Classification of Errors in Text

Jan Bušta, Dana Hlaváčková, Miloš Jakubíček, and Karel Pala

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xbusta,ydana,xjakub,pala}@fi.muni.cz
http://nlp.fi.muni.cz/

Abstract. This paper presents two classifications of errors in Czech texts. As a basic resource we use the corpus (Chyby – Errors) which has been continuously developed from 1999–2000 ([1]). The corpus text contains various kinds of errors such as spelling, typographical, grammatical, semantic, lexical, and stylistic ones. They have been corrected manually and annotated according to the classification of errors (annotation scheme) developed for this purpose. For the annotation we implemented a tool named WinCorr.
We mention the first annotation scheme and discuss the second one which has been designed recently to obtain more adequate description of the errors occurring in texts. We also discuss the principles on which both classifications are based.

Key words: errors in text; classification of errors

1 Introduction

In any text written by humans there always occur errors in spelling, grammar, semantics, style and typography. Not only this: *if humans correct errors in texts, they are not able to remove all of them*. That is why publishing houses and editorial boards have to employ readers and proof-readers whose task is to find errors in texts, correct them and finally produce printed texts of the best possible quality.

At present the prevailing majority of texts is produced on computers, which in turn are used for typesetting, storing and dissemination via Internet. No wonder that there is also a strong tendency to use computers to correct texts and remove errors from them. Programs called (spelling, grammar, style) *checkers* have appeared, and they allow us to correct some well recognised errors in texts. In some respects, they are more reliable than humans and are able to remove errors of some types completely.

The existing checkers are in some respects quite limited. Thus to be able to analyse all kinds of errors occurring in natural language texts, it is necessary to have a collection of texts (in our case in Czech) containing all kinds of errors. Therefore we decided to build a text corpus that contains various spelling, grammatical, style, semantic, typographical (and possibly other) errors and annotate them in the corpus text. The corpus with annotated errors is named **Chyby** ([1]).

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 109–??, 2009.

In this paper we briefly report on building the Czech corpus Chyby and on how errors have been marked and annotated with the help of the tools (programs) developed particularly for this purpose. There are two of them, the old one is WinCorr [2], the new one is OOCorr [3] (see below).

2 Why the Chyby Corpus?

At a first glance, it might seem that standard general corpora such as BNC [4] or the Czech National Corpus [5], could serve reasonably well for our purposes. After closer inspection of the texts from these resources it appears that the general corpora mostly contain texts that already have been proof-read and corrected (newspaper texts or fiction etc.). They still contain some errors as mentioned above, but their number is rather small since the worst ones have been removed. However, if we watch humans in the process of producing texts spontaneously we observe a different picture. The number of errors in such texts is quite high and some of them are quite severe.

Thus we turned to *spontaneous texts (s-texts)*. These texts were generated by students at FI MU, who take in their Bachelor studies a subject called *Elements of Style*. During the course they have to write two kinds of texts: an essay and an introduction to their Bachelors theses, each of them comprising approx. 600–700 words. The submitted texts have been corrected manually by four teachers and returned to the students who have to prepare final corrected versions of their texts and annotate the marked errors electronically using a program developed for this purpose (WinCorr, OOCorr, see below). The corrected and annotated texts have been used for creating the corpus Chyby by means of the corpus manager *Bonito/Manatee* [6]). At present, the size of the Chyby is approx. 500,000 word forms.

The nature of the texts delivered by the students is in accordance with our idea of s-texts: the number of errors and their types can be considered representative enough. In some cases, the texts are not well written and in our view they contain a large percentage of errors. In 650 words it is sometimes possible to find about 30 bad errors, though not all errors are regularly related to the individual word forms. For example, they involve changing word order, deleting and substituting whole lines or even paragraphs.

3 How to Classify Errors in Text?

The starting point for our classifications of errors in texts and the annotation scheme based on it are the Rules of Czech Orthography [7] and their electronic version [8], an official reference manual published by the Institute of Czech Language, Czech Academy of Sciences. It describes the basic principles of Czech orthography, which in comparison with English are much more phonetically oriented, although they are governed by a number of historical rules as well, especially in what concerns of inflection. The Rules also contain the punctuation rules, which reflect the syntactic segmentation of Czech sentences, e. g. main

and subordinate clauses are typically separated by commas on both sides and commas have to be placed before or after some conjunctions as well. In this respect Czech punctuation is somewhat complicated. This is the reason for a large percentage of the punctuation errors in the students' texts.

As a whole the Rules represent a reference manual based on the empirical rules, the majority of which can be characterised as deterministic (we estimate this amount at approx. 80 %). We have found it reasonable to start with the Rules and in combination with the data obtained from the Chyby, work out a more complete and formal description of the errors occurring in Czech texts together with their detailed categorisation. As far as we know, there is no general classification of errors that may occur in the texts. However, on the Web one can find reports and papers about grammar checkers and their development where overviews of the main types of errors can be found, see e. g. [9,10] or [11].

4 S-texts and Errors in Them

In agreement with rules of Czech orthography ([7]) the errors were originally classified in the following way:

- spelling,
- morphology,
- syntax (grammar),
- punctuation,
- lexical and semantic choice,
- style,
- typography.

This classification contains some subgroups and was used in the tool WinCorr [2]. We have been using it since 2002. During this time it served its purpose decently. However, there appeared various problems (e. g. it cannot handle the new ISO-standardized ODT document format which is used more and more extensively by our students). Thus we decided to revise the tool and develop a new and, hopefully, a more adequate one.

There are also several reasons for designing a new classification, though we are aware that it is possible to design an infinite number of them. We mention here the following points that we have been considering in the revision of the first error classification:

- the classification contains items that are overlapping. This, in our view, can hardly be avoided but it can be minimized. We are approaching it in the new classification.
- some of the spelling errors can be characterized as rather formal. These are mostly errors that can be discovered by a spelling checker.
- there are errors that on one hand can be characterized as spelling ones, but on the other hand also as grammatical (morphosyntactic) ones.
- a special group represents semantic and lexical errors. Their nature is not formal and can be discovered and corrected by humans only.

– the same can be said about stylistic errors though they grow from the language form,
– frequency considerations.

The new classification of errors as they occur in s-texts:

– spelling errors
  • obvious typing errors (recognizable by a spelling checker)
  • other typing errors (*i*/*y*, *s*/*z*, that cannot be recognized by a checker)
  • inflectional noun endings
  • syntactic (valencies, verbs–NPs, adjectives–NPs, agreement in NPs, NP–verbs)
  • capital letters, lowercases
  • compounds (mostly adverbial)
– punctuation (comma, colon, semicolon, dot, triple dot)
  • constituents (usually types of coordination)
  • clauses (relative clause, subject, object, adverbial, coordination)
– lexico-semantic errors
  • MWEs or sentences with broken meaningfulness
  • omitted or missing words
  • incorrect use of the possesives (*svůj*, *váš*,...)
  • incorrect choice of lexical items
– stylistic errors
  • incorrect register (colloquial, archaic, slang)
  • repeated expressions (demonstratives, adverbs, particles)
  • cumulation of the nouns ending with *-ní*
  • passive vs. reflexive passive
  • incorrect word order
  • clumsy expressions (MWEs, sentences)
  • too long sentences
– typographical errors
  • local errors: spaces (in acronyms), hyphens, inverted commas, brackets, one character consonant prepositions, incorrect characters
  • overall document layout: incorrect document structure, wrong paragraphization and hyphenation, orphans and widows, rags and rivers
  • incorrect choice of visualization means: inappropriate typeface, font properties or typesetting combination, low readability of text, wrong disposition of non-text items etc.

5 Annotation Scheme and Tags

In the previous section we indicated what kinds of errors we distinguish and want to annotate in the corpus Chyby. The next step is the design of the annotation scheme, which allows us to mark the errors and their types in the corpus text.

The original annotation scheme developed for the Chyby distinguishes the following types of errors:

- *Spelling*, errors which can be relatively well recognised in the texts and tools exist for their recognition (spelling checkers).
  Example: ***s****kouška* instead of correct ***z****kouška (examination)* or *standar**t**ní* instead of *standar**d**ní (standard)*.
  tag: `errtype=prav-pism`,
- *Typographical* errors consist in the incorrect use of various characters such as inverted commas, hyphens, placement of spaces, or single letter consonant prepositions at the ends of lines, etc.
  Example: *4 MB* instead of *4MB*,
  tag: `errtype=prav-mez`,
- *Morphological* and *syntactic* errors consist in using wrong endings in the inflected words (nouns, adjectives, pronouns, numerals, verbs and adverbs). There is, in fact, overlapping between those two types of errors, because the wrong ending (morphological error) causes an error in grammatical agreement on the syntactic level.
  Example: the incorrect ending in the noun group *dvě**mi** způsoby (in two ways)*. Similarly the agreement of subject and verb is violated in the cases like *ženy šl**i*** instead of *ženy šl**y** (women went)*.
  tag: `errtype=ms-nom`
- Clear *syntactic* errors consist in using incorrect verb valencies. Czech verbs in their valency frames strictly require concrete cases as complements, e. g. verb *zabít (to kill)* requires subject in nominative, object in accusative and if the instrument of killing is mentioned it has to be expressed by instrumental case.
  Example: in *Cizinec zabil chlapci nože. (The stranger killed boy knives)* the cases are used incorrectly. Only *Cizinec zabil chlapce nožem (The stranger killed the boy with knife)* is the correct use of the valency frame for *zabít (to kill)*.
  `tag: errtype=ms-val`
- *punctuation* errors follow from missing or incorrect placement of commas or other delimiters (!, ?, ;) in the sentences. In Czech, punctuation rules reflect the syntactic structure of the sentence, commas typically separate the main and subordinate clauses and are obligatory, especially with some conjunctions. The frequency of the punctuation errors in the Chyby is consequently quite high.
  Example: *Student ví že musí složit zkoušku. (The student knows that he has to pass the exam.)* The missing comma in front of *že* has to be inserted *Student ví**,** že musí složit zkoušku*.
  `tag: errtype=intp-pvety`
- *semantic (lexical)* errors include cases where expressions are incorrectly used, causing violation of semantic meaning fullness.
  Example: *rektor fakulty (Rector of the Faculty* – the correct expression is *děkan fakulty (Dean of the Faculty)*
  `tag: errtype=sem-slovo`
- *stylistic* errors represent a collection of the various violations such as inappropriate use of colloquial slang or jargon expressions, archaic or too informal words, repetitions of some expressions within a relatively short context (up

to five sentences). As stylistic errors we also classify the repetitions of some words (*také (also)*) in short contexts, superfluous use of demonstrative pronouns (determiners), abundant use of passive constructions, long chains of noun groups, especially the prepositional ones, and ambiguous uses of anaphoric pronouns, i.e. errors in co-reference. We have developed detailed subclassification of stylistic errors but here we show only two groups related to the substandard uses of some expressions.
Example 1: incorrect slang expression *spakovaný soubor* instead of *komprimovaný soubor (compressed file)*
`tag: errtype=styl-subst,`
Example 2: archaic form of the infinitive *nalézti* as opposed to the standard form *najít (to find)*
`tag: errtype=styl-nadst`

The final format is an XML application. The `<corr>` elements are used for error annotation.

6 Tools for Tagging Errors in the Texts

The tagging of errors is a tedious task which we have tried to make as simple as possible. Each student is responsible for his/her own document and his/her final course grade is partly based on the quality of the tagging of previous errors in the essay. It corresponds to the level of comprehension of each particular grammatical phenomenon.

Our first tool developed for this purpose (WinCorr) was implemented as a standalone text editor for the RTF document format. It has the advantage of being fully in control of users behaviour. On the other hand, it has a relatively poor functionality, it is not multiplatform and restricts the users in their choice of a text editor. Besides, there was also a set of Microsoft Word macros implementing similar functionality, which, however, had similar disadvantages and maintenance problems.

The new OOCorr application implements the functionality of a simple corrector as an extension in the environment of the OpenOffice.org Writer text editor system. This enables users to employ an arbitrary text editor for writing their texts provided that it is able to store the document in one of wide range of document formats supported by OpenOffice.org Writer (e. g. DOC, ODT, RTF, HTML etc.). Moreover, it benefits from the multiplatformity and rich functionality of the OpenOffice.org system which is being continuously developed and enhanced.

The function of this program is demonstrated by the screenshots in Figures 1, 2 and 3, where the error marking process is shown.

7 New Annotation Scheme/Error Classification

Since the time the original corpus Chyby has been built, the error classification was stable even if there were some problematic places. Decisions on how to

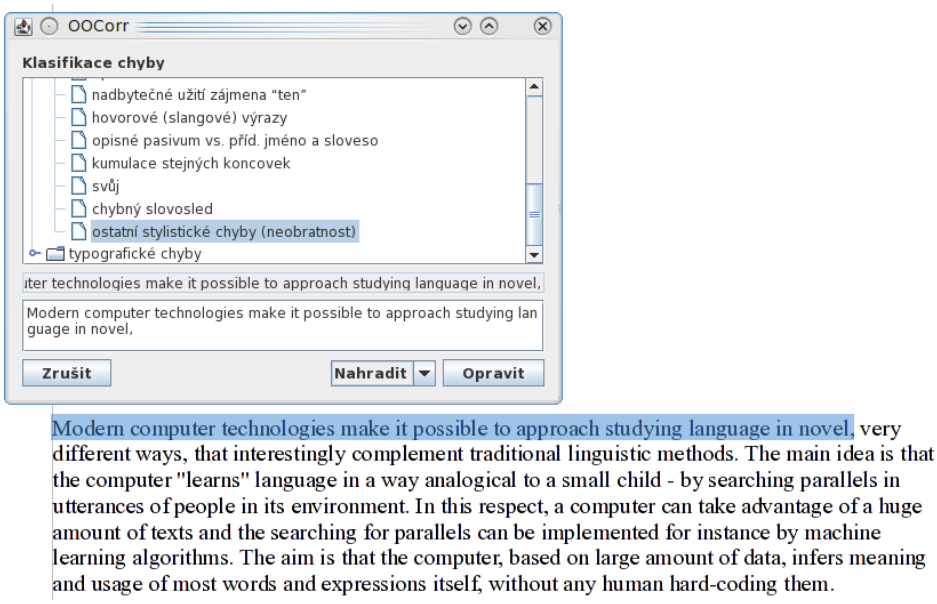

Fig. 1. Marking the text and choosing the error classification.

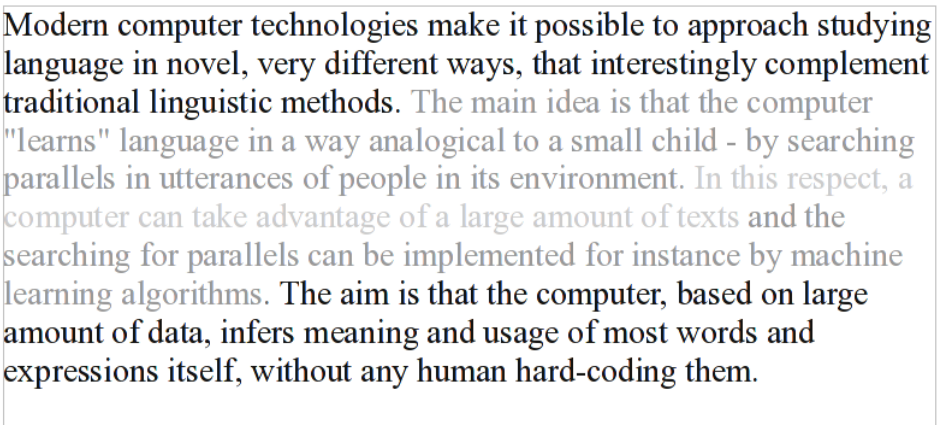

Fig. 2. OOCorr allows you to mark included errors; they are colored for better recognition.

classify a specific error have not been self-evident in some cases. There were two (or more) possibilities for correct classification.

Our wish is to simplify the classification, so that all the errors can be placed into one case only. That leads to building a new classification for error annotation.

– *Spelling (simple)*, error recognizable by a checker (`errtype=preklep`);

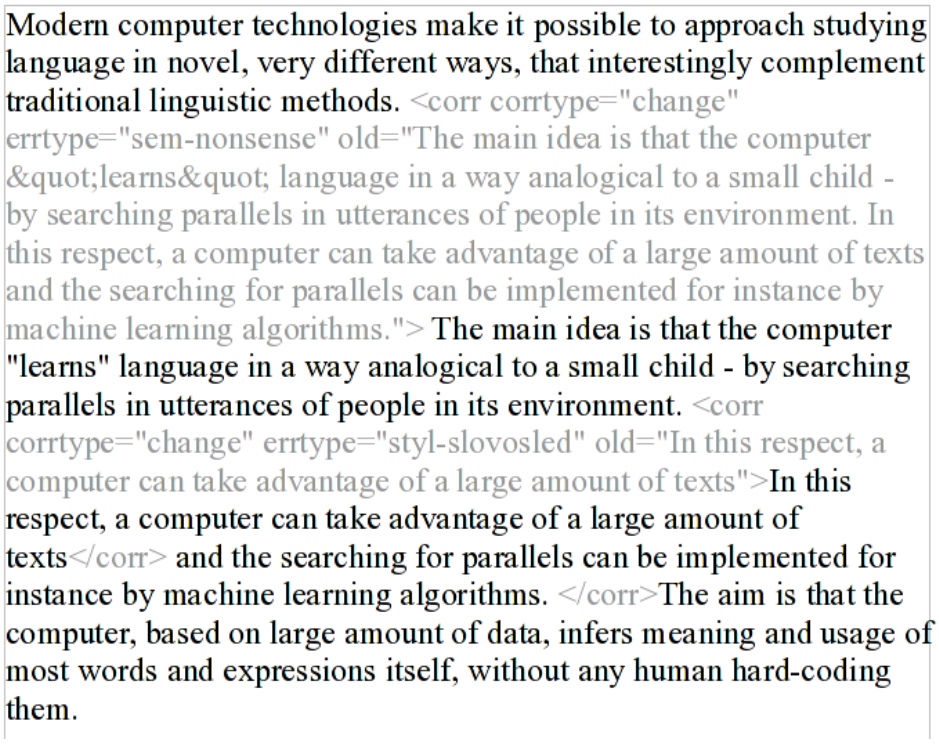

Fig. 3. The principle of the OOCorr is a feature of OpenOffice.org Writer, which is suited to handle "hidden text style". In this style the information about the error, its type and way of correction is saved.

- *Orthography*: error can not be recognized by a checker: punctuation error (`errtype=prav-interp`, *i/y* in specified words (`errtype=prav-iy`), *s/z* in specified words (`errtype=prav-sz`), using the right form of pronouns (`errtype=prav-mneme`), capital letters and lowercase (`errtype=prav-mala-velka`), composites (`errtype=prav-sprezky`), foreign words (`errtype=prav--prejata`), other orthographic errors (`errtype=prav-jine`);
- *Typography*: hyphen and dash (`errtype=typo-spojovnik`), hyphenation (`errtype=typo-delenislov`), division of text into paragraphs (`errtype=typo--odstavce`), spaces(`errtype=typo-mezery`), prepositions at the end of rows (`errtype=typo-predlozky`), brackets, quotes, overall layout and graphical outlook (`errtype=typo-jine`);
- *Morpho-syntactic errors*: morphologically wrong form (`errtype=synt-morf`), error in agreement (`errtype=synt-shoda`), verb and noun valencies (`errtype=synt-vazba`), possessives related errors (`errtype=synt-zamena`), other (`errtype=synt-jine`);

- *Lexico-semantic errors*: non meaningful expression (`errtype=sem-vyraz`), nonsensical or untrue steatment (`errtype=sem-nonsense`), and other semantics(`errtype=sem-jine`);
- *Stylistic*: word repeating (`errtype=styl-opakovani`), redundant use of demonstratives (`errtype=styl-ten`), using slang expressions (`errtype=styl--hovor`), cumulation of the same noun endings (`errtype=styl-koncovky`), incorrect stylistic word order (`errtype=styl-slovosled`), and other stylistic mistakes (`errtype=sytl-jine`).

8 The Differences

The new system is not error annotation specification dependent. There is a possibility to change the classification without changing the program just using another XML error definition file where all needed information is provided. The usage of more than one classification for different purposes is allowed as well as different language-specific settings.

To provide possibility of nesting errors (in up to three levels), the corr tag has been changed. Currently this tag is specified as a pair XML tag which means that, as opposed to the previous version, the *words* attribute which defined the length of the corrected text is not necessary anymore.

```
<corr errtype='string' corrtype='string' old='old text'>
new repaired text
</corr>
```

Fig. 4. New concept of the *corr* tag.

Simplified classification (only six main categories) helps us to build better corpus. It will be faster and easier for students to decide how to annotate their errors. Of course, precise annotation is crucial for getting accurate statistics from the corpus.

9 Results Based on the New Error Classification

At present we are not able to offer a detailed comparison of the Chyby with a standard corpus like DESAM [12] to see what differences exist in the distribution of the errors.

It is not surprising that the most frequent errors in the Chyby are stylistic ones (see Figure 5). The reason for this lies in the fact that the creators of the texts in the Chyby are students who are learning how to write. However, it is also true that the principles of good writing belong to the most neglected issues in the Czech high schools.

Error Group	count	%
Spelling (simple)	2,347	13.04
Morpho-syntactic	1,689	9.39
Spelling (other)	867	4.82
Lexico-semantics	2,536	14.09
Punctuation	3,837	21.32
Stylistic	4,184	23.25
Typography	2,165	12.03
unsorted	371	2.06
Total	**17,996**	100.00

Fig. 5. Error classification group statistics in the Chyby corpus.

The formal nature of stylistic errors is not very thoroughly explored even though they can be reliably identified in the texts. However, attempts to build a formal recognition procedure for them have been successful only partially.

The second most frequent error type is punctuation. Its high frequency is caused by the relative complexity of the Czech punctuation orthography rules and by the fact that the students do not possess the necessary writing skills at this level. The lexical and semantic errors also display a high frequency (3rd in order) for the same reasons. Recognition procedures for them, however, do not exist so far and they can be processed only manually.

10 Conclusions

In this paper we describe a Czech text corpus (Chyby) containing various kinds of errors – spelling, typographical, grammatical, style, lexical, etc. Resources for the Chyby come from the student's texts, reviews and essays written for the subject *Elements of Style*. They are corrected by the teachers and returned to the students who tag the marked errors and insert the respective corrections electronically into their texts. In this way the annotated corpus has been created.

The classification of the errors as they occur in the Chyby and the annotation scheme is presented together with the description of the tools used for inserting the tagged errors into the texts. The new tool developed for this purpose is OOCorr [3].

The present size of the corpus Chyby is approx. 500,000 word forms. It can be seen that the most frequent errors are stylistic ones – 23.25 %, followed by punctuation errors – 21.32 %, and lexical errors – 14.09 %.

The building of the Chyby and the analysis of the errors in the texts is a part of the larger project in the NLP Laboratory at FI MU whose goal is:

- to explore all types of errors that occur in the spontaneous texts,
- depending on the frequency and nature of the errors, to analyse whether effective procedures for an automatic correction can be designed,

- experiments not reported here (to be published in another paper) have already been performed so as to formulate an algorithm for automatic correcting punctuation errors using full parsing,
- to better map the area of stylistic errors and estimate what error detection rules can be developed in this respect for Czech texts.

Acknowledgement This research has been supported by Czech Ministry of Education, under the Grant LC536 and NPV 2C06009.

References

1. Pala, K., Rychlý, P., Smrž, P.: Text corpus with errors. In: Text, Speech and Dialogue: Sixth International Conference, TSD 2003, Berlin, Springer Verlag (2003) 90–97.
2. Kukačka, M.: Correcting errors in WinCorr. (Student Project at the Laboratory of Natural Language Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic) (2000).
3. Moravec Jaroslav: Korekturní rozšíření pro OpenOffice.org (bakalářská práce). Faculty of Informatics, Masaryk University, Brno (2009).
4. Burnard, L., ed.: Users Reference Guide for the British National Corpus. Oxford University Computing Service (1995).
5. Kocek, J., Kopřivová, M., Kučera, K., eds.: Český národní korpus – úvod a příručka uživatele (Czech National Corpus – Introduction and Users Guide). FF UK – ÚČNK (2000).
6. Rychlý, P., Smrž, P.: Manatee, Bonito and word sketches for Czech. In: Proceedings of the Second International Conference on Corpus Linguisitcs, Saint-Petersburg, Saint-Petersburg State University Press (2004) 124–132.
7. Hlavsa, Z., et al.: Akademická pravidla českého pravopisu (Rules of Czech Orthography). Akademia, Praha (1993).
8. ÚJČ AV ČR, FI MUNI: Internetová jazyková příručka. [online] `http://prirucka.ujc.cas.cz/` (2008).
9. Carlberger, J., Domeij, R., Kann, V., Kuntsson, O.: A Swedish grammar checker. `http://citeseer.nj.nec.com/305098.html` (2000).
10. Wei, Y.H., Davies, G.: Do grammar checkers work? `http://www.camsoftpartners.co.uk/euro96b.htm` (2002).
11. Negrilo, A.D., Fernandéz-Domínguez, J.: Error tagging systems for learner corpora. RESLA – Spanish Journal of Applied Linguistics **16** (2006) 83–102.
12. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530.