Comparing majka and MorphoDiTa for Automatic Grammar Checking

Jakub Machura, Helena Geržová,
Markéta Masopustová, and Marie Valíčková

Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno, Czech Republic
{415795,400133,428801,415295}@mail.muni.cz

Abstract. Developing a grammar checker requires the most accurate morphological analysis. We have been using the majka analyzer and DESAMB tagger so far, but due to certain obstacles to disambiguation, we encountered many errors in morphological analysis. Nowadays, there are several tools that achieve comparable results. Therefore, it was beneficial to test the one which is well-kept and open-source – the MorphoDiTa system. For the detection of grammatical, stylistic and punctuation errors we use mainly special grammars built into the SET parser and this paper presents results based on outputs of both morphological analyzers.

Keywords: syntactic analysis, SET, grammar checker, punctuation, comma, homonymy, grammatical agreement, subject-predicate agreement, colloquial expressions, zeugma

1 Introduction

To write a text without any grammatical, spelling, or typographical mistake[1] is one of the main features of high-standard typed text. Nowadays, users of language more often create demand for having software which would reliably detect and correct various kinds of mistakes in texts.

In the Czech environment are known two commercial grammar checkers: 1. Grammar checker built into the Microsoft Office, developed by the Institute of the Czech language [14] and 2. Grammaticon checker made by the Lingea company [1]. From February 2019 the Masaryk University in collaboration with the Charles University, the Institute of the Czech language, and the Seznam company has started a new project of developing an automatic online language checker [5].

This paper is aimed at a description of some well-know as well as particular obstacles in the morphological analysis. The paper also contains an important result: comparison and evaluation of two systems for the morphological analysis.

The structure of the paper is in the following way: The next section superficially describes several tools used for automatic text analyses and

[1] In this paper we use terms error and mistake synonymously as equivalents of *chyba* in Czech.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 3–14, 2019. © Tribun EU 2019

thoroughly examines two examples of phenomena which cause us obstacles during the morphological analysis. Then follows the comparison and evaluation of two examined systems.

2 Some Components For Automatic Language Checking

2.1 The SET System

For the main purposes of the new project is used the SET parser developed by Kovář [10]. In order to detect more complicated grammatical mistakes automatically (e. g. the subject-predicate agreement, punctuation errors, . . .), any grammar checker should work with an output of the morphological analysis which means that for every single word in the sentence structure must be assigned lemma and morphological tag. Nowadays, two mainly used conceptions exist on how to represent grammar information of Czech words – attributive, and positional tag system. An advantage of the SET parser is not only an ability to work upon with the attributive as well as the positional tag sets, but the SET system (which was primarily designed as the syntactic parser [8]) contains functionalities which deal with partial grammar checking.

2.2 Tools Used For the Morphological Analysis and Their Obstacles

Up to now, we have used `unitok` tokenizer [11] and for the morphological analysis, the analyzer majka [17] and subsequent disambiguation is operated by the DESAMB [16] tagger. The analysis using both tools brings sufficient results, yet some inaccuracy has occurred. Of course, there are factors which prevent absolute accuracy for the automatic morphological analysis, such as homonymy of word forms, especially for flective languages, or grammar mistakes caused by users of language:

Ex. 1

S1: Muž, který *je*$_1$ pravděpodobně unesl, *je*$_2$ běloch.

(*The man, who probably kidnapped them*$_1$, *is*$_2$ *a white man.*)

The analyzer majka and DESAMB tagger give following output for S1 (1st column word, 2nd column lemma and 3rd column tag):

```
<s>
Muž                 muž                 k1gMnSc1
<g>
,                   ,                   kIx,
který               který               k3yRgMnSc1
je1                 on                  k3xPp3gNnSc4
pravděpodobně       pravděpodobně       k6eAd1
unesl               unést               k5eAaPmAgMnS
<g/>
,                   ,                   kIx,
je2                 být                 k5eAaImIp3nS
běloch              běloch              k1gMnSc1
<g/>
.                   .                   kIx.
</s>
```

The word je in S1 are tagged as a pronoun (k3xPp3gNnSc4), and as a form of the verb to be (k5eAaImIp3nS).

Nevertheless, many users of Czech, even natives speakers, often forget putting the second comma which set an inserted subordinate clause apart from a main clause from the right side:

S2: *Muž, který je$_1$ pravděpodobně unesl je$_2$ běloch.*

The analyzer majka and DESAMB tagger now provide this output for S2:

```
<s>
Muž                 muž                 k1gMnSc1
<g/>
,                   ,                   kIx,
který               který               k3yRgMnSc1
je1                 on                  k3xPp3gNnSc4
pravděpodobně       pravděpodobně       k6eAd1
unesl               unést               k5eAaPmAgMnS
<g/>
je2                 on                  k3xPp3gNnSc4
běloch              běloch              k1gMnSc1
<g/>
.                                       kIx.
</s>
```

The absence of comma from the right side of the subordinate clause caused the tagger to choose a wrong (but justifiable) tag for the *je*$_2$ in the S2.

Ex.2

To deal with the case homonymy within a paradigm of one noun could represent even much more difficult task for any tagger compared with word form homonymy. For instance, the word form *koření* is homonymous for six cases of the singular and four cases of the plural form of noun *koření* (spice) (See the Table 1), and at the same time could be a form of the verb *kořenit* (spice up) (See the Table 2). Additionally, the instrumental case of the singular form and the accusative case of the plural form are also homonymous (*kořením*). Homonymous forms in Tables 1 and 2 have highlighted background.

Table 1: Paradigm of the noun *koření* (spice).

Paradigm of *koření*		
	singular	plural
nominative	koření	koření
genitive	koření	koření
dative	koření	kořením
accusative	koření	koření
vocative	koření	koření
locative	koření	kořeních
instrumental	kořením	kořeními

Table 2: Conjugation of the verb *kořenit* (Spice Up) – Present Tense.

Conjugation of *kořenit*		
	singular	plural
1^{st} person	kořením	kořeníme
2^{nd} person	kořeníš	kořeníte
3^{rd} person	koření	koření

S3: V druhém šuplíku najdeš správné **koření**.
(*In the second drawer, you will find the right spice.*)

The analyzer majka and DESAMB tagger provide the output for S3:

```
<s>
V                   v                   k7c6
druhém              druhý               k4xOgInSc6
šuplíku             šuplík              k1gInSc6wH
najdeš              najít               k5eAaPmIp2nS
správné             správný             k2eAgNnSc1d1
koření              koření              k1gNnSc1
<g/>
.                   .                   kIx.
</s>
```

In the S3, the DESAMB tagger wrongly identified the case as the nominative for the noun *koření* (the verb *najít* requires an object in the accusative case though). Moreover, the finite verb *najdeš* according to the ending *-š* contains information that a subject is a 2^{nd} person of the singular and the subject *ty* (you) is visibly unexpressed. Therefore, no other constituent in the clause should have a form in the nominative case.

S4: Běž a rychle kup **koření** v supermarketu.
(*Run and buy quickly spice/spices in the supermarket.*)

The analyzer majka and DESAMB tagger provide the output for S4:

```
<s>
Běž                 běžet               k5eAaImRp2nS
a                   a                   k8xC
rychle              rychle              k6eAd1
kup                 kup                 k1gInSc1
koření              kořenit             k5eAaImIp3nS
v                   v                   k7c6
supermarketu        supermarket         k1gInSc6
<g/>
.                   .                   kIx.
</s>
```

In the S4, the noun *koření* was incorrectly tagged as a verb instead of a noun in the accusative case and the imperative form *kup* of the verb *koupit* got the tag as a noun in the nominative case.

We also noticed that the DESAMB tagger sometimes matches some adjectives and pronouns as nouns (See Ex. 3). Nevertheless, they stand in the position of premodifier followed by a noun. Thus, there could not be any syntactic reason to tag them as nouns.

Ex.3

Do své hospody vezmu jen slušné zákazníky, ne **žádné** [k1gMnPc4] vagabundy.
(*I will take only polite customers to my pub, not any vagrants.*)

Žádný [k1gMnSc1] kout světa není bezpečný.
(*There is no place in the world that is safe.*)

S **dalším** [k1gNnSc7] naším vlastníkem uzavřeli dohodu.
(*They made an agreement with our other proprietor.*)

Dokázal uhrát **stejnou** [k1gFnSc4] plichtu i s Francií.
(*It managed to play a tied score also with France.*)

2.3 Making Use of the MorphoDiTa system

Collaboration with linguists from the Charles University and inaccuracies described above led us to start thinking about a possibility to use the MorphoDiTa[2] – a complex tool for the morphological analysis which is used especially at The Institute of the Czech National Corpus and The Institute of Theoretical and Computational Linguistics of Faculty of Arts, Charles University. The open-source MorphoDita [18] is an acronym from the Morphological Dictionary and Tagger. It uses an accessible and updated morphological dictionary MorfFlexCZ [4] and it is composed of several modules (a tokenizer, Morphological Generation, Morphological Analysis and Morphological Tagger [15]). Additionally, the MorphoDiTa system achieves one of the best results in accuracy to assign a tag in comparison to other Czech systems [16].

2.4 Positional - attributive tags conversion

The MorphoDiTa system works with positional tag set. As was mentioned earlier, the SET parser also has functionality which allows processing morphological tags in positional format. The `--posttags` switch provides conversion of positional tags into the attributive format (See the process of conversion below).

The output of the morphological analysis provided by the MorphoDiTa system and follow-up conversion:

S1: *Muž, který je pravděpodobně unesl, je běloch.*

```
                                                    --posttags
Muž            muž            NNMS1-----A----      k1gMnSc1eA;cap
,              ,              Z:-------------      kI
který          který          P4YS1----------      k3yRgMgInSc1
je             on             PPXP4--3-------      k3xPg.nPc4p3
pravděpodobně  pravděpodobně  Dg-------1A----      k6d1eA
unesl          unést          VpYS---XR-AA---      k5mAgMgInSp.mReA
,              ,              Z:-------------      kI
je             být            VB-S---3P-AA---      k5mInSp3mIeA
běloch         běloch         NNMS1-----A----      k1gMnSc1eA
.              .              Z:-------------      kI
```

It should be noted that the conversion is not going on in the ration 1:1 which means that not every single tag from the positional tag set has a corresponding tag in the attributive tag set.

[2] The open-source version is available on `https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-43CD-0`.

3 Partial Grammar Checking Using the MorhoDiTa

3.1 Punctuation Checking

Automatic punctuation checking (finding a place where a missing comma should be inserted, or removal of an incorrectly written comma) belongs to a much more complicated grammar task. The SET system has functionality which reaches one of the best results in finding a place where a missing comma should be inserted [9]. Testing and comparison (majka + DESAMB versus MorphoDiTa) were made on the DESAM corpus [13] that contains 61 098 commas. For the purpose of the task to find a place where a missing comma should be inserted, every single comma was removed from the corpus and the SET system works with a plain text without any comma. Generally, the MorhoDiTa did not bring better results than the majka and the DESAMB tagger. We assume, though, the MorphoDiTa deals with case homonymy a bit better, but this assumption needs deeper research (See the Table 3).

Table 3: Results of the comparison – Punctuation checking. TP – True Positives (correctly found commas); FP – False Positives (incorrectly found commas); FN – False Negatives (missed commas), P – Precision; R – Recall. 1. Rules which deal with commas after the connector; 2. Rules for multiple sentence members mostly based on case agreement; 3. Rules for multiple sentence members mostly based on case agreement + information about collocation from the corpus csTenTen17 [19].If we consider whole rules as complex determining the insertion of commas, the majka and the DESAMB tagger win both in precision and recall. However, it is worth noticing that the MorphoDiTa has better precision in the detection of groups of nouns in coordinating relation (which share a case form), but with a lower value of recall.

Total of commas: 61 098	**majka + DESAMB**					**MorhoDiTa**				
Rules	**TP**	**FP**	**FN**	**P (%)**	**R (%)**	**TP**	**FP**	**FN**	**P (%)**	**R (%)**
All rules	33 833	2 457	27 265	**93,23**	**55,37**	33 808	2 741	27 290	92,50	55,33
1. Connector	32 806	2 256	28 292	**93,57**	53,69	32 805	2 609	28 293	92,63	53,69
2. Coordination	1 025	224	60 073	82,07	**1,68**	1 005	145	60 093	**87,39**	1,64
3. Coordination	1 034	94	60 064	91,67	**1,69**	804	56	60 294	**93,49**	1,32

3.2 Automatic detection of zeugma

The diploma thesis of Geržová [3] deals with automatic detection of zeugma. In Czech, this term means that one expression is in semantic or syntactic relation with two other paratactically connected expressions (e.g. two verbs), but the whole structure is grammatically defective [7] as in the example below.

Ex.:
Potvrzujete a souhlasíte s tím, že žádný software není bez vad.
(*We confirm and agree with an idea that no software is without any fault.*)

In the thesis [3], Geržová focused mainly on verbal coordinations. For this purpose we created eighty-three rules based on the assumption that the first verb with the obligatory subject in coordination with the second verb does not have a suitable addition in the sentence. Together the rules had precision 63,36 % and recall 54,78 % [3].

However, many of the false positives were caused by inaccurate disambiguation, especially if there was ambiguity in-between cases (as in the example below).

Ex.:
Řekl bych, že **věc** `/k1gFnSc4/` **chápe a rozumí grafům**.
(*I would say he gets the point and understands graphs.*)

Therefore we supposed that a more accurate morphological analyzer could increase precision and recall. The precision of the rules in the above-cited work was measured on the first 100 million lines of the corpus csTenTen17 [19]. Recall was obtained from another file ("test_set_with_errors_2") with errors of this type.

To compare majka (DESAMB) and MorphoDiTa, we chose twenty verbs and adequate rules for the detection of zeugma. Against the diploma thesis [3], we tested precision for this time on a file ("test_set_mixed_1") that contained one thousand sentences for each tested verb (rule). These sentences included coordinating structures consisting of a tested verb and another verb. With this method we obtained more defect structures of this type, because as we learned in theses [3] zeugma is not a very frequent phenomenon. The results are included in Table 4.

Table 4: TP test_set_mixed_1 – true positives found in file "test_set_mixed_1"; FP test_set_mixed_1 – false positives incorrectly marked as zeugma in file "test_set_mixed_1"; Precision test_set_mixed_1 – based on results from file "test_set_mixed_1"; TP test_set_with_errors_2 – true positives found in file "test_set_with_errors_2"; FN + TP – number of all zeugmas in the tested file "test_set_with_errors_2".

	test_set_mixed_1			test_set_with_errors_2		
	TP	FP	Precision (%)	TP	FN + TP	Recall (%)
majka + DESAMB	314	57	84,64	227	483	47,00
MorphoDiTa	359	50	87,78	225	483	46,58

According to the results of the analysis, values of precision and recall were similar for both tested analyzers. The rules using the MorphoDiTa morphological analyzer had a few percent higher precision and a half percent worse recall.

3.3 Automatic detection of errors in subject-predicate agreement

Another part of comparison and evaluation was made on sentences that contain errors of subject-predicate agreement. Results of the majka analyzer on this kind of data were discussed in [12].

During the testing, we looked at the subject-predicate agreement with a simple subject that was expressed within a given clause. We realised the complexity of the task, therefore, we decided from the testing to exclude examples where a subject is expressed elsewhere within the sentence or is not expressed at all. At the same time, we removed phrases contained errors that are not covered by the existing rules.

We used a file with 124 sentences, of which 34 were correct and 90 contained one or more errors of subject-predicate agreement. The table of results (majka, MorphoDiTa) is attached to the end of this subchapter (Table 5). The first part of the testing revealed that majka + DESAMB behave cautiously – rather avoid to make a false report about the incorrect subject-predicate agreement, but at the same time, lots of real mistakes are ignored. On the other hand, the MorphoDiTa reports more mistakes, even though lots of them are evaluated as false reports.

In the case of majka and DESAMB, the majority of the false reports are caused by the wrong disambiguation – the DESAMB tagger often identifies a subject as a noun in accusative form and then the SET parser assesses the noun as an object [5].

Table 5: Comparison of majka and MorphoDiTa on the identification of subject-predicate agreement. TP – revealed error when SET correctly labeled a subject as predictive on predicate and labeled it "subject-bad"; FP – so called fake mistake where a "subject-bad" tag was wrongly labbeled to another member in the clause. FN – missed errors in subject-predicate agreement.

	TP	FP	FN	Precision (%)	Recall (%)
majka	29	15	65	65,9	30,9
MorphoDiTa	40	48	54	45,5	42,6
MorphoDiTa (after repair)	40	12	54	76,9	42,6

A deeper inspection of MorphoDiTa's results revealed that the positional-attributive conversion does not provide complete results. The attributive system uses unambiguous tags for verbs which prevent the homonymous understanding – the verb form *pomohly* (they helped) with the tag `k5eAaPmAgFnP` clearly refers to a plural feminine subject (the attribute `gF`). However, the positional system allows using of ambiguous tags – the verb form *pomohly* has the tag `VpTP---XR-AA---` where the third position T applies to the feminine as well as the masculine inanimate gender. On that account, the `--posttags` switch gives the tag `k5mAnPgIgFnPp.mReA` to the output which the SET is not able to process (double attribute `gIgF`) and during the analysis, SET works only with the first attribute `gI`. At the end, the SET announces a false report.

If SET could work with all possible interpretations gained from a tag, we suppose whole analysis will get better results. With this supposition, we manually fixed the way how the `--posttags` switch evaluates tag matching. After that, SET takes into account more possible interpretations if they are recorded in a tag. Subsequent testing reduced the number of false reports (FP) from 48 to 12 which rapidly increased the accuracy (See the table 5). No other results were affected by this adjustment.

3.4 Colloquial expressions in written texts

The last part of testing was focused on stylistic. Details about this module were presented in [5] and [12]. With regard to the type of the rules, there were not as many problems as in other modules because these rules are not so dependent on morphological analysis.

The biggest issue was colloquial expressions in written text - e.g. *hezkej nábytek* (a nice furniture), where MorphoDiTa had no match. It is caused by the `--posttags` switch: in majka atributive system there is a part of the tag containing `wH`, which means conversational/colloquial [6]. However, the `--posttags` switch does not convert this part of the tag to MorphoDiTa's fifteenth position, which holds stylistic variant [2].

Other rules have more or less same results as presented in [5] and [12], so we do not consider them important to mention.

4 Conclusion

Testing did not prove that the MorphoDiTa system would arrange a big difference in results. MorphoDiTa mildly wins in automatic detection of zeugma and detects errors in subject-predicate agreement with better accuracy. However, the majka analyzer with the DESAMB tagger provides better precision and recall in general evaluation of the automatic insertion of missing commas. The detection of multiple sentence members and the detection of errors in subject-predicate agreement indicate that MorphoDiTa deals with case homonymy better than DESAMB tagger.

The using of the MorphoDiTa system could be advantageous since the system works with the maintained dictionary and is updated on a regular basis. Therefore, it would be practical to tune up the `--posttags` switch which will be able to convert ambiguous positional tags. Nevertheless, we also see the room for improvement of tools that we have used up to know and which would lead to satisfying outcome: 1. to implement linguistic rules that would improve disambiguation of the DESAMB tagger or to develop/find better tagger; 2. to update the dictionary which is used by the majka analyzer.

In conclusion, we would like to mention a paradox that partly affects our work: An excellent automatic grammar checker needs the best possible output of the morphological analysis. But in case an analyzer should provide the best analysis, it requires a text with a minimum of mistakes.

Acknowledgements This work was 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/1061/2018) and by the Technology Agency of the Czech Republic under the project TL02000146.

References

1. Behún, D.: Lingea Grammaticon – přísný strážce jazyka českého, `https://www.interval.cz/clanky/lingea-grammaticon-prisny-strazce-jazyka-ceskeho/`
2. Cvrček, V., Richterová, O.: seznamy:tagy – příručka ČNK (2017), `http://wiki.korpus.cz/doku.php?id=seznamy:tagy&rev=1497540816`
3. Geržová, H.: Automatická detekce negramatických větných konstrukcí pro češtinu. Master's thesis, Masaryk University, Faculty of Arts, Brno (2019 [cit 2019-10-30]), Available from: `https://is.muni.cz/th/fuz2y/`
4. Hajič, J., Hlaváčová, J.: MorfFlex CZ (2013), `http://hdl.handle.net/11858/00-097C-0000-0015-A780-9`, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
5. Hlaváčková, D., Hrabalová, B., Machura, J., Masopustová, M., Mrkývka, V., Valíčková, M., Žižková, H.: New online proofreader for czech. In: Slavonic Natural Language Processing in the 21st Century. p. 56–69. Tribun, Brno (in printing)
6. Jakubíček, M., Kovář, V., Šmerk, P.: Czech morphological tagset revisited. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011 pp. 29–43 (2011)
7. Karlík, P.: Zeugma. In: CzechEncy - Nový encyklopedický slovník češtiny (2017), `https://www.czechency.org/slovnik/ZEUGMA`
8. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for czech. In: Language and Technology Conference. pp. 161–171. Springer (2011)
9. Kovář, V., Machura, J., Zemková, K., Rott, M.: Evaluation and improvements in punctuation detection for czech. In: International Conference on Text, Speech, and Dialogue. pp. 287–294. Springer (2016)
10. Kovář, V.: Partial Grammar Checking for Czech Using the SET Parser, pp. 308–314. Springer (2014). https://doi.org/10.1007/978-3-10816-2_38
11. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: RASLAN. pp. 71–75 (2014)
12. Novotná, M., Masopustová, M.: Using syntax analyser set as a grammar checker for czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018. pp. 9–14 (2018)

13. Pala, K., Rychlý, P., Smrž, P.: Desam—annotated corpus for czech. In: International Conference on Current Trends in Theory and Practice of Computer Science. pp. 523–530. Springer (1997)
14. Petkevič, V.: Kontrola české gramatiky (český grammar checker). Studie z aplikované lingvistiky-Studies in Applied Linguistics **5**(2), 48–66 (2014)
15. Pořízka, P.: Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje. Vydavatelství Filozofické fakulty Univerzity Palackého v Olomouci (2014)
16. Šmerk, P.: Towards morphological disambiguation of Czech. Ph.D. thesis, Masaryk University, Faculty of Informatics, Brno (2007)
17. Šmerk, P.: Fast morphological analysis of czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007. pp. 13–16 (2009)
18. Straková, J., Straka, M., Hajič, J.: 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 (2014)
19. Suchomel, V.: cstenten17, a recent czech web corpus. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018. pp. 111–123 (2018)