The Tagged Corpus (SYN2010[1]) as a Help and a Pitfall in the Word-formation Research

Klára Osolsobě

Ústav českého jazyka FF MU

Arna Nováka 1, 602 000 Brno, Czech Republic

osolsobe@phil.muni.cz

Abstract

Today, language corpora are the primary source of linguistic observation. The purpose of this paper is to illustrate some of the problems associated with word-formation research based upon morphologically tagged synchronous corpora. Three problems emerged during work on the linguistic handbook *Dictionary of affixes used in Czech – Slovník afixů užívaných v češtině* (Šimandl et al, 2016): a) tokenization, b) lack of the morphological dictionary, and c) POS tagging. This paper describes the solutions utilized by the authors of the dictionary in response to the above listed problems. These solutions, used in SAUČ, simplistic as they may seem, resulted in particular suggestions for improvement of the automatic morphological analysis of Czech, conducted as a part of the NovaMorf project (Osolsobě et al. 2017).

1. Introduction

The *Dictionary of Affixes used in Czech* (hereinafter referred to as SAUČ) is a new manual, which as well as the printed version issued by the Karolinum publishing house, is also available in a free electronic version (http://www.slovnikafixu.cz/). The entries are the product of thirteen contributory authors which reflect in minor variations due to authors' individual writing styles.

The dictionary is sorted alphabetically by the first letter of the particular affix (the header of the entry located on the left – prefix, in the middle – associated affix, on the right – suffix). A brief morphological characterization of the words formed by the respective affix (information about the inflection and alternations) follows. The text section summarizes information about the structural meanings / word classes corresponding to the analysed affix, the individual meanings being numbered. Respective entries referring to native and loaned affixes include, a so-called, "frequency report".[2] Both parts are based on the analysis of data accessible through the SYN2010 language corpus. In the textual part, the authors also relied on a variety of sources (native speakers intuition or opinion, other corpora, internet).

The SAUC preface further states: "When we use this dictionary, we can concentrate upon the affix system depending on their frequency or productivity."

When examining affixes, the lexeme, as a unit of the language system (*langue)*, is analysed. Therefore, the lemmatised and POS tagged corpus would seem to be helpful; however the lemmatisation and POS tagging are the result of automatic morphological analysis and therefore it is of importance that SAUČ' users have knowledge of automatic morphological analysis. We shall demonstrate, using specific examples taken from the SAUČ, how the results of automatic tagging become pitfalls for corpus based linguistic research (part 2.). In conclusion, (part 3.) of the article, will demonstrate the benefits of working with data for further development of automatic morphological analysis tools, specifically within the NovaMorf project.

[1] Corpus SYN2010 is a synchronous representative corpus of contemporary written Czech containing 100 million text words. For more information, see http://wiki.korpus.cz/doku.php/cnk:syn2010.

[2] Part of entry differentiated by a font type. At the beginning there is corpus query, by which the corpus data had been obtained, the number of hits and their relevance. A section "20 most frequent lemmas" (for sparsely documented affixes, all lemmas are listed) follows. For each lemma reference is made to the meaning (corresponds to numbered meanings in the text above) and the number of occurrences (frequency) of the lemma in the analysed corpus.

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 55–59, Prague, Czechia, 19-20 September 2019.

2. Three steps of the automatic analysis[3]

Automatic morphological analysis generally involves three steps, namely, tokenization, the assignment of linguistic interpretations in the form of a lemma[4] and a tag[5], and disambiguation (if the word form analysed is both word and / or morphologically ambiguous / homonymous, and therefore the interpretations assigned are greater in number based on the dictionary).

All three of these steps are done by the engine automatically and have an impact on the final form of the morphological interpretation (lemma and POS / tag), and hence on all linguistic research based on the corpus. In the following sections, we will show how the use of the results of the automatic analysis has affected the work on the SAUČ.

2. 1 Pitfall No 1: The tokenization[6]

The tokenization is the process of demarcating and possibly classifying sections of a string of input characters. The resulting tokens are then passed on to some other form of processing. The process can be considered a sub-task of automatic morphological analysis. In corpus linguistics, a number of difficulties related to the reduction of the word form to a graphically defined unit (a string of defined alphabet characters separated from both sides by separators) are dealt with as a part of Multiword Expressions (MWE)[7] processing.

The affixes described in SAUČ are usually graphically a part of a single lexeme. Unlike German, there are no separable prefixes – '*trennbare Präfixe'* in Czech, but there are some cases that are somewhat analogous. Some Czech adverbs originate from prepositional phrases of names (nouns: *na konec → nakonec – ultimately*), nominal forms of adjectives: *do cela → docela – quite*, *z blízka → zblízka – close*, pronouns: *po tom → potom – then*, *přede vším → především – above all*, or numerals: *za prvé → zaprvé*, *z prvu → zprvu – first*). The creation of the compound adverb (*adverbiální spřežka*) is typically gradual. The completion of the process of the adverbialization is not only the graphical realization of a compound adverb as one graphical unit, but it is no more possible to insert another word form between the preposition and the nominal form. Usually the two ways of writing (two graphical units / one graphical unit) coexist.[8] The formation of these types of adverbs could be considered as associated affixation, in which the original preposition takes the role of the new prefix and the original ending takes the role of the new suffix. However for the POS tagging the preposition that is not graphically united with some newly created adverb is an independent unit tagged as a preposition and its nominal part is very often not identified (A lot of nominal parts of compound adverbs are no more used, and therefore the lemmas are not included in the dictionary of respective morphological analyser cf. Žižková, 2017.).

An example is the entry ***na- -o*** (http://www.slovnikafixu.cz/heslar/na-%20-o).

The text states: "Typically, an adverb is characterized by dual writing, cf. *na černo / načerno* (*black, blackly*), *na hrubo / nahrubo* (*coarse, coarsely*), *na měkko/ naměkko* (*soft, softly*). The first way of writing is here essential. Whereas only "written together variants" are included in the frequency

[3] https://www.czechency.org/slovnik/MORFOLOGICK%C3%81%20ANAL%C3%9DZA
[4] https://www.czechency.org/slovnik/LEMMATIZACE
[5] https://www.czechency.org/slovnik/TAGSET
[6] https://www.czechency.org/slovnik/TOKENIZACE
[7] https://www.aclweb.org/aclwiki/index.php?title=Multiword_Expressions. In the Czech environment, the entries of lexicon units that go beyond the graphic unity of the word form is systematically examined by prof. František Čermák, founder and a long-time director of the Institute of the Czech National Corpus (cf. a series of entries dedicated to phraseology, idiomatics and collocations in Czech: https://www.czechency.org/slovnik/autor/F%C4%8C).
[8] See: http://prirucka.ujc.cas.cz/?slovo=natvrdo#bref1.

report." The same way is also followed in analogous entries (e. g. ***do- -a*** http://www.slovnikafixu.cz/heslar/do-%20-a). Yet this way can be considered questionable with regard to both frequency and productivity research. The results of the frequency report do not show the frequency of the adverbs formed by the fusion of the preposition and the nominal form (often the name is not documented as a separate word outside the collocation with the corresponding preposition), but only the frequency of one of the graphical variants (in addition, only variants stored in the automatic analyser dictionary, see below).

The second example covers the entries describing such affixes that form verbs by prefix and reflexive particle *se/si*, e. g. *myslet → zamyslet se* (*think → reflect* as an intransitive verb). Let's have a look at the entry *za- se* (http://www.slovnikafixu.cz/heslar/za-%20se). We read in the text: "The statistical report was created by manual editing, with limited data accuracy guarantees."

What's going on? The free word order in Czech allows that the reflexive particle *se/si* can be separated by several word forms from its corresponding verb. The proper place of the particle *se/si* from the verb to the left side is not limited (it is driven by the principles of sentence stress). The proper place of the particle *se* from the verb to the right side is at the maximum on the third position (between the verb and *se* only some short words – clitics can be inserted). But the manual selection of each word order variant would be very time-consuming and probably inaccurate . We considered, for the sake of frequency report, the two most frequent word order variants (variants with the particle *se* immediately before / after the verb). It is, admittedly, a simplistic approach.

2. 2 Pitfall No 2: Assigning lemma + tag interpretation based on the morphological dictionary[9]

The second step of the automatic morphological analysis is to assign all interpretations based on the dictionary of the automatic morphological analyser. The lemmatization results depend on the scope and content of the respective dictionary. Although the dictionary is extensive and growing, the number of hapax[10] expressions in any new text is constantly variable. The productivity measuring is dictionary-dependent.

Now we can return to the case of compound adverbs mentioned above. In the dictionary only some (presumably codified) compound adverbs are stored. If we repeat the query **lemma="na.*o"** (http://www.slovnikafixu.cz/heslar/na-%20-o) with the omission of the morphological tag specification, we obtain more relevant lemmas (e. g. adverbs as *natěsno – tight* or *tightly, nakratičko – short* or *shortly, naneurčito – vague*) that will not appear in the frequency report. These are words of low frequency, but they correspond to the model of such type of compound adverbs in Czech and show its productivity. The productivity picture based on the results of automatic analysis is inaccurate.

Is it possible to overcome the limitations of a dictionary? At this point we can focus on the entry ***-oš*** (http://www.slovnikafixu.cz/heslar/-o%C5%A1).

It is clear from the queries[11] that not only data based on the results of POS tagging were taken into account. For an entry that describes an affix, with which expressive words (hypocoristic proper names as *Miloš, Leoš, Antoš*)[12] are derived, this is an appropriate strategy. The examples given to illustrate the second query would indicate, that if we were not doing so, productivity would be significantly skewed.

[9] https://www.czechency.org/slovnik/ANOTACE

[10] https://www.czechency.org/slovnik/HAPAX

[11] Query [lemma=".*oš" & tag="NN[MI].*"] gives 125 lemmas, 69 are relevant. Query [lemma="(.*oš)|(.*oš[eiů])|(.*oších)|(.*ošům) & tag="X.*"] gives 282 words, 36 relevant lemmas.

[12] The dictionary was built particularly for analysing written language. In the dictionary of automatic morphological analyser the expressive vocabulary (e. g. hypocoristic proper names) is rather neglected.

2. 3 Pitfall No 3: The disambiguation[13]

The last step of automatic morphological analysis is disambiguation (the process of identifying which interpretation of a word is used in context). Its results depend on the method of disambiguation. The biggest problem here is homonymy[14] (cf. Petkevič, 2015). In the case of word-formation, the problem of homonymy affects cases of part of speech transition, polyfunctional affixes, and overgeneration of formal query. Corpus analysis results are „disambiguation-addicted".

The problem of homonymy illustrates the entry ***-cí*** (http://www.slovnikafixu.cz/heslar/-c%C3%AD). We read in the text: "… adjectives formed by suffix *-cí* are in many cases nominalised – they have the meaning of (3) agentive names".

In the frequency report, this meaning (the meaning number (3)) is not differentiated by the number of occurrences of nominalised usage. The reasons behind this decision are as follows: 1) in the dictionary of the automatic analyser the nominalised adjectives are stored rather unsystematically. Except for some frequent lexemes (e. g. *vedoucí – leading* or *leader*, *kolemjdoucí – passing* or *passenger*, etc.), most lemmas have only one (adjective) POS interpretation (cf. Richterová, 2017, Žižková, 2019). 2) The potency of adjectives to transform into nouns is almost unlimited, and moreover, the boundaries can't be defined only in terms of the dictionary, since in many cases we have to deal with contextual ellipses. 3) The disambiguation (see Figure 1) is far from satisfactory (the hints on lines number 3, 5, 8, 10 are disambiguated wrongly).

Figure 1

1	člověk (nebo někdo , kdo jako člověk vypadá)	cestujici/cestujici/AGFP4-----A-----	s Nadiankou mnohem podezřelejší . Auto zastavilo mezi poli zarostlými
2	ho zaměni za celodenní permanentku . Službu může využít každý	cestujici/cestujici/NNMS1-----A-----	, který se zdržuje na Slovensku a zjisti , že
3	Na Dálném východě je kontrola jízdenek zdlouhavá záležitost , protože	cestujici/cestujici/AGIP1-----A-----	mají lístky poschovávané na těch nejpodivnějších místech . Kim předložil
4	nesmí být příliš . Zřízení vízové povinnosti pro české občany	cestujici/cestujici/AGMP4-----A-----	do Kanady zásadně utlumí příliv Romů s naději klepajících na
5	"! Zatraceně ! "" Doktor podal zabalené nemluvně jednomu z"	cestujicich/cestujici/AGMP2-----A-----	, pak se začal posouvat směrem k dusící se ženě
6	kličku , hlasové ovládání , televizi v hlavových opěrkách pro	cestujici/cestujici/AGMP4-----A-----	na zadních sedadlech a další vymožeností . V každém případě
7	, a ustoupil od dveří , aby udělal místo druhému	cestujicimu/cestujici/NNMS3-----A-----	, který právě vystupoval z limuzíny . Malone ztuhl .
8	letenky by závisela na celkové váze jak zavazadla , tak	cestujiciho/cestujici/AGMS2-----A-----	. Podle mluvčího americké Letecké asociace (ATA) Davida
9	aut . Aby nedělní změny nevyvolaly chaos , doporučují dopravci	cestujicim/cestujici/NNMP3-----A-----	spoléhat se hlavně na takzvané páteřní linky . V oblasti
10	faktorů hodnocených v analýze scénářů . V případě požáru jsou	cestujici/cestujici/AGMP1-----A-----	v tunelu vystavení účinkům tepla a toxickým a dráždivým plynům

The problem of overgeneration is illustrated by the lines below the frequency report referring to the lemmas, which doesn't correspond to the words created by the affix. For example in the entry ***sou- -í*** (http://www.slovnikafixu.cz/heslar/sou-%20-%C3%AD) 28 lemmas (e. g. *soustředění – concentration, soužití – coexistence, soutěžení – competition, soužení – suffering/problem, sousedství – neighborhood, soukromí – privacy*) were excluded. We would like to finish with a (politically incorrect) language joke based on overgenerated segmentation: "*Několik soch je **sou-soš-í**, několik žen je **~~sou-žen-í~~/s-ouž-en-í**"* ("*Several sculptures create a sculptural group, several women create a problem."*).

3. Conclusion

Our goal was to show that the authors of the *Dictionary of affixes used in Czech* were very well aware of the limits of working with the results of automatic part of speech tagging. We added detailed commentaries concerning simplistic solutions for the dictionary readers. All data (unless otherwise indicated) referring to the corpus are taken from the reference corpus (SYN2010) and the method of data mining is sufficiently described at the beginning of the frequency report (the corpus query). Therefore, every dictionary user can repeat the query with the same reference corpus or with different data (other corpora). The SAUČ as a whole meets the requirements for empirical testability of the presented results as required by the corpus linguistics. Despite the above-mentioned simplistic solutions, it is not disputed that without using the results of automatic tagging, any way of creating the *Dictionary of affixes used in Czech* would be a) incomparably more time-consuming, b) more expensive and c) in its result less objective.

[13] https://www.czechency.org/slovnik/DISAMBIGUACE%20/%20DESAMBIGUACE

[14] https://www.czechency.org/slovnik/HOMONYMIE

Nevertheless the problems which emerged during work on the *Dictionary of affixes used in Czech* become a starting point for research oriented towards improving automatic morphological tagging (see more Žižková, 2017, 2019). A detailed morphological description of word forms based on the data gained during the work on SAUČ is reflected in the *NovaMorf* project (Osolsobě et al. 2017).

Acknowledgement

This work was supported by the project MUNI/A/1061/2018 Čeština v jednotě synchronie a diachronie – 2019.

References

Hajič, J., Hlaváčová, J. 2016. *MorfFlex CZ*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

Karlík, P. – Nekula, M. – Pleskalová, J. 2016. *Nový encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha. https://www.czechency.org/slovnik/.

Křen, M. – Bartoň, T. – Cvrček, V. – Hnátková, M. – Jelínek, T. – Kocek, J. – Novotná, R. – Petkevič, V. – Procházka, P. – Schmiedtová, V. – Skoumalová, H. *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2010. http://www.korpus.cz.

Osolsobě, K. – Hlaváčová,J. – Petkevič, V. – Šimandl, J. – Svášek, M. 2017. Nová automatická morfologická analýza češtiny. *Naše řeč* 100 (4): 225-234.

Petkevič, V. 2015. *Morfologická homonymie v současné češtině.* Nakladatelství Lidové noviny, Praha.

Pravdová, M. – Svobodová, I. 2014. *Akademická příručka českého jazyka.* Academia, Praha. http://prirucka.ujc.cas.cz/.

Richterová, O. 2017. *Od slovesa ke jménu a předložkám Deparcipiální formy v češtině: forma, funkce, konkurence. From Verbs to Nouns and Prepositions. Deparcipial Forms in Czech: Form, Function, Complementarity*. Ph.D. Thesis. FF UK, Praha.

Šimandl, J. (ed.). 2016. *Slovník afixů užívaných v češtině*. Karolinum, Praha. http://www.slovnikafixu.cz/index.

Žižková, H. 2017. Compound Adverbs as an Issue in Machine Analysis of Czech Language. *Jazykovedný časopis* 68 (2): 396-403.

Žižková, H. 2019. *Slovnědruhové přechody jako problém automatické morfologické analýzy. Part of speech transitions as a problem of automatic morphological analysis*. Ph.D. Thesis. FF MU, Brno.