Automatically Created Noun Definitions for Czech

Marie Stará

Faculty of Arts, Masaryk University
Arna Nováka 1, 602 00 Brno, Czech Republic
413827@mail.muni.cz

Abstract. This paper comments on the automatically created noun definitions. The definitions were created using the results of the Sketch Grammar developed with the aim to gather data for them. These data (Word Sketches) are combined using Python script to form the definition.

Keywords: dictionary definition; corpora; word sketch

1 Introduction

When a monolingual dictionary is created, one of the most complicated tasks to be dealt with is the creation of definitions. Though corpora are used in lexicography since the 1980s as a source of examples (and there were efforts to automatically find definitions), there has been (at least to my knowledge) no attempt to create a tool that would make it possible to *create* definitions automatically.

The purpose of this paper is to show automatically created definitions of Czech nouns and evaluate whether they can be used in a dictionary as they are, or if they can only serve as a basis for human-made definition.

2 Construction of definitions

As is stated below, it is not possible to create such definitions as those we can find in human-made dictionaries. I am using the word definition even though it is more of a set of hints for understanding a word. The definitions are created by composing Word Sketches of the given word together; I am using existing Word Sketches and adapting them to suit the needs of the definition creation, thus making my own Sketch Grammar. This Sketch Grammar was used with the 5-billion-token czTenTen12 corpus.

To construct a definition, I download Word Sketches of the given word in JSON format and use a script in Python to form the pieces of information together. The script takes first three words with highes score for each relation and merges them together in groupes described below.

Each definition consists of several parts, each of which is formed by one or more Word Sketch relation.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 63–68, 2018. © Tribun EU 2018

The most common definitions of nouns consist of genus proximum (hypernym of the given word) and differentia specifica (what distinguishes the word from its synonyms). [1,2] I am using the hypernym relation as well as the relation of synonymy; since it is not possible to reliably establish the hypernymy and/or synonymy for every word in the corpus, I am using the relation of very loose synonymy in my definitions. It is formed by combining results of Word Sketches *coord* (coordination; searching for words connected either by the conjunctions *a* (and) or *nebo* (or), or *ani–ani* (nor–nor) or *buď–nebo* (either–or), *a_jine* (and similar), *například* (for example), and *hypo_hypero*[1] (hyponymy—hypernymy). That means almost every noun is allocated with one or more words of similar meaning.

Below, the relevant part of definition of *pes* (dog) is showed.

> *podobný význam má zvíře, mazlíček, zvířectvo, dítě, člověk, plemeno, jezevčík*
> (similar meaning has animal, pet, fauna, child, human, breed, Dachshund)

Quite similar is the combination of *adj_modif* (adjective modifiers) of the given word and *slovo_je* (the word is).

> *pes může být hlídací, zakopaný, lovecký, pes, zvíře, přítel*
> (a dog can be a watchdog, burried, hunting, a dog, an animal, a friend)

Different but still quite frequent in definitions is the "part of" relation (partitive) [2]. For finding parts I'm using the *slovo_má* (a word has) and *skládá_se_z* (is consisting of) relations. Related is the *skládá_se_z_2* (is consisting of 2) relation which finds the words that consist of the given word; the skládá_se_z and skládá_se_z_2 relations are symmetrical. Similar to skládá_se_z_2 is the *kdo_co_má_slovo* (who/what has a word) relation; both relations should, optimally, find holonyms of the given word.

> *pes může mít pes, srst, vodítko*
> (the dog can have a dog, a hair, a string)
> *pes, majitel, soused může mít pes [psa]*
> (a dog, an owner, a neighbor can have a dog)

Another piece of the definition consists of verbs to which the defined word is either a subject(*je_podmět*) or an object in accusative (*je_předmět_4*) or instrumental (*je_předmět_7*). The valency is important for the definition, as it is a stable pattern of usage and therefore helps us understand the meaning of the unknown word. [4]

It could be argued that using only accusative and instrumental is not enough and that the genitive and dative forms should be used as well. There are two reasons for excluding them. Firstly, the genitive and dative objects have a lower frequency than the accusative and instrumental ones. Secondly, the cases are

[1] This relation was introduced by Baisa and Suchomel in [3].

often not appropriately recognised by the tagger, probably because of many cases of case syncretism in Czech.[2]

> *pes (se) může štěkat, štěknout, chcípnout; je možné jej/ji venčit, vyvenčit, pořídit a jím/jí vrtět, nakrýt, venčit*
> (dog can bark, make one bark, die; it can be being walked, walked, acquired; you can wag it, it can be used at stud (breeded))

There are two more relations I use in the definitions. These are *gen* (genitive following the given noun) and *instr* (instrumental following the given noun). It is useful only in some cases (mostly due to many wrong PoS tags), but I aim for good recall more than for good precision. (The reason is that I try to find definitions even for words with low frequency, which results in a lot of garbage data with the frequented ones.)

> *pes čeho: stář*[3]*, demokracie, plemeno*
> (dog of an age, a democracy, a breed)
> *pes (s) kým/čím: mikročip, povel, psovod*
> (dog with a microchip, a command, a dog handler)

3 Evaluation of definitions

I evaluated the definitions on a set of 78 nouns. The words were chosen based on various criteria. The most apparent criterion is frequency: words with both high and low frequency are included. There are words which seem to be easy to define (e.g. *pes – zvíře, které štěká*, dog – an animal which barks) and those which are harder to explain. The complexity of the explanation is connected to whether the word being defined is an abstractum or a concretum (abstract words being more complicated to define). In the set, there are words with one and more meanings. There are also synonyms included as well as words creating a scale – diminutives and augmentatives. Some words were picked ad hoc to ensure the test set is differentiated enough.

3.1 Examples

There are few words, for which the Word Sketches do not yield sufficient data. *Cestující* (someone who is travelling) is not recognised as a noun, but only as an adjective, therefore it does not contain data for the definition. *Barabizna* (augmentative expression for a house) is assigned only adjective modifiers and verbs to which it is either subject or object, due to its low frequency. Some other words with low frequency are not possible to define using my approach, for example, *barik* (oak [barrel for winemaking]) or *exposé* (an expose), the only word in the set for which I found only irrelevant data.

[2] *nominative–accusative*: inanimate masculine in both singular and plural; plural of feminine and neuter
genitive–accusative: singular of animate masculine
dative–locative: plural of masculine (both animate and inanimate), feminine, neuter

[3] this is an example of wrong lemma, it should be spelled as *stáří*

barik
podobný význam má sičák[4]
(similar meaning has a crook)
barik může být zatříděný, ovocitý, zapracovaný, víno
(oak can be classed, fruity, strong, wine)
barik (se) může ochutnat, potlačit, slušet; jím/jí překvapit, ovlivnit
(oak can be tasted, suppressed, matching; you can surprise with it, you can influence it)
barik čeho: aroma, vůně, chuť
(oak of aroma, smell, taste)
exposé
podobný význam má spacesa, spaces, dashboard, najetí, program
(similar meaning has spaces, dashboard, pointing (the cursor), programme)
exposé může být chvaličský, peterssonův, a-l, ročenka, kratochvíle, inovace
(expose can be subservient, peterssons, a-l, a yearbook, an amusement, an innovation)
exposé může mít svazek, roh, omezení
(exposé can have ligament, horn/corner, restriction)
stejskal, senzor, kláves[5] *může mít exposé*
(stejskal (a surname), a sensor, a key(board) can have an expose)
exposé (se) může namapovat, ozřejmit, sjednocovat; je možné jej/ji salariésit, namapovat, chromat a jím/jí napodobit, narušit, zobrazit
(expose can be mapped, explained, unified; it can be *salariesed*, mapped, *chromed*; you can imitate, disturb, show it)
exposé čeho: přescent[6]*, eleanora, zaorálek*
(expose of eleanor, zaorálek (a surname of a politician)
exposé (s) kým/čím: kinematografie, líčení
(expose with a cinematography, make-up)

On the other hand, the meaning of *trdliště* (fish breeding ground) can be deduced from the automatically created definition, even though it is a very uncommon expression.

trdliště
podobný význam má zimoviště, jikra, úkryt, chvojí
(similar meaning has winter quarters, fish egg, hiding, branches)
trdliště může být lipaní, vysbírán, lososí, pach, lov, samice
(it can be of graylings, picked up, of salmons, smell, hunt, female
trdliště může mít orlice, průměr
(it can have eagle, diameter)
makrela, ryba, populace může mít trdliště
(mackarel, fish, population can have fish breeding ground)

[4] misspelled in corpus, should be syčák
[5] misspelled, should be klávesa
[6] a word play

trdliště (se) může vlákat, spásat, hloubit; je možné jej/ji vybagrovat, devastovat, poničit
(it can allure sth, be grazed, be deepened; it can be excaveted, devastated, destroyed)
trdliště čeho: bistrino, lipan, losos
(fish breeding ground of bistrino (a name), a grayling, a salmon)
trdliště (s) kým/čím: sediment, peřej, většina
(fish breeding ground with a sediment, chute, majority)

Another problematic group is abstract expressions, e.g. *vypjatost* (the "being extreme" property), *dobro* (well-being, the good), *léto* (summer, year). On the other hand, other abstract expressions have more or less acceptible definition, for example *nenávist* (hatred) or *nicota* (nothingness).

3.2 Evaluation

The most reliable data result from valency followed by the loose synonymy and word-is relation. The least reliable are the instrumental and genitive of the given word with only 14 and 34 good results, respectively.

The partitive relations (holonymy and meronymy) [5] has comparatively better results for finding parts of the given word than for finding its holonyms. (This difference might be caused by the test set.)

One of the reasons the definitions are not good enough to be used without any editing is considerably high frequency of wrongly identified lemmata (typically recognizing adjectives as a 3rd person of a verb, e.g. *(mainská) mývalí kočka* (Maine Coon), where the adjective *mývalí* (racoon-like) is identified as a verb ([a cat is] racooning). Another reason is the above-mentioned wrong case identification.

It is worth noting that many of the ill-defined words are not included in the most up-to-date Czech monolingual dictionary [6]. Some of the definitions presented there are, moreover, hard to decipher even for native speakers.[7]

All in all, the definitions are not good enough to be presented in a dictionary without any editing. Nevertheless, they could be very well used as a basis for forming new user-friendly definitions.

4 Conclusion

With the corpus data containing mistakes in lemmata as well as tags, it is nearly impossible to automatically create definitions which would not need any editing. It is, however, possible to make a good basis for lexicographers to work on. This approach could be used in other languages significantly simplifying the process of dictionary.

[7] I asked a non-native speaker with C1 level Czech, and he could not understand about 20 % of the presented definitions. I would argue that monolingual dictionary that does not explain is not very good.

Table 1: Percentage of good/bad/no results for each group of relations

	good results (%)	bad results (%)	no result (%)
similar meaning has (the loose synonymy)	91.03	6.41	2.56
word can be	93.59	6.41	0.00
word can have (meronymy)	66.66	23.08	10.26
sth can have *word* (holonymy)	50.00	33.33	16.67
valency	92.31	5.13	2.56
genitive	43.59	50.00	6.41
instrumental	17.95	60.26	21.79

Acknowledgements. This work has been partly supported by the project of specific research Čeština v jednotě synchronie a diachronie (Czech language in unity of synchrony and diachrony), MUNI/A/0862/2017.

References

1. Atkins, S., Rundell, M. The Oxfor Guide to Practical Lexicography. Oxford University Press, New York (2008)
2. Landau, S.: Dictionaries: the art and craft of lexicography. Cambridge University Press, Cambridge (2001)
3. Baisa, V., Suchomel, B.: Corpus Based Extraction of Hypernyms. [online] (accessed August 29, 2018)
4. Hanks, P.: How people use words to make meanings: Semantic types meet valencies. [online] (accessed October 14, 2018)
5. Hladká, Z., Dočekal, M.: MERONYMNĚ-HOLONYMNÍ VZTAH. [online] (accessed November 9, 2018)
6. Filipec, J.: Slovník spisovné češtiny pro školu a veřejnost. 4th edn. Academia, Praha (2005)