Text Corpus with Errors

Karel Pala, Pavel Rychlý, and Pavel Smrž

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

Abstract. This paper presents a description of a Czech text corpus (Chyby) containing various kinds of errors such as spelling, typographical, grammatical, style, lexical. We explain how Chyby has been built, how the errors in it have been discovered, marked and annotated. The classification of the errors is presented and the statistics concerning the types of errors is given. The tools for annotating the errors are also described. To the best of our knowledge, this is first text corpus of this sort prepared for Czech.

1 Introduction

The following common sense truth is valid: *nothing is perfect*. Whenever we produce texts of any kind we make mistakes of various kinds which leads to the next axiom being formulated: *in any text written by humans errors in spelling, grammar and typography will always occur.* That is why publishing houses and editorial boards have to employ readers and proof-readers whose task is to find errors existing in texts, proof-read them and finally produce printed texts of the best possible quality. The third axiom says: *if errors in texts are corrected by humans, all of them will never be removed.*

At present the prevailing majority of texts is produced on computers which in turn are used for typesetting, storing and dissemination them via Internet. Consequently, there is a strong tendency to use computers to correct texts and remove errors. Programs called *checkers* (spelling, grammar, style) have appeared that allow us to correct some well recognised errors in the texts. In some respects they are more reliable than humans and are able to remove some types of errors completely (see e. g. the spelling checker for MS Word and other text processors).

To be able to recognise all kinds of errors occurring in natural language texts, it is necessary to have a collection of texts (Czech in our case) with all the kinds of errors. Thus it was decided to build a text corpus which contains various spelling, grammatical, style, lexical, typographical (and other) errors, mark them and annotate them in the corpus text. The corpus is called **Chyby**.

In this paper we report on building the Czech Chyby corpus and on how errors are marked and annotated with the help of the tools (programs) especially developed for this purpose.

Some hints in this respect can be also found in [1] though his attention is focused on learner corpora where exploring the text errors is not its primary goal.

V. Matoušek and P. Mautner (Eds.): TSD 2003, LNAI 2807, pp. 90–97, 2003.

2 Building Chyby

At first glance it might seem that standard general corpora such as BNC [2] or the Czech National Corpus [3], could serve reasonably for these stated purposes. On 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 the number of the errors is acceptably small and the most serious 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 rather serious.

Thus we realized that *spontaneous texts (s-texts)* would be needed to build Chyby. When looking for s-texts we realized that our students take in the first two semesters of their studies, a subject called *Elements of Style* in which they have to write two kinds of texts – an essay and a review of a selected software product each of them approx. 600-700 words. The texts are corrected manually by two teachers (K. Pala and P. Peňáz) and returned to the students who have to prepare final corrected versions of their texts and annotate the marked errors electronically using two programs developed for this purpose (see below). Corrected and annotated texts are then included into Chyby (run under the corpus manager *Manatee* and the graphical interface *Bonito* [4]). At present the size of Chyby is approx. 410,000 word forms. Each semester about 200 students deliver their texts, thus Chyby is enlarged by some 100,000 word forms each semester.

The nature of the texts delivered by the students accords well with our idea of s-texts: the number of errors and their types are representative. In some cases the texts are not well written and in our view they contain a large percent of errors – in 650 words it is sometimes possible to find about 30 errors though not all errors are regularly related to the individual word forms, for example they include word order reordering, deleting and substituting whole lines. Interesting conclusions can be drawn from these data with regard to the quality of the previous high school education of our students but this analysis would go beyond the framework of the paper.

3 Texts and Errors in Them

Now that we have built Chyby, we are able to start reading its individual documents and discover that they contain errors in:

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

At present the errors are discovered manually, i.e. the two teachers read the student's text and mark the errors with an agreed notation or write in the student's text that the construction is wrong or that the particular expression is semantically not appropriate in the given context. This is the most laborious part of the work since the teachers

have to read the students' creations twice, the second time to check that the students have precisely corrected their errors. If the results of the second round are positive then students' creations are finally evaluated as acceptable. Those two steps are necessary – experience with the students tells us that they always try to get the best results with the least effort.

The annotation of the errors is done by the students electronically, i.e. they use one of the two programs designed for this purpose to insert the errors hand-marked by teacher into the text of their essay or review. Although the students are not expert linguists, the danger of imprecise annotating is reduced by the fact that the teacher's marking quite clearly indicate what type an error belongs to.

4 Annotation Scheme

In the previous section we indicated what kinds of errors can be found in Chyby. The next question is the design of the annotation scheme, in other words how the errors and their types can be described and classified. As far as we know there is no general theory of errors that may occur in the texts though 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. [5] or [6].

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

- *Spelling* errors 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 at 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 breaking valency frames. The Czech verbs in their valency frames require strictly the concrete cases, e. g. verb *zabít* 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, the rules for placing commas are syntactic, commas typically separate the main and subordinate clauses and are obligatory with some conjunctions. The frequency of the punctuation errors in Chyby is consequently rather 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 relations.
  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 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), incorrect use of passive constructions, long chains of noun groups, especially the prepositional ones and ambiguous uses of anaphoric pronouns, i. e. errors in reference and 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.

5 Orthography Rules – Norms

The starting point for our classification of the errors in texts and the annotation scheme based on it are the Rules of Czech Orthography [7], an official reference manual published by the Institute of Czech Language, Czech Academy of Sciences, Prague. It describes the basic principles of Czech Orthography which in comparison with English is much more phonetic though it is also governed by a number of the historical rules, such as in the area 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, etc. In this respect Czech punctuation is somewhat complicated

and in this way a large percentage of the punctuation errors in the student's texts can be explained.

As a whole the Rules represent a reference manual based on the empirical rules that can be characterised as basically deterministic (we estimate approx. 80%). We have found that it is reasonable to start with the Rules and in combination with the data obtained from Chyby, work out a more complete and formal description of the errors occurring in Czech texts together with their detailed classification and then to try to offer:

- error detecting rules for the particular classes of errors,
- the complete algorithms for checking and correcting the respective classes of errors in the texts as will be developed in the second stage,
- a theory of text errors as general as possible.

6 Tools for Tagging Errors in the Texts

The tagging of errors is a tedious task which we have tried to rationalise for all involved. 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/understanding of each particular grammatical phenomenon.

There are three ways to correct mistakes and tag the errors in a document and the students can choose the most appropriate for them:

1. All errors can be corrected and tagged manually, without software. This is done especially with the documents typeset in TEX format;
2. Students working with Microsoft Word can take advantage of a set of MS Visual Basic macros (called CorrMacros) that simplify error correcting and tagging;
3. Students can load their texts into a special application called WinCorr [8] that manages the whole process of error corrections and tagging.

The WinCorr application implements the functionality of a simple text editor in the environment of MS Windows. It works with the plain text and RTF (Rich Text File) data formats only (all other text formats could be easily transformed through the standard MS Clipboard). The function of this program is demonstrated by Figures 2 and 1.

The operation of the MS Word macros mimics the behaviour of WinCorr. It adds the Correction button that opens a dialogue window where the user defines the error type, how to perform the correction and the appropriate replacement. The user can select a mistake in the text by clicking the mouse or moving the cursor to the appropriate position and push the Correct button.

7 Results

The following table shows error counts for each error type in Chyby.

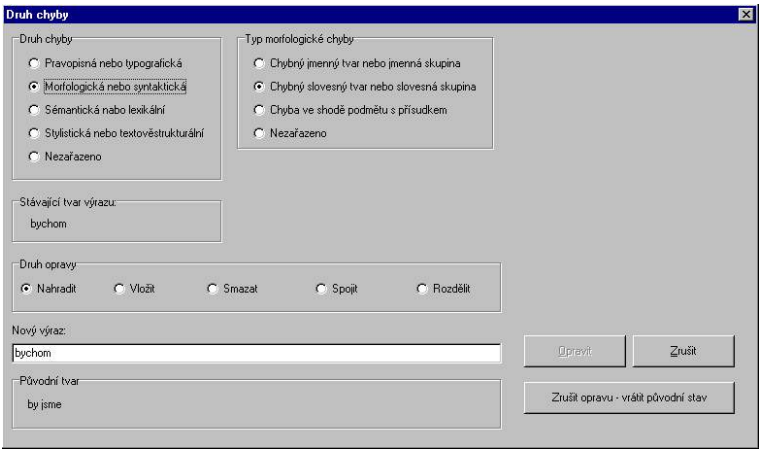

Fig. 1. Error tagging in WinCorr.

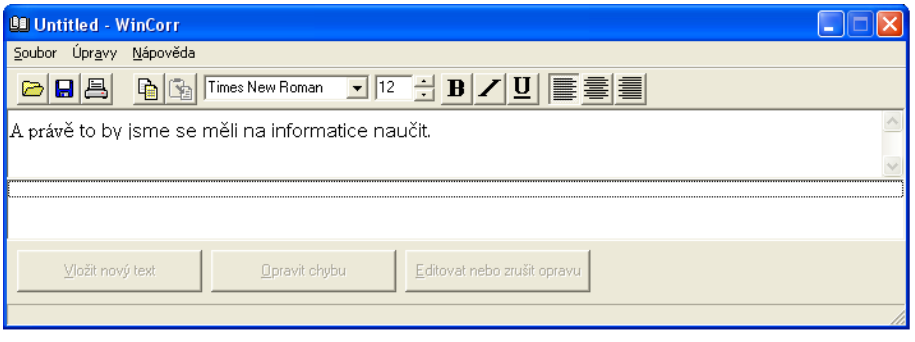

Fig. 2. A sentence correction in WinCorr.

Error Type		**%**
Spelling	1207	14.0
Typographical	1072	12.4
Morphological & syntactic	808	9.4
Punctuation	1830	21.2
Lexical	1553	18.0
Stylistic	2036	23.6
Other	133	1.5
Total	**8639**	100.0

At present we are not able to offer a detailed comparison of Chyby with a standard corpus like DESAM [9] to see what differences exist in the distribution of the errors.

It is not surprising that the most frequent errors in Chyby are stylistic ones. The reason for this lies in the fact that the creators of the texts in Chyby are students who

are learning how to write. However, it is also true that the principles of good writing are among the most neglected issues in the Czech high schools.

The nature of stylistic errors is not very thoroughly explored even though these errors can be reliably identified in the texts: their detection can only be formalised with difficulty.

The second most frequent error type is punctuation. Their high frequency is caused by the relative complexity of Czech punctuation orthography rules and by the fact that the students do not possess the necessary writing skills yet. In our view the lexical errors also display a high frequency (3rd in order) for the same reasons.

We have been consistently surprised by students not using spelling checkers for correction of their texts which is why the frequency of spelling errors is also quite high.

8 Conclusions

In this paper we describe a Czech text corpus (Chyby) containing various kinds of errors – spelling, typographical, grammatical, style, lexical, etc., and how it has been built in the NLP Laboratory at FI MU. Resources for 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 is created.

The classification of the errors as they occur in Chyby and the annotation scheme is presented together with the description of the tools used for inserting the tagged errors into the texts. The tools are two programs – WinCorr [8] and CorrMacros containing the annotation scheme and allowing classification and consequently editing of the errors in the texts.

The present size of Chyby is approx. 410,000 word forms. It can be seen that the most frequent errors are stylistic ones – 23.6 %, followed by punctuation errors – 21.2 %, and lexical errors – 18.0 %.

The building of 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,
- preliminary experiments not reported here (a topic for another paper) have already been performed so as to formulate an algorithm for automatically correcting punctuation errors which is based on the constraint grammars and rules written in the Karlsson and Voutilainen fashion using shallow parsing techniques [10],
- to better map the area of stylistic errors and to estimate what error detection rules could be developed in this respect for Czech texts.

Acknowledgement

This research has been supported by Czech Ministry of Education, Research Program CEZ:J07/98:143300003.

References

1. Leech, G.: Learner corpora: what they are and what can be done with them. In Granger, S., ed.: Learner English on Computer. Addison Wesley Longman, London and New York (1998) xiv–xx.
2. Burnard, L., ed.: Users Reference Guide for the British National Corpus. Oxford University Computing Service (1995)
3. Kocek, J., Kopřivová, M., Kučera, K., eds.: Český národní́ korpus – úvod a příručka uživatele (Czech National Corpus – Introduction and Users Guide). FF UK – ÚČNK (2000)
4. Rychlý, P.: Corpus Managers and Their Effective Implementation. PhD thesis, Faculty of Informatics, Masaryk University, Brno (2000)
5. Carlberger, J., Domeij, R., Kann, V., Kuntsson, O.: A swedish grammar checker. http://citeseer.nj.nec.com/305098.html (2000)
6. Wei, Y.H., Davies, G.: Do grammar checkers work? http://www.camsoftpartners.-co.uk/euro96b.htm (2002)
7. Hlavsa, Z., et al.: Akademická pravidla českého pravopisu (Rules of Czech Orthography). Akademia, Praha (1993)
8. Kukačka, M.: Correcting errors in WinCorr. (Student Project at the Laboratory of Natural Language Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic) (2000)
9. Pala, K., Rychlý, P., Smrž, P.: DESAM – an annotated corpus for Czech. In: Proceedings of SOFSEM'98, Springer (1998)
10. Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A., (Eds.): Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin (1995)