394
kvalitní	superior	0.391	941,081

modifies			coord		
	1,721,643	0.76		62,729	0.03
žák +	20,333	8.19	mocný +	2,997	8.41
starších žáků			Nový +	1,824	8.41
generace +	14,961	7.82	Starého a Nového zákona		
starší generace			zkušený +	2,713	8.05
verze +	19,935	7.81	starší a zkušenější		
bratr +	13,528	7.78	mladý +	7,872	7.42
starší bratr			mohoucí +	342	7.37
člověk +	50,410	7.76	staré a nemohoucí		
syn +	14,009	7.71	nemocný +	582	7.33
město +	32,788	7.70	staré a nemocné		
Starém Městě			pokročilý +	754	7.08
muž +	17,986	7.59	pro starší a pokročilé		
pán +	12,801	7.55	moudrý +	476	6.88
starý pán			starší a moudřejší		
léto +	16,972	7.26	krajový +	246	6.88
let staré			starých a krajových odrůd		

Fig. 7: Word sketches for the word *starý*

6. zastaralý, nemoderní (archaic, outdated)
7. dávno známý, často opakovaný (known for a long time, often repeated)
8. předešlý, bývalý (former)
9. dávný (ancient)
10. stejný jako dříve, původní (same as before, original)
11. delší dobu něco konající, osvědčený, zkušený (doing sth for a long time, time-proven, experienced)

Thesaurus places the word *nový* in the first place (Table 7); it is an antonym (as well as *mladý*). The only synonymic meaning in the first ten results is *původní* (original) which does not describe the primary meaning of *starý*. The word sketch data are not of much use either, as shown in Figure 7; it shows who and what can be old (in column *modifies*) which is almost anything. Antonyms *nový* and *mladý* can be useful which are also in thesaurus.

Table 8: Thesaurus results for *opilý*

Lemma	Translation	Score	Freq
ožralý	drunk	0.404	6,877
místný	local	0.321	87,162
sympatický	likable	0.314	107,989
podnapilý	tipsy	0.304	10,123
naštvaný	angry	0.303	52,656
sedící	sitting	0.302	34,073
vyděšený	scared	0.298	24,121
letý	of age	0.297	184,909
unavený	tired	0.289	101,496
nebohý	pathetic	0.282	17,633

modifies			adv_modifier			coord		
	23,620	0.51		8,641	0.19		1,492	0.03
řidič +	3,260	7.91	namol +	342	10.16	zfetovaný +	185	10.46
bezdomovec +	356	7.50	namol opilý			zdrogovaný +	100	10.19
řidička +	180	7.18	věčně +	423	8.65	zhulený	24	8.19
. Opilá řidička			věčně opilý			zkouřený	16	8.08
koráb +	144	7.12	notně +	155	8.04	pomočený	14	7.70
Opilý koráb			notně opilý			nadrogovaný	9	7.50
námořník +	178	7.00	domů +	244	7.40	sfetovaný	7	7.11
opilý námořník			domů opilý			intoxikovaný	6	6.60
mladík +	656	6.91	totálně +	248	7.37	sjetý	12	6.33
výtržník	81	6.49	totálně opilý			omámený	9	6.14
šofér	84	6.42	silně +	591	6.93	podchlazený	7	5.95
cyklista +	236	5.86	silně opilý			podnapilý	18	5.59
Opilý cyklista			zjevně +	116	6.46	zapáchající	10	5.50
muž +	2,008	5.59	zjevně opilý			dezorientovaný	7	5.41
opilý muž			značně +	287	5.97	střízlivý	19	5.03

Fig. 8: Word sketches for the word *opilý*

5.3 Opilý (drunk)

Opilý has three meanings defined in SSČ:

1. opojený nadměrným požitím alkoholického nápoje (intoxicated by alcohol)
2. svědčící o tom (indicating sb is drunk)
3. expr. mocně zaujatý, opojený, omámený (really preoccupied, intoxicated)

Table 8 shows that the most similar thesaurus result is the expressive synonym *ožralý*. Word sketch does not show any applicable results, maybe except for *namol (opilý)* (blind drunk) from the *adv_modifier* category, and semantically near words in the *coord* category.

5.4 Adjectives: Summary

While it is possible to define some adjectives using existing corpus tools, in other cases it seems to be more complicated. The difference might be in how many meanings a word can have and how narrow these meanings are.

6 Synsemantics

6.1 Poněvadž (because)

Poněvadž is defined as

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

Thesaurus as well as word sketch offers numerous synonyms (see Table 9 and Figure 9).

6.2 Nebo (or)

Nebo is defined as

> 1. vyj. vztah neslučitelnosti, anebo (expression of contradictoriness)

or

> 2. vyj. vztah mezi dvěma i více možnostmi, i časovými (relation between two and more options)

Here the data are not so clear, thesaurus suggests synonym *či* (Table 10), however, as shown in Figure 10, word sketch only provides few results. Relation *post_inf* may be interesting to look at: if we had also relation *prec_inf*, we may be able to extract useful usage patterns, such as *potvrdit nebo vyvrátit* (to confirm or disprove).

6.3 Synsemantics: Summary

As long as conjunctions are defined only by synonyms, it is possible to obtain some useful results; however, this would inevitably lead to a circular definition. The question is how the definitions of synsemantics should look like. The SSČ-like definitions would be hard to create, but they do not seem to be extremely useful anyway. Usage patterns may be a better way of "defining" these words, and it may be easier to extract them from corpora.

Table 9: Thesaurus results for *poněvadž*

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
jestliže	if	0.250	363,826
bo	because	0.225	48,891
byť	although	0.218	247,035
neboť	because	0.156	767,296
nýbrž	but	0.148	206,053

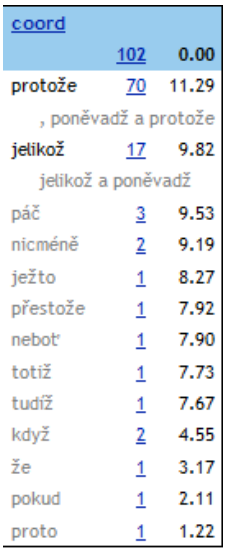

Fig. 9: Word sketches for the word *poněvadž*

7 Conclusions

As shown above, the existing corpus tools are able to find fragments that can be used in definitions. Different parts of speech will require slightly different approaches.

It is impossible to estimate whether automatic definition of synsemantics will prove doable, because their meaning is often too specific to be defined easily (e. g. conjunctions (and, or) and prepositions (in, on)).

Automatic creation of definitions, at least to some extent, should be possible for nouns, verbs, adjectives and adverbs. A special sketch grammar aimed at needs of such definitions may help.

Existing dictionaries (like SSČ) list many meanings which are quite similar and listing all those meanings is redundant. Differentiation between meanings that should be distinguished is another task.

Table 10: Thesaurus results for *nebo*

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
)		0.840	28,670,317
on	he	0.834	32,844,994
:		0.817	27,939,280
ale	but	0.816	25,330,812
ani	neither	0.803	5,674,148

post_inf			coord		
	343,361	0.03		1,579	0.00
vyvrátit +	1,400	6.90	buď +	988	13.61
potvrdit nebo vyvrátit			buď a nebo .		
počkat +	1,740	6.70	přesto +	148	9.82
, nebo počkat			. Přesto a nebo právě proto		
použít +	4,416	6.44	i	27	7.73
, nebo použít			IS a nebo		
brečet +	1,066	6.44	či	8	7.32
smát nebo brečet .			a +	103	7.18
využít +	3,856	6.38	a a nebo		
, nebo využít			že	15	6.58
zkusit +	1,720	6.33	nebo nebo že		
, nebo zkusit			když	10	6.26
zrušit +	1,557	6.30	jestli	12	5.90
nebo zrušit			nebo nebo jestli		
vyměnit +	1,417	6.29	proto	8	4.08
nebo vyměnit			co	8	3.38

Fig. 10: Word sketches for the word *nebo*

Acknowledgments. This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2016* (MUNI/A/0863/2015) and by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. HarperCollins Publishers: The history of COBUILD. [online] ((accesed November 9, 2016))
2. Kovář, V., Močiariková, M., Rychlý, P.: Finding definitions in large corpora with Sketch Engine. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
3. Przepiórkowski, A., Degórski, Ł., Wójtowicz, B., Spousta, M., Kuboň, V., Simov, K., Osenova, P., Lemnitzer, L.: Towards the automatic extraction of definitions in Slavic. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, Association for Computational Linguistics (2007) 43–50

4. Navigli, R., Velardi, P., Ruiz-Martínez, J.M.: An annotated dataset for extracting definitions and hypernyms from the web. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
5. Jin, Y., Kan, M.Y., Ng, J.P., He, X.: Mining scientific terms and their definitions: A study of the ACL anthology. EMNLP-2013 (2013)
6. Klavans, J.L., Muresan, S.: Evaluation of the DEFINDER system for fully automatic glossary construction. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2001) 324
7. Čermák, F., Blatná, R.: Manuál lexikografie. 1st edn. H & H, Jinočany (1995)
8. Filipec, J.: Slovník spisovné češtiny pro školu a veřejnost. 4th edn. Academia, Praha (2005)
9. Kilgarriff, A.: "I don't believe in word senses". Computers and the Humanities **31**(2) (1997) 91–113