J. Yaghob (Ed.): ITAT 2015 pp. 73–80
Charles University in Prague, Prague, 2015

ITAT

Improvements to Korektor: A Case Study with Native and Non-Native Czech

Loganathan Ramasamy[1], Alexandr Rosen[2], and Pavel Straňák[1]

[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
[2]Institute of Theoretical and Computational Linguistics, Faculty of Arts
Charles University in Prague

Abstract: We present recent developments of Korektor, a statistical spell checking system. In addition to lexicon, Korektor uses language models to find real-word errors, detectable only in context. The models and error probabilities, learned from error corpora, are also used to suggest the most likely corrections. Korektor was originally trained on a small error corpus and used language models extracted from an in-house corpus WebColl. We show two recent improvements:

- We built new language models from freely available (shuffled) versions of the Czech National Corpus and show that these perform consistently better on texts produced both by native speakers and non-native learners of Czech.
- We trained new error models on a manually annotated learner corpus and show that they perform better than the standard error model (in error detection) not only for the learners' texts, but also for our standard evaluation data of native Czech. For error correction, the standard error model outperformed non-native models in 2 out of 3 test datasets.

We discuss reasons for this not-quite-intuitive improvement. Based on these findings and on an analysis of errors in both native and learners' Czech, we propose directions for further improvements of Korektor.

1 Introduction

The idea of using the context of a misspelled word to improve the performance of a spell checker is not new [10]. Moreover, recent years have seen the advance of context-aware spell checkers such as *Google Suggest*, offering reasonable corrections of search queries.

Methods used in such spell checkers usually employ the *noisy-channel* or *window-based* approach [4]. The system described here also belongs to the *noisy-channel* class. It makes extensive use of language models based on several morphological factors, exploiting the morphological richness of the target language.

Errors detected by such advanced spell checkers have a natural overlap with those of rule-based grammar checkers – grammatical errors are also manifested as unlikely n-grams. Using language models or even complete SMT approach [8] for grammatical error correction is also becoming more common, however all the tasks and publications on grammar correction we have seen so far expect pre-corrected text in terms of spelling. See also [15] and Table 1 in [14] for what types of errors were subject to correction at the CoNLL 2013 and 2014 Shared Tasks on English as a Second Language.

We make no such optimistic expectations. As we show in Section 2 there are many types of spelling errors both in native speakers' texts and in learner corpora. The error distributions are slightly different, though.

Richter [12] presented a robust spell checking system that includes language models for improved error detection and suggestion. To improve the suggestions further, the system employs error models trained on error corpora. In this paper we present some recent improvements to Richter et al.'s work in both respects: improved language models in Section 3 and task-dependent, adapted error models in Section 4. We apply native and non-native error models on both native and non-native datasets in Section 5. We analyze a portion of the systems output in Section 6 and provide some insight into the most problematic errors that various models make. Finally, we summarize our work and list potential scope for further improvements of Korektor components in Section 7.

2 Error Distribution for Native *vs* Non-Native Czech

Richter [11, p. 33] presents statistics of spelling errors in Czech, based on a small corpus of 9500 words, which is actually a transcript of an audio recording of a novel. The transcription was done by a native speaker. Following [1], the error analysis in Table 1 is based on the classification of errors into four basic groups: substitution, insertion, deletion/omission and swap/transposition/metathesis. Although the figures may be biased due to the small size of the corpus and the fact that it was transcribed by a single person, we still find them useful for a comparison with statistics of spelling errors made by non-native speakers.

In Table 2 the aggregate figures from Table 1 (in the last column headed by "Native") are compared with figures from an automatically corrected learner corpus ("SGT", or CzeSL-SGT) and a hand-corrected learner corpus ("MAN", or CzeSL-MAN). The taxonomy of errors is derived from a "formal error classification" used in those two corpora, described briefly in Section 4.[1] In this table we follow [3] in treating errors in diacritics as dis-

[1]See [7] for more details about the classification and the http://utkl.ff.cuni.cz/learncorp/ site, including all information about the corpora.

Error Type	Frequency	Percentage
Substitution	224	40.65%
– horizontally adjacent letters	142	25.77%
– vertically adjacent letters	2	0.36%
– z → s	6	1.09%
– s → z	1	0.18%
– y → i	10	1.81%
– i → y	10	1.81%
– non-adjacent vocals	13	2.36%
– diacritic confusion	21	3.81%
– other cases	19	3.45%
Insertion	235	42.65%
– horizontally adjacent letter	162	29.40%
– vertically adjacent letter	13	2.36%
– same letter as previous	14	2.54%
– other cases	46	8.35%
Deletion – other cases	58	10.53%
Swap letters	34	6.17%
TOTAL	551	100.00%

Table 1: Error types in a Czech text produced by native speakers

	SGT	MAN	PT	Native
Insertion	3.76	3.52	10.45	42.65
Omission	1.39	9.20	17.12	10.53
Substitution	31.30	37.67	12.82	36.84
Transposition	0.16	0.19	3.69	6.17
Missing diacritic	50.19	40.40	37.66	
Addition of diacritic	12.69	8.60	1.67	
Wrong diacritic	0.51	0.43	0.92	3.81

Table 2: Percentages of error types in a Czech text produced by non-native speakers, compared to Portuguese and Czech native speakers

tinct classes, adding their statistics on native Brazilian Portuguese for comparison in the "PT" column.

The high number of errors in diacritics in non-native Czech and native Portuguese in comparison with native Czech can be explained by the fact that native speakers of Czech are aware of the importance of diacritics both for distinguishing the meaning and for giving the text an appropriate status. The high number of errors in diacritics in learner texts is confirmed by results shown in Table 3, counted on the training portion of the "CzeSL-MAN" corpus by comparing the uncorrected and corrected forms, restricted to single-edit corrections.[2] The distribution is shown separately for the two annotation levels of CzeSL-MAN: somewhat simplifying, L1 is the level where non-words (forms spelled incorrectly in any context) are corrected, while L2 is the level where real-word errors are corrected (words correct out of context but incorrect in the syntactic context). For more details about CzeSL-MAN see Section 4.1.

[2] I.e., without using the "formal error types" of [7].

Error type	L1		L2	
Substitution[3]	22,695	84.36%	30,527	84.15%
– Case	1,827	8.05%	5,090	16.67%
– Diacritics	14,426	63.56%	13,367	43.79%
Insertion	1,274	4.74%	1,800	4.96%
Deletion	2,862	10.64%	3,809	10.50%
Swap	72	0.27%	143	0.39%
Total	26,903	100.00%	36,279	100.00%

Table 3: Distribution of single edit errors in the training portion of the CzeSL-MAN corpus on Levels 1 and 2

Substituting...	Frequency	Substituting...	Frequency
a for *á*	5255	*y* for *ý*	780
i for *í*	3427	*á* for *a*	695
e for *ě*	1284	*u* for *ů*	635
e for *é*	1169	*y* for *i*	482
i for *y*	1077	*í* for *ý*	330
í for *i*	1005	*z* for *ž*	297

Table 4: The top 12 most frequent substitution errors in the CzeSL corpus

As an illustration of the prevalence of errors in diacritics in non-native Czech, see Table 4, showing the 12 most frequent substitution errors from L1 in Table 3. There is only one error which is not an error in a diacritic (the use of the *i* homophone instead of *y*).

3 Current Improvements for Native Czech Spelling Correction

The original language model component of Korektor [12] was trained on *WebColl* – a 111 million words corpus of primarily news articles from the web. This corpus has two issues: (i) the texts are not representative and (ii) the language model from this data could not be distributed freely due to licensing issues. To obviate this, we evaluate Korektor using two new language models built from two corpora available from the *Czech National Corpus* (*CNC*): (i) SYN2005 [2] and (ii) SYN2010 [9]. Both have the size of 100 million words each and have a balanced representation of contemporary written Czech: news, fiction, professional literature etc.

We use the error model and the test data (only the *Audio* data set) described in [12]. *Audio* contains 1371 words with 218 spelling errors, of which 12 are real-word errors.

[3] The two error types below are actually subtypes of the substitution error.

For the CNC corpora, we build 3^{rd} order language models using *KenLM* [6].

The spell checker accuracy is measured in terms of standard precision and recall. The precision and recall measures are calculated at two levels: (i) error detection and (ii) error correction. These evaluation measures are similar in spirit as in [17]. For both levels, precision, recall and other related measures are calculated as: $Precision(P) = \frac{TP}{TP+FP}$, $Recall(R) = \frac{TP}{TP+FN}$, and $F-score(F1) = \frac{2*P*R}{P+R}$, where, for error detection,

- **TP** – Number of words with spelling errors that the spell checker detected correctly
- **FP** – Number of words identified as spelling errors that are not actually spelling errors
- **TN** – Number of correct words that the spell checker did not flag as having spelling errors
- **FN** – Number of words with spelling errors that the spell checker did not flag as having spelling errors

and for error correction,

- **TP** – Number of words with spelling errors for which the spell checker gave the correct suggestion
- **FP** – Number of words (with/without spelling errors) for which the spell checker made suggestions, and for those, either the suggestion is not needed (in the case of non-existing errors) or the suggestion is incorrect if indeed there was an error in the original word
- **TN** – Number of correct words that the spell checker did not flag as having spelling errors and no suggestions were made
- **FN** – Number of words with spelling errors that the spell checker did not flag as having spelling errors or did not provide any suggestions

The results for error detection and error correction are shown in Tables 5 and 6, respectively. Maximum edit distance, i.e., the number of edit operations per word is set to values from 1 to 5. In the case of error detection, the best overall performance is obtained for the SYN2005 corpus when the maximum edit distance parameter is 2, and there is no change in results for the edit distance range from 3 to 5. Of the two CNC corpora, SYN2005 consistently provides better results than SYN2010 corpus. Differences in the vocabulary could be the most likely reason.

Even in the case of error correction, the best overall performance is obtained for SYN2005 with 94.5% F1-score. We can also see that WebColl performs better in 3 out of 5 cases, but we should also note that this happens when we include top-3 suggestions in the error correction. Otherwise the SYN2005 model consistently provides better scores. We have also experimented with pruned language models and obtained similar results.

LM train data	Max. edit distance	P	R	F1
WebColl		94.7	90.8	92.7
SYN2005	1	95.7	90.8	**93.2**
SYN2010		94.7	89.9	92.2
WebColl		94.1	95.4	94.8
SYN2005	2	95.0	95.9	**95.4**
SYN2010		94.1	95.0	94.5
WebColl		94.1	95.4	94.8
SYN2005	3	95.0	95.9	**95.4**
SYN2010		94.1	95.0	94.5
WebColl		94.1	95.4	94.8
SYN2005	4	95.0	95.9	**95.4**
SYN2010		94.1	95.0	94.5
WebColl		94.1	95.4	94.8
SYN2005	5	95.0	95.9	**95.4**
SYN2010		94.1	95.0	94.5

Table 5: Error detection results with respect to different language models

4 Work in Progress for Improving Spelling Correction of Non-Native Czech

One of the main hurdle in obtaining a new error model is the availability of annotated error data for training. Many approaches are available to somehow obtain error data automatically from sources such as the web [16]. The error data obtained from the web may be good enough for handling simple typing errors, but not for the more complicated misspellings a learner/non-native speaker of a language makes. However, these approaches can be successfully used to obtain general purpose spell checkers. One resource which could be of some value to spell checking is the learner corpus. Unlike native error corpus, the learner corpus of non-native or foreign speakers tend to have more errors ranging from orthographical, morphological to real-word errors. In this work, we try to address whether error models from texts produced by native Czech speakers can be applied to errors from non-native Czech texts and vice versa. We also derive error analysis based on the results.

4.1 CzeSL — a Corpus of Czech as a Second Language

A learner corpus consists of language produced by language learners, typically learners of a second or foreign language. Deviant forms and expressions can be corrected and/or annotated by tags making the nature of the error explicit. The annotation scheme in CzeSL is based on a two-stage annotation design, consisting of three levels. The level of transcribed input (Level 0) is followed by the level of orthographical and morphological corrections (Level 1), where only forms incorrect in any context are treated. The

LM train data	Max. edit distance	top-1			top-2			top-3		
		P	R	F1	P	R	F1	P	R	F1
WebColl	1	85.2	89.9	87.5	90.9	90.5	90.7	93.3	90.7	92.0
SYN2005	1	87.9	90.1	**89.0**	92.3	90.5	**91.4**	93.7	90.7	**92.2**
SYN2010	1	86.0	89.0	87.5	91.8	89.6	90.7	92.3	89.7	91.0
WebColl	2	84.2	94.9	89.2	91.0	95.3	93.1	93.2	95.4	94.3
SYN2005	2	86.8	95.5	**91.0**	91.8	95.7	**93.7**	93.2	95.8	**94.5**
SYN2010	2	85.0	94.4	89.5	91.4	94.8	93.1	92.3	94.9	93.5
WebColl	3	84.2	94.9	89.2	91.0	95.3	93.1	93.2	95.4	**94.3**
SYN2005	3	86.8	95.5	**91.0**	91.4	95.7	**93.5**	92.7	95.8	94.2
SYN2010	3	85.0	94.4	89.5	90.9	94.8	92.8	91.8	94.8	93.3
WebColl	4	84.2	94.9	89.2	91.0	95.3	93.1	93.2	95.4	**94.3**
SYN2005	4	86.8	95.5	**91.0**	91.4	95.7	**93.5**	92.7	95.8	94.2
SYN2010	4	85.0	94.4	89.5	90.9	94.8	92.8	91.8	94.8	93.3
WebColl	5	84.2	94.9	89.2	91.0	95.3	93.1	93.2	95.4	**94.3**
SYN2005	5	86.8	95.5	**91.0**	91.4	95.7	**93.5**	92.7	95.8	94.2
SYN2010	5	85.0	94.4	89.5	90.9	94.8	92.8	91.8	94.8	93.3

Table 6: Error correction results with respect to different language models

result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. All other types of errors are corrected at Level 2.[4]

This annotation scheme was meant to be used by human annotators. However, the size of the full corpus and the costs of its manual annotation have led us to apply automatic annotation and find ways of its improvement.

The hand-annotated part of the corpus (CzeSL-MAN) now consists of 294 thousand word tokens in 2225 short essays, originally hand-written and transcribed.[5] A part of the corpus is annotated independently by two annotators: 121 thousand word tokens in 955 texts. The authors are both foreign learners of Czech and Czech learners whose first language is the Romani ethnolect of Czech.

The entire CzeSL corpus (CzeSL-PLAIN) includes about 2 mil. word tokens. This corpus comprises transcripts of essays of foreign learners and Czech students with the Romani background, and also Czech Bachelor and Master theses written by foreigners.

The part consisting of essays of foreign learners only includes about 1.1 word tokens. It is available as the CzeSL-SGT corpus with full metadata and automatic annotation, including corrections proposed by Korektor, using the original language model trained on the WebColl corpus.[6] In the annotation Korektor detected and corrected 13.24% incorrect forms, 10.33% labeled as including a spelling error, and 2.92% an error in grammar, i.e. a 'real-word' error. Both the original, uncorrected texts and their corrected version was tagged and lemmatized, and "formal error tags," based on the comparison of the uncorrected and corrected forms, were assigned. The share of 'out of lexicon' forms, as detected by the tagger, is slightly lower – 9.23%.

4.2 The CzeSL-MAN Error Models

We built two error models from the CzeSL-MAN corpus – one for Level 1 (L1) errors and another for Level 2 (L2) errors. As explained in Section 4.1 above, L1 errors are mainly non-word errors and L2 errors belong to real-word and grammatical errors, but still include form errors that are not corrected at L1 because the faulty form happens to be spelled as a form which would be correct in a different context. Extracting errors from the XML format used for encoding the original and the corrected text at L1 is straightforward. The only thing needed is to follow the links connecting tokens at L0 (the original tokens) and L1 (the corrected tokens) and to extract tokens for which the links are labeled as correction links. In the error extraction process, we do not extract errors that involve joining or splitting of word tokens at either level (Korektor does not handle incorrectly split or joined words at the moment).

L2 errors include not only errors identified between L1 and L2 but also those identified already between L0 and L1, if any. This is because L2 tokens are linked to L0 tokens through L1 tokens, rather than being linked directly. For example, consider a single token at Levels L0, L1 and L2: *všechy* (L0) $\xrightarrow{formSingCh,incorBase}$ *všechny* (L1) $\xrightarrow{agr}$ *všichni* (L2). The arrow stands for a link between the two levels, optionally with one or more error labels. For the L1 error extraction, the extracted pair of an incorrect token and a correct token is (*všechy*, *všechny*) with the error labels (*formSingCh*, *incorBase*), and for the L2 error extraction, the extracted error and correct token pair is

[4] See [5] and [13] for more details.

[5] For an overview of corpora built as a part of the CzeSL project and relevant links see http://utkl.ff.cuni.cz/learncorp/.

[6] See http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-en.pdf.

Error	CzeSL-L1		CzeSL-L2	
	train	test	train	test
single-edit	73.54	72.24	67.02	69.30
multi-edit	26.46	27.76	32.98	30.70

Table 7: Percentage of single and multi edit-distance errors in the train/test of L1 and L2 errors.

(*všechy*, *všichni*) with the error labels (*formSingCh, incorBase, agr*). For the L2 errors, we project the error labels of L1 onto L2. If there is no error present or annotated between L0 and L1, then we use the error annotation between L2 and L1. The extracted incorrect token is still from L0 and the correct token from L2.

Many studies have shown that most misspellings are single-edit errors, i.e., misspelled words differ from their correct spelling by exactly one letter. This also holds for our extracted L1 and L2 errors (Table 7). We train our L1 and L2 errors on single-edit errors only, thus the models are quite similar to the native Czech error model described in [11]. The error training is based on [1]. Error probabilities are calculated for the four single-edit operations: *substitution, insertion, deletion,* and *swap*.

5 Experiments with Native and Non-Native Error Models

For the native error model (*webcoll*), we use the same model as described in [12]. For the non-native error models, we create two error models as described in Section 4.2: (i) *czesl_L1* – trained on the L1 errors (CzeSL-L1 data in Table 8) and (ii) *czesl_L2* – trained on the L2 errors (CzeSL-L2 data in Table 8). We partition the CzeSL-MAN corpus in the 9:1 proportion for training and testing.

The non-native training data include more errors than those automatically mined from web. The training of non-native error models is done on single-edit errors only (refer Table 7 for the percentage of errors used for training). For the language model, we use the best model (SYN2005) that we obtained from Section 3.

We perform evaluation on all kinds of errors in test data. We also set the maximum edit distance parameter to 2 for all our experiments. We arrived at this value based on our observation in various experiments. We run our native and non-native models on the test data described in Table 9, and their results are given in Table 10. Error correction results are shown for top-3 suggestions.

In error detection, in terms of F1-score, *czesl_L2* model posts better score than the other two models for both native and non-native data sets. When it comes to error correction, the native model *webcoll* seems to perform better in 2 out of 3 data sets, and the next better performer being the *czesl_L2* model. One has to note that, the non-native models are not tuned to any particular phenomenon such as capitalization or keyboard layouts, so there is still some scope for improvements on the non-native error models. While *webcoll* and *czesl_L2* models help each other in the opposite direction, i.e., the performance of native model on the non-native data and vice versa, the *czesl_L1* model works better only on the CzeSL-L1 dataset. In other words, since L1 error annotation did not involve complete correction of the test data of CzesL-MAN, they can be used, for instance, the correction of misspellings that do not involve grammar errors.

Train data	Corpus size	#Errors
WebColl	111M	12,761
CzeSL-L1	383K	36,584
CzeSL-L2	370K	54,131

Table 8: Training data for native and non-native experiments. The errors include both single and multi-edit errors.

Test data	Corpus size	#Errors
Audio	1,371	218
CzeSL-L1	33,169	3,908
CzeSL-L2	32,597	5,217

Table 9: Test set for native and non-native experiments. The errors include both single and multi-edit errors.

6 Discussion

We manually analyzed a part (the top 3000 tokens) of the output of Korektor for the CzeSL-L2 test data for all the three models. We broadly classify the test data as having *form* errors (occurring between the L0 and L1 level), grammar (*gram*) errors (occurring between L1 and L2) and accumulated errors (*form+gram*, where errors are present at all levels – between L0 and L1, and L1 and L2). The CzeSL-L2 test data can include any of the above types of errors. About 23% of our analyzed data include one of the above errors. More than half of the errors (around 62%) belong to the *form* errors and about 27% belong to the *gram* class. The remaining errors are the *form+gram* errors.

In the case of *form* errors, both the native (*webcoll*) and the non-native models (*czesl_L1* and *czesl_L2*) detect errors at the rate of more than 89%. Form errors may or may not be systematic and they are easily detected by all the three models. Most of the error instances in the data can be categorized under either missing/addition of diacritics, or they can occur in combination with other types of errors, for instance, *přítelkyně* was incorrectly written as *přatelkine*.

Model	Error detection									Error correction								
	Audio			CzeSL-L1			CzeSL-L2			Audio			CzeSL-L1			CzeSL-L2		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
webcoll	95.0	95.9	95.4	81.8	81.7	81.7	91.0	65.0	75.9	93.2	95.8	94.5	71.7	79.6	**75.4**	78.0	61.5	**68.8**
czesl_L1	95.0	96.8	**95.9**	82.2	82.2	**82.2**	91.1	64.4	75.4	93.7	96.7	**95.2**	70.2	79.8	74.7	75.5	60.0	66.8
czesl_L2	95.0	96.8	**95.9**	81.2	82.7	81.9	90.9	65.4	**76.1**	93.7	96.7	**95.2**	68.2	80.0	73.6	74.9	60.9	67.2

Table 10: Error models applied to native and non-native Czech

```
Error label: "form:formCaron0 + formSingCh
+ formY0 + incorBase + incorInfl"
Error token: přatelkine
Gold token: přítelkyně
webcoll: přatelkine
czesl_L1: <suggestions="přítelkyně|pritelkyne
|přátelíme">
czesl_L2: <suggestions="přítelkyně|pritelkyne
|přátelíme">
```

In the case of *gram* errors, most of the errors are undetected. Out of 193 *gram* errors in our analyzed data, the percentage of errors detected by the models are: *webcoll* (15.5%), *czesl_L1* (9.3%) and *czesl_L2* (15.0%). Most of the grammar errors involve agreement, dependency and lexical errors. The agreement errors are shown in Table 11. Except for a few pairs such as *jedné* → *jednou* (incorrect → correct), *mě* → *mé*, *který* → *kteří*, *teplí* → *teplý*, most of the error tokens involving agreement errors have not been recognized by any of the three models.[7]

Dependency errors (e.g. a wrongly assigned morphological case, missing a syntactic governor's valency requirement) such as $roku_{\text{GEN}}$ → $roce_{\text{LOC}}$ 'year', $kolej_{\text{ACC}}$ → $koleji_{\text{LOC}}$ 'dormitory', $roku_{\text{SG}}$ → $roky_{\text{PL}}$ 'year', $restauraci_{\text{LOC}}$ → $restaurace_{\text{NOM}}$ 'restaurant' have not been recognized by any of the models. The pair mi_{DAT} → $mě_{\text{ACC}}$ 'me' has been successfully recognized by all the three models and the correct suggestion listed in the top:

```
Error label: "gram:dep"
Error token: mi
Gold token: mě
webcoll: <suggestions="mě|mi|ji|mu|i">
czesl_L1: <suggestions="mě|mi|ji|mu|si">
czesl_L2: <suggestions="mě|mi|ji|mu|ho">
```

For instance, the pair *ve* → *v* 'in' (vocalized → unvocalized) has been recognized by the *webcoll* and *czesl_L2* models, but not by the *czesl_L1* model. When it comes to grammar errors, *webcoll* and *czesl_L2* have better performance than *czesl_L1*. It was expected, because the *czesl_L1* model was not trained on grammar errors.

When the error involved a combination of *form* and *gram* errors, all the three models tend to perform better. Most of the *form+gram* errors were recognized by all the three models: *webcoll* (85%), *czesl_L1* (86%) and *czesl_L2* (89%). For instance, the error pair **zajímavy* → *zajímavé* 'interesting' that was labeled at both L1 and L2 level was successfully recognized by all the models, and the correct suggestions were listed in the top. There were

incorrect usage	correct usage	category	gloss
$bavím_{\text{SG}}$	$bavíme_{\text{PL}}$	number	enjoy
byl_{SG}	$byly_{\text{PL}}$	number	was → were
byl_{SG}	$Byly_{\text{PL}}$	number	was → were
$Chci_{\text{1ST}}$	$Chce_{\text{3RD}}$	person	want → wants
$chodím_{\text{SG}}$	$chodíme_{\text{PL}}$	number	walk
$Chtěla_{\text{FEM}}$	$Chtěl_{\text{MASC}}$	gender	wanted
$dívat_{\text{INF}}$	$dívá_{\text{3RD}}$	verb form	to see → sees
$dobré_{\text{FEM}}$	$dobří_{\text{MASC.ANIM}}$	gender	good
$dobrý_{\text{MASC}}$	$dobrá_{\text{FEM}}$	gender	good
$druhý_{\text{NOM}}$	$druhého_{\text{GEN}}$	case	2nd, other
$hezké_{\text{PL}}$	$hezký_{\text{SG}}$	number	nice
je_{SG}	$jsou_{\text{PL}}$	number	is → are
$jednou_{\text{INS}}$	$jedné_{\text{LOC}}$	case	one
$jich_{\text{GEN}}$	je_{ACC}	case	them
$jsem_{\text{SG}}$	$jsme_{\text{PL}}$	number	am → are
$jsme_{\text{PL}}$	$jsem_{\text{SG}}$	number	are → am
$jsou_{\text{PL}}$	je_{SG}	number	are → is
$který_{\text{SG}}$	$kteří_{\text{PL}}$	number	which
$leželi_{\text{MASC.ANIM}}$	$ležely_{\text{FEM}}$	gender	lay
$malý_{\text{SG}}$	$malé_{\text{PL}}$	number	small
$malých_{\text{GEN}}$	$malé_{\text{ACC}}$	case	small
$mé_{\text{ACC}}$	$mí_{\text{NOM}}$	number	my
$Mě_{\text{PERS.PRON}}$	$Mé_{\text{POSS.PRON}}$	POS	me → my
$miluju_{\text{1ST}}$	$miluje_{\text{3RD}}$	person	love → loves
$mohli_{\text{MASC.ANIM}}$	$mohly_{\text{FEM}}$	gender	could
$nemocní_{\text{PL}}$	$nemocný_{\text{SG}}$	number	ill
$nich_{\text{LOC}}$	$ně_{\text{ACC}}$	case	them
$oslavili_{\text{MASC.ANIM}}$	$oslavila_{\text{NEUT}}$	gender	celebrated
$pracovní_{\text{NOM}}$	$pracovním_{\text{INS}}$	case	work-related
$pracuji_{\text{1ST}}$	$pracuje_{\text{3RD}}$	person	work → wants
$Studovali_{\text{MASC.ANIM}}$	$studovaly_{\text{FEM}}$	gender	studied
$teple_{\text{ADV}}$ cc	$teplé_{\text{ADJ}}$	POS	warmly → warm
$teplí_{\text{PL}}$	$teplý_{\text{SG}}$	number	warm
$tří_{\text{GEN}}$	$tři_{\text{ACC}}$	case	three
$tuhle_{\text{FEM}}$	$Tenhle_{\text{MASC}}$	gender	this
$typické_{\text{FEM}}$	$typická_{\text{NEUT}}$	gender	typical
$velké_{\text{PL}}$	$velký_{\text{SG}}$	number	big

Table 11: Some of the agreement errors in the analyzed portion of the CzeSL-L2 test data

[7] The category glosses should be taken with a grain of salt: many forms can have several interpretations. E.g. $oslavili_{\text{MASC.ANIM}}$ → $oslavila_{\text{NEUT}}$ 'celebrated' could also be glossed as $oslavili_{\text{PL,MASC.ANIM}}$ → $oslavila_{\text{SG,FEM}}$.

many errors that were successfully recognized, but the correct suggestions did not appear in top-3, such as, **nechcí* → *nechtěl* 'didn't want', **mym* → *svým* 'my', **kamarad* → *kamaráda* 'friend', **vzdělany* → *vzdělaná* 'educated'.

Based on the results in Table 10 and the manual error analysis in this section, we can make the following general observations:

- Non-native Czech models can be applied to native test data and obtain even better results than the native Czech model (Table 10).
- From the manual analysis of the test outputs of both native and non-native Czech models, the most problematic errors are the grammar errors due to missed agreement or government (valency requirements). Some of the grammar errors involve most commonly occurring Czech forms such as *jsme, byl, dobrý, je, druhý*.
- Both native and non-native error models perform well on spelling-only errors.
- The CzeSL-MAN error data include errors that involve joining/splitting of word forms that we did not handle in our experiments. We also skipped word order issues in the non-native errors which are beyond the scope of current spell checker systems.

7 Conclusions and Future Work

We have tried to improve both the language model and the error model component of Korektor, a Czech statistical spell checker. Language model improvements involved the employment of more balanced corpora from the Czech National Corpus, namely SYN2005 and SYN2010. We obtained better results for the SYN2005 corpus.

Error model improvements involved creating non-native error models from CzeSL-MAN, a hand-annotated Czech learner corpus, and a series of experiments with native and non-native Czech data sets. The state-of-the-art improvement for the native Czech data set comes from the non-native Czech models trained on L1 and L2 errors from CzeSL-MAN. Surprisingly, the native Czech model performed better for non-native Czech (L2 data) than the non-native models. This we attribute to the rich source of learner error data, since the texts come from very different texts: Czech students with Romani background, as well as learners with various proficiency levels and first languages. Another potential reason could be the untuned nature of the non-native error models that may require further improvement.

As for future work aimed at further improvements of Korektor, we plan to explore model combinations with native and non-native Czech models. We would also like to extend Korektor to cover new languages so that more analysis results could be obtained. To improve error models further, we would like to investigate how the more complex grammar errors such as those in agreement and form errors such as joining/splitting of word forms can be modeled. Further, we would like to analyze non-native Czech models, so that Korektor can be used to annotate a large Czech learner corpus such as CzeSL-SGT more reliably.

References

[1] Church, K., Gale, W.: Probability scoring for spelling correction. Statistics and Computing **1(7)** (1991), 93–103

[2] Čermák, F., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Křen, M., Novotná, R., Petkevič, V., Schmiedtová, V., Skoumalová, H., Spoustová, J., Šulc, M., Velíšek, Z.: SYN2005: a balanced corpus of written Czech, 2005

[3] Gimenes, P. A., Roman, N. T., Carvalho, A. M. B. R.: Spelling error patterns in Brazilian Portuguese. Computational Linguistics **41(1)** (2015), 175–183

[4] Golding, A. R., Roth, D.: A window-based approach to context-sensitive spelling correction. Machine Learning **34** (1999), 107–130 10.1023/A:1007545901558.

[5] Hana, J., Rosen, A., Škodová, S., Štindlová, B.: Error-tagged learner corpus of Czech. In: Proceedings of the Fourth Linguistic Annotation Workshop, Uppsala, Sweden, Association for Computational Linguistics, 2010

[6] Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, 187–197, Edinburgh, Scotland, United Kingdom, 2011

[7] Jelínek, T., Štindlová, B., Rosen, A., Hana, J.: Combining manual and automatic annotation of a learner corpus. In: Sojka, P., Horák, A., Kopeček, I., Pala, K., (eds.), Text, Speech and Dialogue – Proceedings of the 15th International Conference TSD 2012, number 7499 in Lecture Notes in Computer Science, 127–134, Springer, 2012

[8] Junczys-Dowmunt, M., Grundkiewicz, R.: The AMU System in the CoNLL-2014 Shared Task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 25–33, Baltimore, Maryland, Association for Computational Linguistics, 2014

[9] Křen, M., Bartoň, T., Cvrček, V., Hnátková, M., Jelínek, T., Kocek, J., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V., Skoumalová, H.: SYN2010: a balanced corpus of written Czech, 2010

[10] Mays, E., Damerau, F. J., Mercer, R. L.: Context based spelling correction. Information Processing & Management **27 (5)** (1991), 517–522

[11] Richter, M.: An advanced spell checker of Czech. Master's Thesis, Faculty of Mathematics and Physics, Charles University, Prague, 2010

[12] Richter, M., Straňák, P., Rosen, A.: Korektor — a system for contextual spell-checking and diacritics completion. In: Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), 1019–1027, Mumbai, India, (2012), Coling 2012 Organizing Committee

[13] Rosen, A., Hana, J., Štindlová, B., Feldman, A.: Evaluating and automating the annotation of a learner corpus. Language Resources and Evaluation — Special Issue: Resources for language learning **48 (1)** (2014), 65–92

[14] Rozovskaya, A., Chang, K. -W., Sammons, M., Roth, D., Habash, N.: The Illinois-Columbia System in the CoNLL-2014 Shared Task. In: CoNLL Shared Task, 2014

[15] Rozovskaya, A., Roth, D.: Building a state-of-the-art grammatical error correction system, 2014

[16] Whitelaw, C., Hutchinson, B., Chung, G. Y., Ellis, G.: Using the web for language independent spellchecking and autocorrection. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing – Volume 2, EMNLP'09, 890–899, Stroudsburg, PA, USA, Association for Computational Linguistics, 2009

[17] Wu, S. -H., Liu, C. -L., Lee, L. -H.: Chinese spelling check evaluation at SIGHAN Bake-off 2013. In: Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 35–42, Nagoya, Japan, Asian Federation of Natural Language Processing, 2013