Using Syntax Analyser SET as a Grammar Checker for Czech

Marie Novotná, Markéta Masopustová

Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno, Czech Republic
{428801,415295}@mail.muni.cz

Abstract. Checking the grammaticality of written text is one of the essential and highly desirable tasks of natural language processing. One of the very common mistakes in Czech texts are errors in agreement or using colloquial expressions in written texts. Based on the analysis, we created new rules for the grammar of the SET syntax analyser to use it not only as an analyser but also as a grammar checker. Then we tested their functionality. The side effect of the work was also the identification of possible complications, deficiencies of the tools and partly also suggestions for their solution.

Keywords: syntactic analysis; SET; spell checker; grammar checker; grammatical agreement; subject-predicate agreement; compound subjects; attributive adjective-noun agreement; colloquial expressions

1 Introduction

One of the basic tasks of natural language processing is checking the grammaticality of texts. Grammar checkers check the formal and grammatical accuracy of text written in a natural language based on the rules of the language in question. Since the complexity of spelling, grammatical and stylistic features varies in different languages, the level of grammar checkers differs. While the spell check is already relatively well solved, the problem is more complicated at other levels of language.

For our work, we have chosen areas of language in which users of language often fail. One of them is grammatical agreement, which is often written ungrammatically, since it is not noticeable in standard spoken Czech (i.e. the difference in writing *i/y*), or it uses different endings in colloquial form, which are informal for written texts. Since the subject-predicate agreement has already been partly solved earlier (see Chapter 2), we focused only on sentences in which the subject was multiple (consisting of two names), and the attributive adjective-noun agreement when attribute stands before the name. We also worked on selected common mistakes in the area of word formation, stylistics and syntax. These are often found in the commonly spoken language, however, in a written language, they are considered as faulty constructions. Based on this analysis, we created rules for the grammar of the syntax analyser SET [1], which makes it possible to identify the mistakes in the texts.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 9–14, 2018.

2 Related Work

For the Czech language, there are several different grammar checkers. The problem of spelling is solved well, whether in stand-alone programs or web applications or as part of text editors (e.g. Microsoft Office). In case of grammar checkers, only two commercial products are known, namely Grammaticon from Lingea [2] and Kontrola české gramatiky, developed at Czech Language Institute of the Czech Academy of Sciences as part of the Microsoft Word editor [3]. There is not much information about these grammar checkers, mostly only advertising posts. Nevertheless, Grammaticon is no longer supported today (the support ended in 2014), and Kontrola českého gramatiky has, according to the author's words, limited functionality since its launch. The only known stylistic checker for Czech was part of Grammaticon and had only minimal functionality [4].

As mentioned in the previous chapter, attempt to use the SET analyser as a grammar checker already exists [5]. There was a simple rule for detecting an error in a subject-predicate agreement. This subject was referred to as subject-bad. The rule was added to the existing one, and then the new grammar was tested on 26 sentences with 11 mistakes in subject-predicate agreement, which came from the tests of pupils of the first grade of elementary schools and were manually identified and classified [6]. This rule, however, has been able to work only with a simple subject, so in our work, we have continued with the rules for a multiple subject. Also, we extended the coverage of the grammar checker to include other mistakes, which we will introduce in Chapters 4-6.

3 The SET parser

The Syntactic Engineering Tool, introduced in [1], is based on partial segmentation of sentence. It works with implemented grammar, which is made out of rules for searching for the relations between tokens, or sentence members. These rules are then compared with the input sentence, and all relevant records are counted, their weight is evaluated, and the "heaviest" rules are applied. The result of such an analysis is the sentence with the labelled parts of speech and the relationships between them. Examples of rules are given in [1].

Our task was to create rules to correctly identify the part of speech in which the error lies and to mark it appropriately.

4 Attributive adjective-noun agreement

In the case of the attributive adjective-noun agreement, we limited the experiments to a simple adjective attribute standing before the noun. A new label has been introduced to indicate the wrong attribute *modifier-bad*. Primarily, we have searched for attributes that are widely used in spoken or informal language formations but do not belong to the written text.

In some cases, finding the error attribute was a simple task, because the information about the colloquial expressions was already mentioned in the

morphological tag on which the syntactic analysis is based (e.g. *malýmu klukovi*). In many cases, it was the format of an adjective that is formal for another case than that it was used (e.g. *o obědový pauze*). These mistakes were due to a case divergence. A common mistake in Czech is also in the use of the dual ending of the attribute if it is plural (non-zero) number (e.g. *barevnýma pastelkama*). However, due to the current limitations of the tools, we have not been able to solve this problem. In total, four new rules have been created to detect an error in the attributive adjective-noun agreement.

Following example shows one of the rules. This rule marks adjective as modifier-bad if the adjective was marked as colloquial in the morphologic analysis.

```
TMPL:   $MALYHOMU $...* noun MARK 0DEP 2 PROB 6001
        LABEL modifier-bad

$MALYHOMU(tag): k2.*wH
```

The rules were tested with 230 sentences with the attributive adjective-noun agreement selected from the Skript2012 [7] corpora, either in the correct or wrong form, approximately in a ratio of 1:1. It was correctly marked as *modifier-bad* 92 of the 136 wrong sentences, and a false positive appeared only once. The results are shown in Table 1, further comments can be found in [8].

In an analysis of the results, we found that a relatively large part of the attributes in which the error was not revealed was pronouns that we did not focus on (e.g. type *v kterým příkladu*), so the actual coverage inadequate cases could have been more successful.

5 Multiple subject-predicate agreements

Since the subject-predicate agreement has been dealt with earlier [5], we have focused on the multiple subjects, regarding the limitations of the tool used to the subject consisting of two nouns. SET allows to create coordination, but failed to allocate its morphological tag. In the syntactic analysis, it usually found a coordination rule, and each component of the multiple subjects was joined to it, but it evaluated the subject-predicate agreement for each name separately and subsequently assigned the coordination a label which had the highest weight according to the rules. The simplest solution to this problem would be to implement the SET rules for assigning a morphological tag to all coordination

Table 1: Results of attributive adjective-noun agreement.

TP	FN	FP	recall	precision
92	44	1	0,68	0,99

(for example, if there is at least one member of the coordination family, consider coordination as the gender of male animated).

Because SET does not allow coordination for the above mentioned morphological tag assignment, we had to deal with the problem differently. We have created entirely new rules for creating coordination directly linked to the predicate. New rules increased the number of different combinations, so it was necessary to create a relatively large number of rules for different gender and numbers of subjects and predicate. Given the complexity of this problem, the rules were set relatively "roughly". In the future, however, we expect the analyser to be adjusted, and when the coordination can be managed more efficiently, so we expect to change our rules in order to work more reliably. As a result of this part of the work, there were 24 rules for identifying the error in subject-predicate agreement and 35 rules to indicate the correctness of such compliance.

Following example shows one of the rules which are made for detecting multiple subject-predicate agreements. In this agreement on the first position is noun masculine animated and on the second position is any noun in nominative followed by predicate with ending *-i* (masculine animated in plural). If this pattern is found, multiple subject-predicate is marked as <*cood-s*> during syntactic analysis.

```
TMPL:   $SUBJ_M $...* $AND $...* $SUBJ $...* $PREDi MARK 0 2 4
        HEAD 2 DEP 6 PROB 10000

$PREDi(tag): k5.*gM.*nP
$SUBJ(tag): k1.*c1 $SUBJ_M(tag): k1.*gM.*c1
$AND(word): a i nebo ani $...*(tag not): k5 k8
```

We tested these rules on selected sentences of the Skript2012 [7] and CzeSL-SGT [9] corpora, which contained multiple subjects. Of the 39 errors, 24 were revealed, a false positive appeared 18 times in 131 correct sentences. The results are shown in Table 2, further comments can be found in [8].

When we tried analysis on the random sentences from the Internet, the results were even worse, but the exact numbers are unknown. The grammar checker's unreliability in this regard was mainly due to erroneous morphological disambiguation, but also to other factors (for example, we have failed to limit the rules to verbs in past tense that work on most of the mistakes in agreement). We continue to work on these issues to make the rules applicable in practice.

Table 2: Results of multiple subject-predicate agreements.

TP	FN	FP	recall	precision
24	15	18	0,62	0,57

6 Colloquial expressions in the written texts

Within the next module, we focused on the occurrence of spoken language elements in the written texts, because colloquial expressions are formal only in its spoken version. The first problem was the word order, namely the order of enclitics and prepositions. In Czech, the sentences are beginning with the enclitics, e.g. **Si pořídím nové kolo.*, rated as informal, whereas the sentences with two prepositions behind each other are considered as confusing for the participant of the speech, because the bond is broken, e.g. *Dospěl k pro něj těžké otázce*. We also tried to solve the excess of the demonstrative pronouns in the sentence and repeating the same expressions within one sentence. Also pleonasm, i.e. redundant repetition of the same (for example, *dárek zadarmo*), absurd superlatives (e.g. *nejzákladnější*), wrong use of the word *jakýkoli*, forgetting a double conjunctions (e.g. *buď-(a)nebo*), wrong usage of pronoun *který/jenž* and finally, the occurrence of words or forms of words rated as spoken lexicons in written form. The last subcategory were so-called other mistakes where we included mistakes such as the bad writing of words *výjimka* and *permanentka*, addressing another case than the fifth, incorrect form of the word *datum*, hypercorrection in the nominative of life masculine pattern muž (e.g. **reprezentanté*) and misuse of conjunction *mimo*.

The result is an extensive set of rules. For our evaluation, we built corpus of 1200 sentences without error and 400 sentences with an error. Generally speaking, the simple rules have had great success, and the more complex rules were worse, which is the result that we expected. The results are shown in Table 3 and discussed in more detail in [10].

Table 3: Results of colloquial expressions in the written texts.

rule	TP	FN	FP	precision	recall
enclitics correct sentences	136	0	0	1	1
enclitics bad senteces	41	0	309	0,117	1
prepositions	35	3	5	0,875	0,921
demonstrative nouns	59	1	1	0,983	0,983
repetition of words	46	0	0	1	1
pleonasms	131	17	0	1	0,885
superlatives	11	0	0	1	1
pronoun *jakýkoli*	20	0	0	1	1
double conjunctions	54	7	19	0,740	0,885
gender *který* correct sentences	151	0	1	0,993	1
gender *který* bad sentences	37	3	133	0,218	0,925
colloquial expressions	16	1	8	0,667	0,941
other	117	0	0	1	1

7 Conclusion

Our project aimed to create new rules for the SET grammar checker, which extends its functionality not only as a syntactic analyser but also as a Czech grammar checker. We have tried to partially cover the area of the attributive adjective-noun agreement, multiple subject-predicate agreements and other selected language issues. We are aware that the range of errors covered by our work has been considerably limited and that it needs to be expanded even more for the needs of the grammar checker. Similarly, it is possible to work on improving the precision and recall.

Creating a grammar checker for such a grammatically demanding language as the Czech language is not an easy task. However, we are convinced that if enough attention is paid to the problem and the tools are continually improved; we can make it to the ideal result slowly and in small steps.

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; project no. MUNI/A/0862/2017).

References

1. Kovář, V., Horák, A., Jakubíček, M. In: Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. Springer (2011) 161–171
2. Behún, D.: Lingea grammaticon – přísný strážce jazyka českého
3. Petkevič, V. In: Kontrola české gramatiky: český grammar checker. Univerzita Karlova v Praze, Filozofická fakulta (2014) 48–86
4. Pala, K.: Stylistický korektor (2017)
5. Kovář, V. In: Partial Grammar Checking for Czech Using the SET Parser. Springer (2014) 308–314
6. Trifanová, B.: Analýza chyb v diktátech žáků po absolvování 1. stupně ZŠ [online] (2013 [cit. 2018-10-30])
7. Šebesta, K.: Skript2012: akviziční korpus psané češtiny – přepisy písemných prací žáků základních a středních škol v ČR (2013)
8. Novotná, M.: Automatická detekce chyb v gramatické shodě v češtině [online]. Master's thesis, Masaryk University, Faculty of Arts, Brno (2018 [cit. 2018-10-30])
9. Šebesta, K.: CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 (2014)
10. Masopustová, M.: Automatická analýza srozumitelnosti textu [online]. Master's thesis, Masaryk University, Faculty of Arts, Brno (2018 [cit. 2018-10-30])