Comments on Czech Morphological Tagset

Karel Pala

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pala@fi.muni.cz

Abstract. In the area of natural language processing the appropriate morphological annotation is necessary. In this paper we offer some comments on the Czech morphological tagset as used in the analyzer Majka that has been developed in NLP Centre (CZPJ) both for academic and commercial purposes relatively recently. We try to argue that the existing approach to the morphological annotation is not in agreement with the language reality and that the used solutions are motivated rather technically than theoretically. We think that it makes sense to consider some changes of the presently applied annotation principles that might, if applied, to improve the annotation accuracy.

Keywords: part-of-speech tagging; morphological analysis

1 Introduction

Morphology is as a rule the base for a number NLP applications and it is obvious that its descriptive adequacy is heavily determined by the annotation principles and consequently by the tagset depending on them.

For Czech, two tagsets have been available since the 90's, developed by two leading NLP groups: one produced in the Institute of Formal and Applied Linguistics at the Charles University in Prague [1] and another one in the NLP Centre at the Masaryk University in Brno [5].

In this paper we refer to comments on the version of the second tagset together with the underlying morphological database used by the analyzer Majka [7,8]. The tagset can be found in the Appendix B of the paper by Jakubíček et al. [2]

2 Annotation principles

The question that is essential: should the principle on which the used morphological annotation is based be holistic or partial? The question is not touched in Jakubíček's (et al) paper but it can be seen that linguistically complex expressions not behaving compositionally as e.g. *Karlovy Vary*, *vzhledem k (with regard to)*, *jestliže, ... pak (if ... then)*, *a to (and this)* or *budu číst (I will read)* are simply taken apart and later put together again, thus being analyzed twice. The approach

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 15–19, 2018. © Tribun EU 2018

does not reflect the language reality and consequently generates a significant number of disambiguating errors. The main argument used for it is based on technical aspects of the annotation only and in our view is evading the essence of the annotation problem in Czech as such.

There is a convincing frequence evidence proving that so far used partial approach to the morphological annotation can hardly lead to an essential improvement of the present situation with annotation accuracy.

In our view the indicated problems with the disambiguation accuracy certainly include the following parts of speech:

PoS	tag	CzTenTen	Desam
adverbs	k6	278,172,710	49,469
conjunctions	k7	447,920,261	99,432
particles	k8	324,980,597	52,951
verbs	k5	694,012,081	126,067

As a data we use corpora CzTenTen12 with 4,175,089,441 tokens and also Desam with 874,354 tokens which show that the mentioned four parts of speech display really high frequences. This means that they (k5, k6, k7, k8) represent a large source of the disambiguation difficulties because of their strong mutual polyfunctionality. Apart from the high frequencies there is another relevant cause of the disambiguation errors (possibly up to 10 %) – many of these parts of speech are MWES as the examples above show, e.g. frequency of *a to* is 3,014,697 in CzTenTen12. The indicated simple facts can be generalized to support the thoughts about the possible changes of the so far used partial disambiguation strategy.

2.1 An example with MWE *a to* (and this)

We are well aware that proposing changes to an existing and established tagset can be understood as an unpopular step that implicates compatibility issues with the older tagset version(s). However, we would like to show an example taken from the corpus Desam that indicated holistic approach to MWEs with regard to disambiguation is worth of consideration.

We have chosen Czech MWE conjuction *a to (and this)* and found that it is tagged in manually disambiguated corpus Desam in a rather conflicting way as the Table 1 shows. We can see that 100 tokens of *a to* is tagged in different ways, particularly, we can see that *a (and)* is tagged as k8, k9 or not tagged at all. Similarly, the tagging of the second part of MWE *to (this)* varies considerably too as k3, k8 and k9, or not tagged at all.

Notice that the indicated conflicts would disappear if we decide to treat *a to (and this)* (and similar expressions) as one MWE unit which should be in this case tagged as a conjunction (k8).

Table 1: Various tags for *a to* and their frequency in Desam corpus.

not tagged	13
k8 + k3	60
k9 + k3	19
k8 + k9	3
k9 + k9	5

2.2 A Comment on Verbs (k5)

Few words should be said about verbs. They display the highest frequency among the mentioned parts of speech (in CzTenTen12 694,012,081 tokens, in Desam 126 067 tokens), thus they will be affected by the holistic approach relevantly.

In Czech this includes all analytical verb forms, i. e. future tense forms of imperfective verbs *budu číst (I will read/will be reading)*, past tense forms *četl jsem (I read/have read/was reading)*, conditional forms *četl bych (I would read/)*, which should be tagged as single units. It is obvious that it would require the massive re-tagging though we have to admit that the current tagging of the analytical verb forms does not generate too many disambiguation errors. However, if we want to be in concordance with the language reality we have to think seriously about the re-tagging of verbs. On the other hand, we realize that this re-tagging will considerably influence the structure of the existing parsers for Czech such as Synt or Set [3].

The question is if anybody can be found who would be able to try to undertake such demanding task. The language reality speaks for treating analytical verb forms consistently as single units, however, the technical consequences for a respective syntactic analysis would be extensive. It is, however, possible to go for a compromise and deal with the verb analytical forms in their current shape.

2.3 A Remark on Adverbs (k6), Prepositions (k7), Conjuctions (k8) and Particles (k9)

These frequent parts of speech are not inflected and as a rule they are difficult to disambiguate thanks to their high mutual ambiguity. They are not easy to classify because they often pass over between the individual categories and they can be disambiguated relativel reliably only in the particular contexts, which is sometimes difficult even for humans. As we could see above, it often happens that the conjunctions are recognized as particles, see *a to* above or adverbs as complex prepositions, e.g. *pokud (unless)*. It would be desirable to go through these ambiguities manually and disambiguate them, however, it is obviously not real because of their high frequency. Perhaps it would be helpful to collect

Table 2: Mapping of the Czech tagset to the Google Universal Tagset

universal tag	description	attributive tags
VERB	verbs (all tenses and modes)	k5.*
NOUN	nouns (common and proper)	k1.*
PRON	pronouns	k3.*
ADJ	adjectives	k2.*, k4.*xO, k4.*xR
ADV	adverbs	k6.*
ADP	adpositions (prepositions and postpositions)	k7.*
CONJ	conjunctions	k8.*
DET	determiners	(none)
NUM	cardinal numbers	k4.*xC
PRT	particles or other function words	k9.*
X	other: foreign words, typos, abbreviations	k0
.	punctuation	kI

all these PoSs and keep them as the particular lists making them a part of the database of Majka analyzer.

In existing Czech grammars, e.g. Karlík et al. [4], we can find subclassifications of the mentioned parts of speech – k6, k7, k8, k9, which are quite detailed but partly a bit overlapping. It would be useful to compare it with the corresponding subclassifications in Jakubíček's paper but this would be topic for a separate paper.

We would like to point out that in Jakubíček's paper some PoS' and their details are left aside, particularly prepositions (k7) and conjunctions (k8). It has to be stressed that they also include a number of MWES that should be re-tagged as single units. An attempt in this direction can be found in [9].

2.4 Mapping to the Google Universal Tagset

We decided to refer here to a mapping to the universal tagset created by joint effort of Google Research and Carnegie Mellon University [6]. The mapping is given in Table 2 and we take it over from Jakubíček's paper.

To remark: we assume that the comparison of the tags above should not be much influenced by the considered change of the annotation principles. In this respect we do not expect any relevant modification of the presented lists of tags.

3 Conclusions

We have offered some comments dealing with theoretically motivated changes to the attributive tagset for Czech language. Implicitly, we react to the paper by Jakubíček et al, however, we think that the revision proposed in it is not sufficient and that there is a time to consider more essential, holistic revision of the tagging principles and consequent re-tagging of existing corpora. We are well aware that in fact such revision would be a painful and expensive new project as well but the challenge is here. This should lead to essential changes in tagging results and not only for Czech language.

The question remains whether the techniques exploiting neural networks will be able to deal with MWES in a holistic and descriptively adequate way. This too represents a topic for a separate paper.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781.

References

1. Hana, J., Zeman, D., Hajič, J., Hanová, H., Hladká, B., Jeřábek, E.: Manual for Morphological Annotation PDT. Tech. Rep. 27, Institute of Formal and Applied Linguistics, MFF UK, Prague, Czech Republic (2005)
2. Jakubíček, M., Kovář, V., Šmerk, P.: Czech morphological tagset revisited. Proceedings of Recent Advances in Slavonic Natural Language Processing pp. 29–42 (2011)
3. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for czech. In: Human Language Technology. Challenges for Computer Science and Linguistics. pp. 161–171. Springer, Berlin/Heidelberg (2011), `http://dx.doi.org/10.1007/978-3-642-20095-3_15`
4. Nekula, M., Rusínová, Z., Karlík, P.: Příruční mluvnice češtiny. Nakladatelství Lidové noviny (1995)
5. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97. pp. 523–530. Springer-Verlag (1997)
6. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. Arxiv preprint ArXiv:1104.2086 (2011)
7. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of the RASLAN Workshop 2009. Brno (2009)
8. Šmerk, P.: Towards Computational Morphological Analysis of Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (August 2010)
9. Žižková, H.: Improving Compound Adverb Tagging. In: Proceedings of the RASLAN Workshop 2018. Brno (2018)