Improving Compound Adverb Tagging

Hana Žižková

Masaryk University, Faculty of Arts,
Arna Nováka 1, 60200 Brno, Czech Republic
zizkova@phil.muni.cz
http://www.phil.muni.cz

Abstract. This paper describes the corpus probe we made to obtain and analyze data with a focus on improving compound adverb tagging. Thanks to our research we gain large amounts of unrecognized units that resemble to compound adverbs. We manually selected 470 units and we examined whether they are listed in existing Czech dictionaries and how they are tagged in corpus if we respread it into multiword expression. We found out that the compound adverb tagging in Czech National Corpus is inconsistent and unsatisfactory, so we proposed three solutions for improving compound adverb tagging.

Keywords: compound adverb; automatic morphological analysis; tagging

1 Introduction

Compound adverbs represent an interesting issue in terms of automatic morphological analysis (AMA). In Czech, the compound adverbs are always formed from a preposition and a noun or a preposition and an adjective or a preposition and a pronoun or a preposition and a numeral or a preposition and an adverb. Recognition of compound adverbs by AMA is difficult because, Czech compound adverbs are written mostly together as one word, but often there exists a multiword expression and their meaning is the same (*na příklad* – *například*) [1]. For instance Dokulil [2] states that: "compound adverbs are formed by compounding frequently occurring words in a sentence, without any change in their form. It is characteristic for them that you can always divide the compound adverb again." For the purposes of this paper it is essential that we write compound adverbs mostly together as one word, but often in parallel compound adverbs there exists a multiword expression. Additionally, a member of the multiword expression can function independently of this expression as a separate word [3]. Multiword expressions can be "defined as expressions which are made up of at least two words and which can be syntactically and/or semantically idiosyncratic in nature. Moreover, they act as a single unit at some level of linguistic analysis." [4]

There are contexts in which one may hesitate whether to use a one-word adverb or a multiword expression (*Obarvit načerno.* vs. *Obarvit na černo.*). Another

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 103–109, 2018.

important feature of the compound adverbs is that when written as two (or more) words, it is not possible to insert another expression between the two words that could develop the unit (*například – na příklad*, but not **na dobrý příklad*).

It is important for the compound adverb to be recognized by AMA in both cases (as a one-word and also as a multiword expression) regardless of whether the codification determines what is the correct spelling of the compound adverb. The automatic morphological analysis takes place in three steps: the first is a division of word forms (tokenization), the second is an assignment of one, but usually more interpretations from the morphological dictionary and the third step it is the disambiguation, which means assigning an interpretation [5].

The AMA recognizes and correctly identifies such compound adverbs which are written as one-word and are listed in the morphological dictionary.

There are many ways how to examine compound adverbs. We decided to make a corpus probe to identify compound adverbs tagged as an unrecognized part of speech in Czech National Corpus SYN v3 corpus[1] [6]. We have chosen unrecognized compound adverbs because they will likely have the same characteristics as those recognized, and we will thus have the data to add into the morphological dictionary. The obtained data were sorted out manually and grouped by their prefix: *do-*, *k-/ku-*, *mezi-*, *na-*, *nad-*, *o-*, *ob-*, *od-*, *po-*, *pro-*, *před-*, *při-*, *s-/sou-*, *u-*, *v-*, *z-*, *zpod-*, *za-* and, consequently, by their ending, because every prefix (previously preposition) can have more than one word ending (e.g. *poanglicku, pořadě, pokrk, pošesté, poprvní*). Afterwards, we were interested whether the AMA recognizes expressions that we have found as a one-word unit with the tag [tag="X.*"] if we respread them into multiword expressions. And if the AMA recognizes them, what tag will it assign them with. So we searched in the corpus gradually for multiword expressions of one-word compound adverbs that we found while processing the first step.

2 Finding

Thanks to the chosen CQL queries,[2] we have obtained a relatively large set[3] of one-word expressions that have the same initial and ending strings as a possible compound adverb. By hand selection, we have identified 470 units that we thought could be compound adverbs. They were not recognized by AMA because they were not listed in the morphological dictionary. Many of the one-word compound adverbs (e.g. *kpředu, odposledka, zmísta, zšeda, předloni,*

[1] At the time the biggest availible corpus in Czech National Corpus.

[2] [tag="X.*" & lemma="po.*"], [tag="X.*" & lemma="do.*"], [tag="X.*" & lemma="k.*"], [tag="X.*" & lemma="ob.*"], [tag="X.*" & lemma="od.*"], [tag="X.*" & lemma="o[^ db].*"], [tag="X.*" & lemma="mezi.*"], [tag="X.*" & lemma="na.*"], [tag="X.*" & lemma="pro.*"], [tag="X.*" & lemma="před.*"], [tag="X.*" & lemma="při.*"], [tag="X.*" & lemma="s.*"], [tag="X.*" & lemma="u.*"], [tag="X.*" & lemma="v.*"], [tag="X.*" & lemma="za.*"], [tag="X.*" & lemma="z[^ a].*"]

[3] More than 30.000 units.

naven, nablint, ...) are recorded in existing dictionaries, so they are not only occasionalism.

Somewhat more complicated situations have been encountered in the case of a compound adverbs in the form of a multiword expressions. We noticed that most of the compound adverbs are recognized by automatic morphological analysis and, from the point of view of word formation, the multiword expression is tagged as a preposition and part of speech from which the compound adverb is formed. Most often they are nouns (e.g. *na mokro, k dobru, ob den, ...*), but we have also noted adjectives (e.g. *na jisto, do pevna, ...*), adverbs (e.g. *na knap, na krátce, na tajno, k stáru, ...*) or numerals (e.g. *ob dva, na vícekrát, po mnohokrát, ...*), pronouns (e.g. *po svých, ...*) and prepositions (e.g. *na podél, na prostřed, ...*). In rare cases, we have registered the preposition and the verbs (e.g *do leskla, k předu, na zrz, z nenadála*).[4]

We found interesting that most of the obtained expressions were a compound of preposition and nouns (nouns, adjectives, pronouns, numerals) in the singular (e.g. *naskok, dočervena, nadálku, ...*), but we also noticed the compound of the preposition and the noun in plural (e.g. *nahony, sdíky, počertech, odvěků*).

We have found many multiword compound adverbs in Idiomatic and phrasal dictionary (DEBDict) [7,8] (e.g. *na světlo, na slovo, nablint, po krk, po čertech, ...*) and some of the analyzed data have shown strong collocations (e.g. *zbarvit do bíla / dobíla; zaostřovat do blízka / doblízka; holení na mokro / namokro; rozválet / vyválet / nakrájet na tenko / natenko; být natuty / na tuty; ...*).

Tagging of multiword compound adverbs as a preposition and seven different part of speech is inconsistent. Especially when comparing multiword expressions tagging such as *na tvrdo* (POS=R, POS=A), *na žluto* (POS=R, POS=N), *na tajno* (POS=R, POS=D). However, this is understandable with respect to the tagset currently used for the SYN corpora series in Czech National Corpus. The currently used tagset does not contain any tag for the compound adverb or its part. We think this is inappropriate.

By analyzing, we found that not all prepositions taken into account in queries form part of compound adverbs, to four (*u, mezi, o, při*) no expression was found according to established criteria.

3 Suggestions

By analyzing the corpus data, we came up with three proposals that could improve the automatic tagging of compound adverbs. The first proposal is the addition to the morphological dictionary, the second is the change of tag, and the third is the addition of strong collocations into the Multiword Expressions Lexical Database.

[4] In Czech it is not possible to follow the verb after the preposition. In the case of an expression *k předu*, this is an error in the disambiguation, since both the POS=D and POS=V interpretations are attributed to the unit *předu*. In the case of *do leskla, na zrz, z nenadála* only the POS=V interpretation is in the morphological dictionary.

3.1 Addition into the Morphological Dictionary

We believe that one of the ways to improve automatic morphological analysis is to add data to the morphological dictionary. We have selected 470 units from the corpus probe, but not all of them are suitable for the morphological dictionary, for several reasons. Some expressions are not considered adverbial, because adverbialization has not occurred, there is only missing space when writing this expression (e.g *ožívot, narozloučenou, . . .*).

We also recorded expressions that are compound, but we do not consider them as an adverb (e.g. *doboha*: interjection). In one case the obtained form resembled compound adverb structure, but we came to the conclusion that it is a verb form (*zamražena*: verb). We have recorded expressions which we do not consider to be compound adverbs and are listed in existing dictionaries as another part of speech (*mezitímco/mezitím co*: conjunction, *naprostřed*: preposition). On the other hand, these units are not listed in morphological dictionary and may be added there as a different part of speech (not compound adverb).

Some expressions are compound adverbs, but we understand them more as occasionalism and they occur in the order of units (e.g. *narub, pokopě, vnedohlednu, . . .*). For this reason, we have set a minimum frequency of 15 occurrences in the corpus SYN v3 to add the word into the morphological dictionary. Otherwise, because 15 occurrences are no longer 0 i. p. m. but 0.01 i. p. m. The random check in the corpus SYN v6 showed that in many cases the occurrences of analyzed compound adverbs are very similar.

We propose to add into morphological dictionary those expressions that are demonstrably compound adverbs, the process of adverbialization is either completed or ongoing, and the occurrence frequency is greater than 15. Furthermore, we propose to add into the morphological dictionary the expressions we have found in existing dictionaries (DEBDict) [7], regardless of the frequency of occurrence and part of speech. We also propose to add those units with frequency higher than 15 which we identified as a different part of speech than adverb. The proposal for addition in the morphological dictionary always contained lemma and part of speech interpretation.

Altogether, 177 units were proposed for addition in the morphological dictionary, their number and the part of speech interpretation was as follows:

POS=D, SUB=s, compound adverb, 103 units, (e.g. *domodra*)
POS=O, SUB=s, oscillating, compound, 43 units, (e.g. *modro*)
POS=C, numeral, 20 units, (e.g. *našestkrát*)
POS=D, adverb, 4 units, (e.g. *tuty*)
POS=R, preposition, 2 units, (e.g.*naprostřed*)
POS=I, interjection, 1 unit, (*doboha*)
POS=J, conjunction, 1 unit, (*mezitím*)
POS=T, particle, 1 unit, (*naviděnou*)[5]

[5] We do not consider *naviděnou* as a compound particle. We proposed this unit to be add into the morpohological dictionary because its one-word form is common. Similar case is e.g. *nashledanou*, also tagged as POS=T.

POS=N, noun, 1 unit, (*podmíru*, lemma *podmíra*)
POS=V, verb, 1 unit, (*zamražena*, lemma *zamrazit*)
Number of proposed units is 177, number of analyzed units is 470.

3.2 Change of Morphological Tag

There are two facts which led us to suggest to change the morphological tag: First, there is no tag in the tagset of the Czech National Corpus [9] that indicates the compound adverb and second, there are many words, which we consider to be compound adverbs, that are marked inconsistently. Examples of inconsistent tagging:

na tvrdo: preposition, adjective
na žluto: preposition, noun
na tajno: preposition, adverb

We find a satisfactory solution in the concept of the NOVAMORF project [10], which proposes a new part of speech type: POS=O: an oscillating part of speech. For POS=O, we consider those forms that are ambiguous whether they are nouns, adjectives, or adverbs (e.g. *sucho, mokro, modro, ...*). We also propose to add a subset of the SUB=s meaning compound to adverbs and numerals.

We suggest therefore to tag one-word compound adverbs as POS=D with specifying a type compound as SUB=s (e.g. *namodro*: POS=D, SUB=s). We propose to tag multiword expressions of type *na modro* as *na* POS=R, *modro* POS=O, SUB=s.

In connection with the introduction of a new tag for a compound word, the question arises as to whether this addition should be added to all the part of speech in which the compound word can occur. These would be adverbs, numerals, as well as interjections, prepositions or conjunctions. With a view to the consistency of tagging, we think the adding the tag for a compound is useful, but only for adverbs and numerals. Compound interjections (e.g. *proboha, doboha, ...*), prepositions (e.g. *naprostřed*) or conjunction (e.g. *mezitím*) are very few.

3.3 Collocations

By analyzing data, we have found that some compound adverbs are found in collocations, some of which are part of phrases and idioms, and are recorded in the Idiomatic and phrasal dictionary (DEBDict) [7,8]. Nowadays the Multiword Expression Lexical Database (MWELD) is built by Petkevič et al. [11] and we find very useful to enlarge this database with our data. Larger the MWELD is, better results in disambiguation can be reached.

4 Conclusion

We have focused on compound adverbs from the automatic morphological analysis point of view. Compound adverb tagging is a non-trivial problem

because compound adverbs are written mostly together as one word, but often in parallel there exists a multiword expression and their meaning is the same (e.g. *na příklad – například*).

We made a corpus probe on corpus SYN v3 and we searched for unrecognized forms that can be considered as being compound adverbs. Thanks to CQL queries, we have obtained large data. We have sorted it manually according to prefix and then by ending. We selected 470 units we consider as compound adverbs. Afterward, we were interested whether the AMA recognizes these expressions if we respread it into multiword expressions. Both one-word and multiword expressions were checked in existing Czech dictionaries. We also focused on strong collocations of chosen units.

By analyzing the corpus data we suggest three solutions to improve the compound adverbs tagging: First is to enlarge morphological dictionary by adding units which are demonstrably compound adverbs and which frequency of occurrence is more than 15 in corpus SYN v3. Altogether we have found 103 units to be added into the morphological dictionary as a compound adverb and others 74 units as others part of speech. Second, we propose in accordance with NOVAMORF project a new compound adverb tagging. We propound a new type of part of speech such as POS=O, oscillating part of speech, and also new subset SUB=s, means compound. We suggest tagging subset compound type not only with adverbs but also with numerals.

We are aware that the proposed solutions do not cover the complete issue of compound adverb recognition, but we believe that the corpus probe and the proposed solutions can contribute to an at least partial improvement of the AMA in this area.

References

1. Internetová jazyková příručka, `http://prirucka.ujc.cas.cz/?id=130`
2. Dokulil, M. Tvoření slov v češtině, díl 1 Teorie odvozování. pp. 22. Nakladatelství Československé akademie věd, Praha (1962)
3. Cvrček, V. Mluvnice současné češtiny. Karolinum, Praha (2010)
4. Multiword Expressions. In: Wiki of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg PA, USA (2016).
5. Petkevič, V. Problémy automatické morfologické disambiguace češtiny. Naše řeč, 97(4/5), pp. 194–207 (2014)
6. Křen, M., Čermák, F., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová V., Skoumalová, H., Šulc, M. Korpus SYN, verze 3 z 27. 1. 2014. Ústav Českého národního korpusu FF UK, Praha (2014)
7. Horák, A., Pala, K., Rambousek, A., Povolný, M. DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Proceedings of the Third International WordNet Conference – GWC 2006. pp. 325–328. Masaryk University, Brno (2006)
8. Čermák, F. Slovník české frazeologie a idiomatiky. Leda, Praha (2009)
9. Hajič, J., Cvrček, V., Chlumská, L. Morfologické značky (tagy) In: Wiki Český národní korpus. FF UK ÚČNK, Praha (2017)

10. Osolsobě, K., et al. Nová automatická morfologická analýza češtiny. Naše řeč, 100(4), pp. 225–234 (2017)
11. Petkevič, V. et al. Lexicon and Multiword expressions. In: Slavicorp2018 Book of abstracts. FF UK, Praha (2018)