„Bohemistyka” 2022, nr 1, ISSN 1642–9893

Dana HLAVÁČKOVÁ , Hana ŽIŽKOVÁ
Masaryk University
Klára DVOŘÁKOVÁ, Markéta PRAVDOVÁ
Czech Academy of Sciences

DOI: 10.14746/bo.2022.1.7

Developing Online Czech Proofreader Tool: Achievements, Limitations and Pitfalls[1]

Keywords: web spell-checker and grammar-checker, automatic proofreading, Czech, orthographical/typographical/grammatical errors, error correction

Abstract

This paper deals with achievements, limitations and pitfalls of developing the Online Czech Proofreader Tool (OCPT). The Tool has been developed in cooperation with the Department of Czech language of the Faculty of Arts of Masaryk University, the Institute of Theoretical and Computational Linguistics of the Faculty of Arts of Charles University, the Czech Language Institute of the Czech Academy of Sciences and Seznam.cz since 2019. The article describes the linguistic data used and tools and modules that constitute the OCPT and indicates the limitations of using an online web-based proofreader tool, especially in areas where mere application of formal rules for language error detection is not sufficient. The article also brings up the drawbacks of developing the OCPT which include occurrence of false-positives.

0. Introduction

Continuous development of computational linguistics brings along improvement of various natural language processing tools. In the

[1] This work has been supported by the Technology Agency of the Czech Republic under the project TL02000146. The work described herein has also used data provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Czech language environment development of programmes for automated correction of Czech texts began in the mid-1980s, when KORA[2] spell-checker was created and applied in "T602" text editor. More tools – capable already to not only check for typos but also for orthographical and grammar errors – were developed around 2000. These are:

- *Czech language grammar checker* developed in the Czech Language Institute of the Czech Academy of Sciences (Petkevič, 2014) and used until now in the Microsoft Word editor and integrated into the Microsoft OfficeTM system,
- a system called *Grammaticon*, developed by the company Lingea, also for purposes of Microsoft Office (not available anymore).

Apart from these two, there is a *Czech language dictionary for orthography and hyphenation checks and thesaurus* for LibreOffice as well as a thesaurus created as an add-on to LibreOffice. Lately, *Correct your spelling & grammar in Google Docs* (*Opravy pravopisu…* 2021) for Czech has appeared and is being developed based on machine learning. However, the latter two cases represent a mere spell-checker.

Since 2019 Department of the Czech Language of the Faculty of Arts of Masaryk University, Institute of Theoretical and Computational Linguistics of the Faculty of Arts of Charles University, Czech Language Institute of the Czech Academy of Sciences, and Seznam.cz have been cooperating on developing the web-based Online Czech Proofreader Tool (OCPT). When completed, the Tool will be available for free to the general public. Currently, a demo version of the OCPT can be accessed at https://korektor.plin.cz/demo.html.

The OCPT aims to correct typos and orthographic, grammatical and typographic errors. It should also indicate certain stylistic flaws and should even provide the user with an explanation by clicking through to *Internet Language Reference Book* (ILRB)(Internet Language Reference Book, 2008–2021). Developing the OCPT on this level of complexity brings many challenges which can be addressed more or less successfully. In this paper we aim to describe some of our achievements, but also chosen limitations and pitfalls.

[2] KORA – linguistic aspects of automated corrections were designed by Klára Osolsobě, their implementation was executed by Zdeněk Bouša and Milan Šárek.

1. Achievements

Development of the OCPT is based on an analysis of the most common errors in Czech language and on lists of the most frequent words and word forms searched for in ILRB. In addition, we have also created our own annotated error corpus.

The OCPT utilizes rule-based error checking and the rules form five distinct correction modules (punctuation, linguistic agreement, ungrammatical sentence constructions, typographical errors and other types of linguistic errors). Typos are identified by a spell-checker equipped with an extensive dictionary of Czech word forms. A large amount of linguistic data used for the development of the OCPT is freely available, was gained through our cooperation within the scope of this project (e.g. data from ILRB) or was created by us from scratch (e.g. anotated error corpus for internal use).

We are testing the success rate of individual rules and our modules continuously. The best results have so far been achieved by the punctuation correction module, using MorphoDiTa tagger, with a precision of 94.97% and a recall of 63.63%. Testing of the OCPT as a whole and evaluation of its success rate is planned for the upcoming months.

2.1. Linguistic data

To a significant extent the quality of a correction of a text depends on the quality and size of the employed linguistic data (dictionaries, lists etc.). As Petkevič (2014, p. 4) states, to develop a programme for a simple spell-check is relatively easy once a complete list of all the existing word forms of the language available.

The necessity of a high quality lexicon coverage is also affirmed by the enquiries recorded in online database of enquiries addressed to *Language Consultation Centre* (LCC)[3] of the Czech Language Institute (Language Enquiry Database, 2016–2021) or by Svobodová (2019, p. 250), who mentions that LCC´s clients sometimes inquire about expressions or word forms underlined by their spell-check.

Considering this we came to the conclusion that we need an extensive dictionary for the OCPT. At the moment the OCPT is using:

– two morphological dictionaries (Majka, MorfFlex)
– its own spell-checker dictionary
– extensive collocation lists
– list of words and word forms included in ILRB
– list of the most frequent expressions entered incorrectly into ILRB
– list of proper nouns comprising anthroponyms, toponyms, oikonyms
– and more

However, when we combined the dictionaries of Majka tagger (Šmerk, 2014), MorfFlex dictionary (Hajič, 2020), and ILRB (Internet Language Reference Book, 2008–2021) dictionary for a use by the OCPT, we obtained a dictionary of 22.8 mil. word forms. With such size, the dictionary was slowing the OCPT down. That is why we decided to reduce the dictionary to 2.5 mil. word forms (it is an intersection of the combined dictionaries of Majka, MorfFlex and ILRB with the frequency list from corpus SYN v9[4]). As a part of this dictionary we also included records of word form frequencies from the corpus Czech Web 2017 (Suchomel, 2018) and we also utilize these for ordering of correction suggestions in the correction prediction feature. Considering taggers, the OCPT either uses Majka dictionary (Šmerk, 2014) or MorfFlex dictionary (Hajič, 2020). The choice of tagger is based on its success rate for different modules.

[3] An example question: "Why does my spell-checker mark the word místostarostka (deputy mayoress) as incorrect?"

[4] When the article was being written, corpus SYN v9 was not published yet.

Another valuable source for development of the OCPT is a list of neologisms provided by Czech Language Institute. This source will update the dictionary, as the list contains words which have not yet been captured in available sources (e.g. words related to the coronavirus epidemic).

2.2. Technical parameters

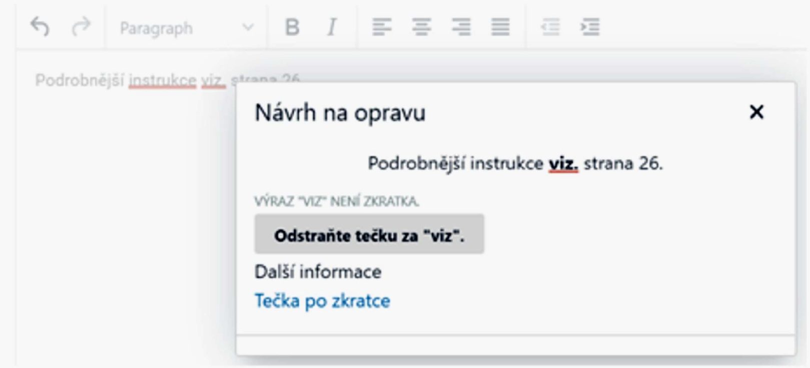
Upon input a text is processed using several tools. Tokenization is done by Unitok (Michelfeit, 2014), the selected sentence segmenter is sentence_separator.pl (Šmerk, 2008). Spell-checking is performed by SymSpell (Garbe, 2020). Morphological analysis is carried out by Majka (Šmerk, 2014) or MorphoDiTa[5] (Straková, 2014) depending on the type of the module. Because selection of the tagger affects the quality of the results achieved for different modules, the taggers are selected to give the best possible results for a given module (Machura, 2019). Syntactic analysis is performed by SET (Kovář, 2011). Further processing of the inputted text involves the processing by the above modules: the typographical module (typographical corrections are described in form of regular expressions and some are performed automatically); the punctuation module (rules governing additions of missing commas and rules for removing incorrectly used commas); the linguistic agreement module (subject-verb agreement, attributes in agreement etc.); the module for ungrammatical sentence constructions (zeugma, attraction, contamination) and the module for phenomena outside the scope of the former four (solves e.g. certain errors in spelling of -mně-/-mě- clusters; excessive repetition of demonstrative pronouns etc.).

The first step upon input of a text is incorporation of automated typographical corrections. After that the text is processed by the tokenizer, then by the tool correcting typos, followed by the sentence segmenter. The next step is the automatic morphological analysis and

[5] We use the version from 2014, since the newest was not published at the time the article was being written..

eventually the text is processed by the individual modules in a collateral manner.

After the text is processed, the OCPT underlines potentially incorrect expressions. This is the moment when interaction between the user and the OCPT must take place. When a user clicks a highlighted word, an error notification appears.



Podrobnější instrukce viz. strana 26

Návrh na opravu

Podrobnější instrukce **viz.** strana 26.

VÝRAZ "VIZ" NENÍ ZKRATKA.

Odstraňte tečku za "viz".

Další informace

Tečka po zkratce

Fig. 1. Error notification

The notification carries information on why the phenomenon is incorrect (*Výraz "viz" není zkratka.* "Viz" is not an abbreviation.) and offers a suggestion how to correct the error (*Odstraňte tečku za "viz".* Remove full stop after "viz".).

Section *Další informace* (More information) contains a link to ILRB. Although Vernon states that "current integrated grammar checkers are not designed to teach grammar, but to assist the writer in the identification of potential problems" (Vernon, 2000, p. 335), the integrated links to ILRB (Internet Language Reference Book, 2008––2021) will offer an explanation of linguistic errors. Thus the OCPT allows users not only to relatively passively accept suggested changes but also to actively improve their knowledge. We anticipate that we will receive feedback in the planned public testing on whether the users find this approach beneficial.

3. Limitations

In some areas of language usage extralinguistic circumstances, context or meaning (not always easily distinguished from each other) must be applied and here the OCPT meets its limits. As Audy Masopustová (2021) observes, people project their deep knowledge of the world into their writing, whereas a computer depends on the given data. Hence, there are language areas, where mere application of formal rules is not sufficient, e.g., capitalization, comma in attributes, writing of compound adjectives, certain cases of subject-verb agreement or certain typographical phenomena.

3.1. Capitalization

We have been verifying preparation of rule-based principles for capitalization, but their application is limited. This could be improved by incorporating lists of proper nouns to the spell-checker dictionary. We have such lists at hand, but they further impair the speed of the spell-checker. Yet not even full lists of proper nouns could guarantee correct solutions. As Svobodová (2015, p. 7) declares, "it is not always easy to decide whether a particular name should be considered a proper noun or a common noun. In many cases a proper noun cannot be identified based on its form, because common nouns can also exhibit the function of a proper noun (*Oko* (Eye) – name of a cinema)." Hence, capitalization may require extralinguistic factors to be taken into account and these, naturally, remain undisclosed to the OCPT.

Despite these limitations the OCPT will be equipped with several existing rules for correction of small groups of toponyms (Vostřelová, 2019) and the spell-checker dictionary will be extended with lists of the most frequent proper names obtained from corpus Czech Web 2017 (Suchomel, 2018).

3.2. Attributes

If a multicomponent attribute is situated in front of the modified noun, punctuation depends on the relationship between its components. When the components' relationship is coordinate, the attribute is a so-called multiple attribute and its components are divided by commas. If one component modifies the following one, it is a so-called gradually modifying attribute and there are no commas. In some cases the use or omission of a comma can even change the meaning: "*druhý mezinárodní filmový festival* (the second international festival = two festivals took place, both international) × *druhý, mezinárodní filmový festival* (the second, international festival = the second festival was international whereas the first was not)" (Pravdová and Svobodová, 2019, p. 148).

An attribute that is situated after its head noun can either be a loose attribute or a close attribute. The difference between the two is that a close attribute "expresses a substantial fact which, in the given situation and context, is essential for unambiguous understanding of the sentence" (Pravdová and Svobodová, 2019, p. 146) and is, therefore, not surrounded by commas, whereas a loose attribute "does not narrow the denotation of the noun which it modifies" (Pravdová and Svobodová, 2019, p. 146) and is, therefore, surrounded by commas. Like with multiple attributes and gradually modifying attributes, also here the use or omission of a comma may indicate a different meaning.

In such cases, then, extralinguistic circumstances and the meaning the authors want to project into their statement play a fundamental role. The OCPT, however, cannot incorporate these aspects.

3.3. Compound adjectives

Another intricate area of orthography is spelling of compound adjectives. The OCPT manages to correct many of them thanks to its large dictionary. What it cannot do, however, is to assess which of the variants is appropriate to use in a specific context. Considering automated corrections, two types of compound adjectives are difficult: compound adjectives designating colours and cases where the use or omission of a hyphen is essential for distinguishing the meaning.

For the first type a colour shade is written as a closed compound (e.g. *žlutozelený nálev čaje* a yellowish green tea infusion), whereas

for two independent colours the compound is hyphenated[6] (*žluto--zelené kostkované šaty* yellow and green checkered dress)(Pravdová and Svobodová, 2019, p. 130–131).

For the second type of compound adjectives the hyphen reflects whether the relationship between the components is coordinate (*politicko-ekonomický* political and economical), or subordinate (*politickoekonomický* politico-economic) (Pravdová and Svobodová, 2019, p. 130).

In these cases the OCPT cannot provide a solution, as it depends to an extent on intended meaning.

3.4. Certain cases of subject-verb agreement

Corrections of the subject-verb agreement are difficult for the OCPT because of semantics. To be able to suggest a fitting solution, the OCPT must first be able to identify the subject. It can recognise a compound subject, e.g. *Inženýr, doktor a policista poseděli v hospůdce* (An engineer, a doctor and a policeman were sitting in a pub), but it is not capable of distinguishing it from a complex subject, e.g. *Tento prozaik, dramatik a publicista napsal několik zásadních děl* (This novelist, playwright and journalist wrote (*sg*.) several fundamental works). For the latter case the OCPT will suggest an inadequate correction to a plural form of the predicate (*napsali* they wrote). An incorrect identification of the subject, or failure to recognize an implicit subject, then leads to a miscorrection like this: *Petr s Karlem se ztratil v lese a hodiny hledali cestu domů* (Petr with Karel got lost (*sg.*) in the wood and were searching their way home for hours), where the OCPT suggests to correct *hledali* (were searching), since it considers *hodiny* (hours) to be the subject. Yet, there are cases where the agreement based on meaning does not constitute a problem. If a double agreement is possible (i.e. agreement based on the form and agreement based on the meaning), the OCPT will adopt both appropriate solu-

[6] An exception being *černobílý* (black and white), written traditionally as a closed compound.

tions, e.g.: *“Zástupy zaměstnanců demonstrovaly před budovou vlády, aby vyjádřily* (i *vyjádřili*) *svůj nesouhlas.”* (Pravdová and Svobodová, 2019, p. 497) (Crowds of employees were demonstrating in front of the government building so that they would express (*inanimate/animate*) their disagreement).

3.5. Selected typographical phenomena

Typographical errors are usually easy to remove, but still, there are cases which the OCPT will not manage to resolve in a satisfactory manner. It will, e.g., fail to distinguish a decimal number (*1,2 km* 1.2 km) from a list of numbers (*Prodávají se v barevném označení 62, 38 a 25.* They are sold in colour codes of 62, 38 and 25.). It may struggle to distinguish a score (*Fotbalový zápas skončil výhrou domácích 6:2.* The football match ended in a 6–2 home win.) from a ratio (*Výsledek 50 : 52 hlasů byl těsný.* The result of 50 to 52 votes was a narrow one.). With a percentage sign the OCPT cannot be trusted to differentiate between a numeral value, *sleva 11 %* (a discount of 11%) and an adjective, *11% (jedenáctiprocentní) sleva* (an 11% discount). The OCPT is also incapable of assessing some abbreviations correctly, which is especially a problem with initialisms, since their number is countless. Also, most emoticons consisting usually of graphic symbols available in a computer keyboard will probably be evaluated as an error.

4. Pitfalls

The greatest difficulties we are facing now in the development of the OCPT are false positives. What needs to be considered for the future is the maintenance of the employed linguistic data (dictionaries, lists etc.) and their updates.

4.1. False positives and missed errors

Another aspect hindering comfortable use of the OCPT is that the OCPT is giving error notifications for expressions and phenomena

that are correct. When the OCPT is given a flawless text to check, it should definitely not return any error notifications. Incorrect error notifications devaluate credibility of the tool. The same happens when a user notices an error in the text but OCPT fails to correct it. While some users would prefer to be notified of undisputable errors only, meaning OCPT would refrain from notifying possible or questionable mistakes and would not bother the user with potential error notifications, other users (such as professional users of the language, editors, proofreaders etc.) would instead appreciate if the checking and error notification would be as comprehensive as possible for all (hence also potential) errors, since such users would be able to assess themselves, whether an error notification is justified and the error needs correcting. Of course, a 100% success rate in error detection and no error notifications for texts without errors would be ideal. To achieve such a state of development still requires a long way to go.

5. Conclusion

The goal of this article was to present the achievements, limitations and pitfalls of the Online Czech Proofreader Tool which has been being developed since 2019. Unlike other spell-checkers, the OCPT shall correct not only typos, but also orthographical, grammatical and typographical errors. Furthermore, thanks to its connection with ILRB, it shall also provide explanations for usage of various language means.

In the Achievements section we described tools employed in the rule-based OCPT and we characterised the linguistic data used. We described the correction process and, using an incorrect sentence as an example, we presented the appearance of the error notification and its linking to ILRB, which offers a detailed explanation of the phenomenon in question. We also declared that interaction between the user and the OCPT is necessary.

We acknowledged some limitations of the automated correction involving mainly those areas of a language, where extralinguistic circumstances, context or meaning play a role. Here, we mentioned capitalization, usage of commas in attributes, hyphenation of compound adjectives, agreement based on meaning and some typographical errors.

What we consider a problem now, is the false-positives (underlining a word that is correct from the language point of view).

To summarize, proofreading of texts is a demanding task in natural language processing, which is why some language problems are difficult to deal with. Despite all mentioned problems and limitations, we attained many achievements in the development of the Tool. The OCPT offers complex corrections including spelling, grammar, typography and even partly improves style of the text. For that reason we believe that the OCPT will become a useful tool for wide range of users.

References

Audy Masopustová, Markéta et al. (2021). Lingvista versus stroj: Rozdíl ve zpracování jazykových rovin – úskalí, možnosti a meze. In *Wyraz i zdanie w językach słowiańskich*, Wrocław.

Language Enquiry Database („Databáze jazykových dotazů"). (2016–2022). Praha: ÚJČ AV ČR. Accessible at: https://dotazy.ujc.cas.cz.

Garbe, Wolf. (2020). SymSpell, version 6.7. Available at: https://github.com/wolfgarbe/symspell.

Hajič, Jan et al. (2020). MorfFlex CZ 2.0. Data/software, LINDAT-CLARIAH. Available at: http://hdl.handle.net/11234/1-3186.

Internet Language Reference Book. (2008–2022). Praha: ÚJČ AV ČR. Accessible at: https://prirucka.ujc.cas.cz/.

Kovář, Vojtěch. et al. (2011). Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pp. 161–171. Berlin/Heidelberg: Springer.

Machura, Jakub et al. (2019). Comparing majka and MorphoDiTa for Automatic Grammar Checking. In *Thirteen Workshop on Recent Advances in Slavonic Natural Language Processing*, RASLAN 2019. Brno: Tribun EU, pp. 3–14.

Michelfeit, Jan et al. (2014). Text Tokenisation Using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 71–75. Brno: Tribun EU.

Opravy pravopisu a gramatiku v Dokumentech Google. (2021). Google. Available at: https://support.google.com/docs/answer/57859?co=GENIE.Platform%3DAndroid&hl=cs#zippy=.

Petkevič, Vladimír. (2014). Kontrola české gramatiky (český grammar checker). *Studies in Applied Linguistics*, 2014(2), pp. 48–86.

Pravdová, Markéta, Svobodová, Ivana (eds.). (2019). *Akademická příručka českého jazyka*. 2nd edition. Praha: Academia, 600 p.

Straková, Jana et al. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Suchomel, Vít. (2018). csTenTen17, a Recent Czech Web Corpus. In *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*, pp. 111–123. Brno: Tribun EU.

Svobodová, Ivana et al. (2015). Psaní velkých písmen v češtině. Praha: Academia, 350 p.

Svobodová, Ivana. (2019). Věrohodnost elektronických zdrojů jazykových dat. *Český jazyk a literatura*, 2018–2019, 69(5), pp. 249–251.

Šmerk, Pavel. (2008). *K morfologické desambiguaci češtiny*. Rigorózní práce. Masarykova univerzita, Fakulta informatiky. Brno.

Šmerk, Pavel. (2014). Tools for Fast Morphological Analysis Based on Finite State Automata. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 147–150. Brno: Tribun EU.

Vernon, Alex. (2000). Computerized Grammar Checkers 2000: Capabilities, Limitations, and Pedagogical Possibilities. *Computers and Composition*, 17, pages 329–349.

Vostřelová, Klára. (2019). *Automatická detekce chyb v psaní velkých písmen v češtině*. Brno: Masarykova univerzita.