Are Dictionary Definitions of Verbs in Corpora?

Discovering Dead Ends in Generating Explanations of Verbs

Marie Stará

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
413827@mail.muni.cz

Abstract. Compared to nouns, there are no clear guidelines on what a definition or explanation of a verb should contain. We search corpora for the information present in Czech and English dictionaries to find out if the data is present and, thus, if it can be used as a guideline for assessing the quality of generated definitions. We show that even though a notable portion of the dictionary data is present in corpora, it does not seem frequent or specific enough to be findable with the methods we previously used.

Keywords: verb, meaning, explanation, corpora

1 Introduction

Defining a word or explaining its meaning is a task relatively easy if the word in question is a noun. Literature[1] covers the topic quite well, and gathering data to create explanations automatically is arguably a doable task (cf. [5,6]). For verbs, however, the situation is different: There is less agreement on what its definition/explanation should consist of, and thus it is more complicated to create a meaningful explanation of a verb.

We do not try to suggest the best approach to explain the meaning of a verb, merely to find out whether corpus data contain the information one can find in dictionary definitions of verbs. Subsequently, one can ask if the generated definitions can (or should) approximate the human-made definitions.

In Section 2, we list the resources we used and briefly describe the structure of our queries and the difference between Czech and English data. In Section 3, we list and discuss our results. In Section 4, we conclude that using existing dictionaries as a standard to assess the quality of generated data is not necessarily an optimal solution.

2 Method

We chose a small set of 19 verbs in Czech and English and gathered their definitions from two dictionaries, Slovník spisovné češtiny [1] and Macmillan

[1] For a summary (in Czech), see [6].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 65–69, 2022. © Tribun EU 2022

Table 1: The number of queries created, found, and found with low frequency for chosen verbs in Czech and English.

	queries	found	low freq.		queries	found	low freq.
bát (se)	8	4	3	fear	5	3	2
dát	19	16	1	give	7	7	3
dostat	6	4	1	get	5	5	1
existovat	3	3	0	exist	7	5	2
fungovat	5	5	5	function	7	5	1
jednat	11	8	1	act	6	5	3
ležet	4	3	1	lie	4	1	1
namlouvat	6	1	1	insinuate	2	0	-
pojistit	6	1	0	insure	3	2	1
skákat	4	2	1	jump	9	3	2
ulovit	4	3	2	hunt	10	7	3
uvést	10	4	2	initiate	4	1	1
večeřet	2	0	-	dine	1	1	1
vřít	5	4	3	boil	4	2	1
zabít	7	5	3	kill	4	2	1
začít	3	3	1	begin	4	4	2
zapomenout	6	2	2	forget	3	3	2
znemožnit	2	1	1	discredit	2	1	1
žít	4	3	3	live	4	3	0

dictionary [2], respectively. As verbs usually have a great number of senses, for verbs with multiple definitions, we used only the first three.

From these definitions, we extracted the words and phrases we considered relevant and searched for them in the czTenTen17 [3] and enTenTen13 [4] corpora[2]. To follow the approach used in [6], the word or phrase in question had to be present in the same sentence as the given headword.

We did not try to find the exact wording of the definitions; we searched for a part of the definition at a time, often using wildcards in place of (possible) modifiers and determiners.

We used CQL queries with the following structure:

```
[lemma = "headword"] []{0,7} [lemma = "other_verb"] within < s/> |
[lemma = "other_verb"] []{0,7} [lemma = "headword"] within < s/>,
```

where the `[lemma = "other_verb"]` consisted of a single verb, a verb and its object, or a more complicated phrase.

For example, (a part of) the definition for *hunt* is *to kill animals for food*. The `[lemma = "other_verb"]` | is, in this case, replaced by: `[lemma = "kill"] [] {0,2} [lemma = "animal"]? [lemma = "for"] [lemma = "food"]`.

Although the method we chose was the same for both languages, the approaches to explaining the meaning of verbs differ in the dictionaries: In Czech, the verb is explained mostly by other (usually multiple) verbs with

[2] The corpora were chosen purely practically, to be big but not too much, for time is a scarce resource.

Table 2: Czech verbs, sorted by percentage of found queries and percentage of low-frequency results.

	% found	% low freq.		% found	% low freq.
existovat	100	0	existovat	100	0
fungovat	100	100	fungovat	16,7	0
začít	100	33,3	začít	84,2	6,3
dát	84,2	6,25	dát	72,7	12,5
vřít	80	75	vřít	66,7	25
ležet	75	33,3	ležet	75	33,3
ulovit	75	66,7	ulovit	100	33,3
žít	75	100	žít	50	50
jednat	72,7	12,5	jednat	40	50
zabít	71,4	60	zabít	71,4	60
dostat	66,7	25	dostat	75	66,7
bát (se)	50	75	bát (se)	50	75
skákat	50	50	skákat	80	75
znemožnit	50	100	znemožnit	100	100
uvést	40	50	uvést	16,7	100
zapomenout	33,3	100	zapomenout	33,3	100
namlouvat	16,7	100	namlouvat	50	100
pojistit	16,7	0	pojistit	75	100
večeřet	0	-	večeřet	0	-

similar meaning and, occasionally, a relevant object (noun or prepositional phrase, usually). Some of the definitions contain examples leading to listing other meanings.

In English, the definitions are more descriptive and approximate actually spoken language. These definitions usually contain an explanatory verb and its object(s) and adverbials. Some are structured as a sentence.

The different approach can be demonstrated on definitions for *hunt*. In English, the first definition is *to kill animals for food or for their skin or other parts, or for sport*; in Czech, it is *získat lovem* (to obtain by hunting).

3 Results

We got at least some results for all the verbs except two (insinuate, večeřet). Generally speaking, the result for English are slightly better, meaning the percentage of queries found is higher, and the percentage of results with low frequency is lower than for Czech. nevertheless, the overall results are similar.

An overview of the results is presented in Table 1 showing the absolute number of queries created and found, together with the number of results with low frequency. As we made no thorough attempt to clear the resulting data, some of our found queries include modal or aspectual verbs, which in some cases modify other verbs not included in the definitions.

Table 3: English verbs, sorted by percentage of found queries and percentage of low-frequency results.

	% found	% low freq.		% found	% low freq.
begin	100	50	live	75	0
dine	100	100	function	71,4	20
forget	100	66,7	get	100	20
get	100	20	exist	71,4	40
give	100	42,9	give	100	42,9
act	83,3	60	hunt	70	42,9
live	75	0	begin	100	50
exist	71,4	40	boil	50	50
function	71,4	20	insure	66,7	50
hunt	70	42,9	kill	50	50
insure	66,7	50	act	83,3	60
fear	60	66,7	fear	60	66,7
boil	50	50	forget	100	66,7
discredit	50	100	jump	33,3	66,7
kill	50	50	dine	100	100
jump	33,3	66,7	discredit	50	100
initiate	25	100	initiate	25	100
lie	25	100	lie	25	100
insinuate	0	-	insinuate	0	-

Tables 2 and 3 show the results sorted according to the percentage of found queries and low-frequency results for Czech and English, respectively. We found no apparent correspondence between the number of absolute and low-frequency results.

4 Conclusion

The corpora do contain some of the dictionary definitions vocabularies. This data is, however, not frequent nor specific enough to be found by the method previously used.

When it comes to whether it makes sense to use the existing dictionary definitions as a benchmark or at least a reference point, the answer is, it depends. For Czech, we lean towards no, as the definitions are mostly lists of synonyms or synonyms with an object. (Unless, of course, we use the dictionary to check the relevance of possible synonyms.) As for English, it is something to consider.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

References

1. Internetová jazyková příručka, https://prirucka.ujc.cas.cz/, [cit. 2022-10-30]

2. Macmillan dictionary, https://www.macmillandictionary.com/, [cit. 2022-10-30]
3. Lexical Computing: Czech web 2017 (cstenten17), https://www.sketchengine.eu/cstenten-czech-corpus/, [cit. 2022-11-04]
4. Lexical Computing: English web 2013 (ententen13), https://www.sketchengine.eu/ententen-english-corpus/, [cit. 2022-11-04]
5. Stará, M.: Automatically created noun explanations for english. In: Aleš Horák, Pavel Rychlý, A.R. (ed.) Proceedings of the Thirteenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2019. pp. 83–87. Tribun EU, Brno (2019)
6. Stará, M.: Automatická tvorba definic z korpusu. Diplomová práce (2019), https://is.muni.cz/th/i9wm5/