Introducing a corpus of non-native Czech with automatic annotation[1]

Alexandr Rosen
Institute of Theoretical and Computational Linguistics
Faculty of Arts, Charles University in Prague

1 Introduction

Texts in a learner corpus can be annotated in two independent ways: (i) by standard linguistic categories: morphosyntactic tags, base forms, syntactic structure and functions, and (ii) by error annotation: corrected word forms (target hypotheses), and categories specifying the nature of errors. Reasonably reliable methodologies and tools are available for linguistic annotation (i) of many languages, as long as the text is produced by native speakers. The situation is different for non-standard language of non-native learners and for error annotation (ii), where manual annotation is quite common. However, with the growing volumes of learner corpora, the need for methods and tools simplifying such tasks is increasing.

In §2 we provide a glimpse of the landscape of learner language annotation. Then, after an overview of existing learner corpora of Czech, including a corpus of texts written by non-native learners of Czech in §3, we show that useful results can be achieved by applying tools developed for standard language. The core part of this contribution (§4) is concerned with the CzeSL-SGT corpus (*Czech as a Second Language with Spelling, Grammar and Tags*), which includes transcripts of essays hand-written by (mostly young) non-native learners of Czech in 2009–2013. The corpus includes about 8.6 thousand texts by nearly two thousand native speakers of 54 languages; altogether about 1 million words (for details about the corpus content see Table 4).

Most texts are equipped with metadata (§4.1). Word forms are tagged by word class, morphological categories and base forms (lemmas). Forms detected as incorrect (including some real-word errors) are corrected by a stochastic spelling and grammar checker and the resulting texts are tagged again. Original and corrected forms are compared and error labels are assigned, based on criteria applicable in a formally specifiable way. All the annotation is assigned automatically (see §4.2).

The corpus is available either for on-line searching using the search interface of the Czech National Corpus (http://korpus.cz), or for download from the LINDAT data repository (http://www.lindat.cz), see §4.3.

Automatic corrections have been evaluated using an existing manually annotated subset of the corpus (the manual annotation includes a target hypothesis about the form, see §5).

Finally in §6 we discuss some challenging aspects of the corpus and its annotation a show perspectives for its development and use.

[1] The paper reports on the CzeSL-SGT corpus, which was built from texts collected within the ESF project CZ.1.07/2.2.00/07.0119 and the Charles University PRVOUK P10 funding program. Work on the corpus itself was supported by the Ministry of Education of the Czech Republic as a part of the Czech National Corpus project LM2011023.

2 Automating the annotation of learner texts

Some annotation tools designed for native language, such as taggers, lemmatizers, parsers, spelling and grammar checkers can be applied to the original text including learner language, or to their corrected version, even though the success rate depends on how much the texts deviate from the standard language. On the other hand, the task of categorizing errors is usually a manual exercise and its methodology is far from established. Some error taxonomies have a more prominent position (Dagneaux et al. 2008; Granger et al. 2002), but there are quite a few other annotation schemes used in practice (for an overview see Štindlová 2013: 71f). Moreover, such taxonomies often assume a target hypothesis and even if motivated mainly by formal, grammar-based criteria, they are designed for a human annotator. For example, the error taxonomy in Rosen et al. (2014) includes errors of two types: (i) non-words, i.e. word forms which are incorrect with respect to literary Czech in any context, and (ii) real-word errors, i.e. word forms identifiable as incorrect only in a specific context. Both types are subdivided into more detailed categories and subcategories; see Table 1 and Table 2.[2] Only the categories printed in boldface can be detected automatically, but the system still assumes that a target hypothesis or a more general category (complex verb) is established manually.

incorrect form	inflection
	stem
	other
foreign word, coinage	coined Czech word
	foreign word
	inflected foreign word
word boundary	split prefix, joined preposition
	wrongly split/joined compound
	other

Table 1: A taxonomy of non-words

agreement	complex verb:	negation	lexis, idiom
government	• **analytical**	**redundant word**	misused grammar category
pronominal reference	• **modal verb**	**missing word**	incurred error
reflexive form	• **copula**	**word order**	word salad

Table 2: A taxonomy of real-word errors

Despite the daunting complexity of assigning such categories by an automatic tool, automatic annotation of learner texts is still a realistic task. In addition to the option of using standard methods and tools designed for native language, applications developed specifically to process learner language are now also available, including intelligent tutoring systems (e.g. Dickinson and Herring 2008; Levy et al. 2014),

[2] An additional error category concerns inappropriate register and style. In comparison to other taxonomies the sample taxonomy may seem rather coarse-grained. However, it does not need to specify details about the individual forms because it assumes morphosyntactic annotation of the text.

automated scoring in language testing (http://www.ets.org, Shermis and Burstein 2013), annotation of learner texts: both linguistic annotation (Dickinson and Ragheb 2009; Nagata et al. 2011; Krivanek and Meurers 2014) and error annotation (Leacock et al. 2010, 2014; Díaz-Negrillo et al. 2013; Gamon et al. 2013a,b; Ng et al. 2013, 2014; Junczys-Dowmunt and Grundkiewicz 2014). However, the efforts listed above are focused on English and error annotation is limited to error correction.[3]

3 Learner Corpora of Czech

Czech is one of the three languages of Merlin, a learner corpus of Czech, German, and Italian (http://www.merlin-platform.eu, Boyd et al. 2014). The main goal of the 2012–2014 project was to build a platform matching the standard proficiency levels of CEFR (Common European Framework of Reference) with language phenomena specific to the level. The corpus includes texts consisting of about 80 thousand word tokens at CEFR levels A1–C1. It is tagged, parsed, on-line searchable and includes rich metadata.

AKCES, the Acquisition Corpora of Czech (http://akces.ff.cuni.cz, Šebesta 2012), is an umbrella project aimed at building written and spoken language resources about the acquisition Czech by both non-native and native learners. The project also maps the Roma ethnolect of Czech (Eckert 2015). Table 3 shows the currently available AKCES corpora.

		Searchable[4]	**Downloadable**[5]	**# tokens**	**Note**
Native	**Written**	SKRIPT 2012	AKCES 1	0.7M	school essays, age 11–19
	Spoken	SCHOLA 2010	AKCES 2	1.0M	transcripts, class interactions, age 6–19
Non-native		CzeSL-plain	AKCES 3	2.3M	essays, also Roma ethnolect, age 9–76, also non-native bachelor theses
		CzeSL-SGT	AKCES 5	1.1M	automatic annotation
		CzeSL-MAN[6]		0.3M	subset of CzeSL-plain, manual annotation
Roma	**Written**		AKCES 4	0.3M	subset of CzeSL-plain (rom)
	Spoken	ROMi 1.0		1.5M	audio and transcripts, various environments, age 12–28

Table 3: Available AKCES corpora

In the following, we focus on the non-native texts of AKCES, i.e., on its part called CzeSL (Czech as a Second Language).[7] CzeSL is a collection of transcribed

[3] Errors in the training data for the 2014 CONLL Shared Task were classified into 28 types (Ng et al. 2014), but the task was to correct the text, not to assign error labels.

[4] From http://kontext.korpus.cz, the corpus search interface of the Czech National Corpus.

[5] From http://lindat.mff.cuni.cz, the LINDAT/CLARIN repository, license Creative Commons BY-(NC-)ND 3.0.

[6] From http://chomsky.ruk.cuni.cz:5125 (beta version).

[7] For historical and technical reasons, the CzeSL-plain and CzeSL-man corpora also include the Roma ethnolect, and CzeSL-plain an additional part consisting of Bachelor theses authored by non-native students.

essays, hand-written by students of Czech at various occasions as a part of the learning process. For a basic overview of the scope of CzeSL see Table 4. Most texts are equipped with metadata about the author and the task.[8] The first languages (L1) of the learners are varied – most of them belong to the Slavic group (65%, mainly Russian, Ukrainian and Polish), followed by non-Indo-European languages (20%, mainly Vietnamese, Chinese and Arabic). Other Indo-European languages (German, English, French) constitute about 10%. The distribution of texts according to CEFR and the L1 groups is shown in Table 5.

Number of texts	8.6K
Number of sentences	111K
Number of words	958K
Number of tokens	1,148K
Number of authors	1,965
Number of native languages	54
Proficiency levels	A1–C2
Age of the authors	9–76
Share of women/men (in the number of words)	5/3 KW
Number of words per text	100–200

Table 4: The CzeSL corpus – sizes and proportions

	Slavic	Indo-European	Non-Indo-European	Unknown	Total
A1	1783	199	622	5	2609
A1+	283	21	11	0	315
A2	1348	269	480	1	2098
A2+	403	54	113	0	570
B1	929	195	357	0	1481
B2	523	115	107	0	745
C1	82	17	24	0	123
C2	0	1	0	0	1
Unknown	291	27	33	324	675
Total	5642	898	1747	330	8617

Table 5: The CzeSL corpus – number of texts by language groups and CEFR levels

The texts are anonymized by replacing personal names with appropriate forms of *Adam* and *Eva*. Names of smaller places (streets, villages, small towns) and other potentially sensitive data are replaced by QQQ. Unreadable characters or words are transcribed as XXX. For more details about the CzeSL corpus see http://utkl.ff.cuni.cz/learncorp/, or, e.g., Štindlová et al. (2013), Rosen et al. (2014), Meurers (2015).

[8] Full metadata are currently available only in the CzeSL-SGT corpus. See http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta_en.html for the complete statistics.

4 An automatically annotated learner corpus – CzeSL-SGT

The CzeSL-SGT corpus (Czech as a Second Language with Spelling, Grammar and Tags) is coextensive with the strictly non-native part of CzeSL. Texts from the “foreign” part of CzeSL-plain (ciz), collected in 2009–2011, are extended by texts collected in 2013. The transcription markup, encoding some properties of the original manuscripts and preserved e.g. in CzeSL-MAN, is discarded. Instead, the final edits of the author are respected.

4.1 Metadata

Most texts are equipped with metadata about the author and the text, available in Czech and English. The Czech National Corpus site offers the Czech version, while the LINDAT data repository offers the entire corpus using their English version. There are 15 items about the author, such as sex, age, L1, CEFR level of proficiency in Czech, duration and method of study, length of stay in the Czech Republic or knowledge of Czech among family member. Additional 15 items concern the task and the text, such as date, time limit, word count, topic, genre, dictionary/textbook allowed or whether it is a part of an exam.[9] Most authors (79%) have written more than one text. Some or even all items may be missing for some texts: identification of the author is present in 96.7% texts, the first language in 96.3% texts.

4.2 Annotation

If a word form in the original input text is recognized by a standard morphological analyzer (Hajič 2004), it is tagged by word class, morphological categories and base forms (lemmas). We use *Morče*, a standard Czech tagger (Votrubec 2005, 2006), trained on native language (the Prague Dependency Treebank, see Hajič 1998). Its success rate varies by text and deteriorates with the amount of deviations from standard Czech (its reported results on native text are 95–96%). For native texts in the Czech National Corpus, the tagger is combined with a rule-based module (Petkevič 2006), but experiments have shown that for non-native texts the rules, assuming correct grammatical structures, increase the error rate.

In parallel to the tagging task, the input text is corrected by *Korektor*, a spelling and grammar checker, combining rule-based morphology with stochastic language and error models (Richter 2010; Richter et al. 2012). For annotating the current version of CzeSL-SGT, the language model was trained on a corpus of native texts collected from the web and the error model on a small custom-built corpus.[10] The tool corrects not only unrecognized word forms (non-words) but also some forms which are incorrect within a given context (real-word errors). From the resulting n-best ranked suggestions with a correction type (spelling or grammar) only the first option is used. However, the present implementation of *Korektor* cannot insert or delete

[9] For a more technical description of the corpus see http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-en.pdf For a list of all attributes and values in Czech and English see http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html. The numbers of documents, listed according to specific attribute values, are given here: http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta_en.html.

[10] Ramasamy et al. (2015) report better results with language models trained on the SYN2005 corpus.

word boundaries (split or join word forms), which is one of the more frequent error types in learner texts.

The corrected text is tagged and lemmatized again. Original and corrected forms are compared and error labels, based on applicable formal criteria, are assigned (Jelínek et al. 2012). In the resulting annotation each token is labelled by the following attributes:

- *word* – original word form
- *lemma* – lemma of *word*; same as *word* if the form is not recognized
- *tag* – morphological tag of *word*; if the form is not recognized: X@-------------
- *word1* – corrected form; same as *word* if determined as correct
- *lemma1* – lemma of *word1*
- *tag1* – morphological tag of *word1*
- *gs* – information on whether the error was determined as a spelling (S) or grammar (G) error; *word* is mostly recognized for grammar errors
- *err* – error type, determined by comparing word and word1 http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1_en.html

Example (1) shows 4 spelling errors in a single sentence. Incorrect forms are in boldface, the second line is the sentence as corrected by *Korektor*. All the ill-formed words are non-words.

(1) ***Tén*** *pes* ***míluje*** ***svécho*** ***kamarada*** – *člověka.*
Ten pes miluje svého kamaráda – člověka.
that dog loves REFL.POSS friend man
'That dog loves his friend – the man.'

Table 6 shows the attribute values for the annotated sentence in the corpus. The three columns headed by the attributes *word*, *lemma* and *tag* concern the original, uncorrected text. An incorrect form is labelled by the morphological analyser bundled with the tagger as unknown (X@), while its *lemma* is identical to *word*.[11] The next triple *word1*, *lemma1* and *tag1* shows its automatically corrected version. *Korektor* specifies the incorrect forms in the *gs* column as spelling errors (S). The analyser and *Korektor* do not always agree about a specific form as a non-word. A more sophisticated word form recognition is currently available in the analyser, so it is safer to trust the tagger's verdict.

word	*lemma*	*tag*	*word1*	*lemma1*	*tag1*	*gs*	*err*
Tén	Tén	X@	*Ten*	ten	PDYS1	S	Quant1
pes	pes	NNMS1	*pes*	pes	NNMS1		
míluje	míluje	X@	*miluje*	milovat	VB-S---3P	S	Quant1
svécho	svécho	X@	*svého*	svůj	P8MS4	S	Voiced
kamarada	kamarada	X@	*kamaráda*	kamarád	NNMS4	S	Quant0
-	-	Z:	-	-	Z:		
člověka	člověk	NNMS2	*člověka*	člověk	NNMS4		
.	.	Z:	.	.	Z:		

[11] Irrelevant suffixes of the positional tags are omitted for space reasons. For a description of the tagset see http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html.

Table 6: Annotation of a sample sentence (1) including spelling errors

Example (2) includes a non-word *nejakij* and a real-word error *postele*. According to the rules used for the manual annotation of CzeSL-MAN, *nejakij* should be corrected to *nějaký*, a form of Literary Czech, rather than to *nějakej*, a Colloquial Czech form preferred by *Korektor* in Table 7 due to the smaller edit distance between the non-word and the corrected form. The form *postele* could be correct in a different context, but it is incorrect within the adverbial of location, where the form *posteli* in the local case is required.

(2) ***Nejakij*** *muž* *spí* *v* ***postele***.
*some man sleeps in bed$_{\text{GEN.SG/NOM.PL/}\mathbf{ACC.PL}\text{/VOC.PL}}$ □
Nějaký muž spí v posteli.
some man sleeps in bed$_{\mathbf{LOC.SG}}$
‘Some guy is sleeping in the bed.’

In the corpus, *nějakej* (correction of the non-word *nejakij*) is correctly tagged as a colloquial form (“6” at position 7 of *tag1*, see Table 7). For the real-word error *postele* the tagger chooses the implausible directional interpretation of the adverbial, where *postele* is accusative plural (“P4” at positions 4 and 5 of *tag*) and the preposition *v* takes an accusative complement (“4” at position 5): ‘Some guy is sleeping into the beds.’ The corrected form and the preposition are tagged correctly (singular, local case “S6”). *Korektor* specifies the error as grammatical “G”, while the error label assigner merely says that there is an error in a single character “SingCh”.

word	*lemma*	*tag*	*word1*	*lemma1*	*tag1*	*gs*	*err*
Nejakij	Nejakij	X@	Nějakej	nějaký	PZYS1-6	S	Caron0
muž	muž	NNMS1	muž	muž	NNMS1		
spí	spát	VB-S---3P	spí	spát	VB-S---3P		
v	v	RR--4	v	v	RR--6		
postele	postel	NNFP4	posteli	postel	NNFS6	G	SingCh
.	.	Z:	.	.	Z:		

Table 7: Annotation of a sample sentence (2) including a real-word error

Table 8 shows how many spelling and grammar errors are corrected in the corpus (depending on the G or S value of *gs*, assigned by *Korektor*) and how many word forms are (un)recognized (depending on *tag* assigned by the tagger). The number of grammar (real-word) errors is relatively high (17.8% of the total number of corrected errors, even if only grammar errors in forms recognized by the tagger are counted). However, the success rate of correcting grammar errors is lower than for spelling errors.

Error type	Frequency	% of total tokens	% of corrected forms
Spelling errors	118,488	10.33%	77.97%
Grammar errors	33,474	2.92%	22.03%
Errors total (grammar and spelling)	151,962	13.24%	100.00%
Spelling errors in unrecognized forms	94,878	8.37%	62.44%
Grammar errors in recognized forms	27,055	2.36%	17.80%

Unrecognized forms total	104,523	9.11%

Table 8: Spelling and grammar errors corrected by *Korektor*;
word forms (un)recognized by the tagger

Table 9 shows a sample of 50 error labels used in the corpus. The labels are assigned by rules comparing the original and the corrected string. Some of them have a strong linguistic basis, other labels are more formal or used as wastebasket categories.

Error type	Error description	Example
Cap0	capitalization: incorrect lower case	*evropě → Evropě*; *štědrý → Štědrý*
Cap1	capitalization: incorrect upper case	*Staré → staré*; *Rodině → rodině*
Voiced0	voicing assimilation: incorrect voiced	*stratíme → ztratíme*; *nabítku → nabídku*
Voiced1	voicing assimilation: incorrect voiceless	*zbalit → sbalit*; *nigdo → nikdo*
VoicedFin0	word-final voicing: incorrect voiceless	*kdyš → když*; *vztach → vztah*
VoicedFin1	word-final voicing: incorrect voiced	*přez → pres*; *pag → pak*
Voiced	voicing: other errors	*protoše → protože*; *hodili → chodili*
Palat0	missing palatalization (*k,g,h,ch*)	*amerike → Americe*; *matke → matce*
Je0	*je/e*: incorrect *e*	*ubjehlo → uběhlo*; *Nejvjetší → Největší*
Je1	*je/e*: incorrect *je*	*vjeděl → věděl*; *vjeci → věci*
Mne0	*me/mne*: incorrect *m*	*zapoměla → zapomněla*
Mne1	*me/mne*: incorrect *mne, mne, mne*	*mněla → měla*; *rozumněli → rozuměli*
ProtJ0	protethic *j*: missing *j*	*sem → jsem*; *menoval → jmenoval*
ProtJ1	protethic *j*: extra *j*	*jse → se*; *jmé → mé*
ProtV1	protethic *v*: extra *v*	*vosm → osm*; *vopravdu → opravdu*
EpentE0	*e* epenthesis: missing *e*	*domček → domeček*
EpentE1	*e* epenthesis: extra *e*	*rozeběhl → rozběhl*; *účety → účty*

Table 9: Selected formal errors in Czesl-SGT

Table 10 lists the top 12 most frequent error labels in the corpus. Note that errors in diacritics are by far the most common. The notorious spelling problem of Czech native speakers – the uncertainty about the use of *i* and *y* – ranks much lower.

Error type	Error description	Example	Freq	%
Quant0	error in diacritics: missing vowel accent	*vzpominám → vzpomínám*; *doufam → doufám*	67181	41.61
SingCh	a single wrong character	*otevřila → otevřela*; *vezmíme → vezmeme*;	25451	15.76
Quant1	error in diacritics: extra vowel accent	*která → která*; *hledát → hledat*	17710	10.97
Caron0	error in diacritics: missing caron	*vecí → věcí*; *sobe → sobě*	13893	8.61
Cap1	capitalization: incorrect upper case	*Staré → staré*; *Rodině → rodině*	11847	7.34
RedunChar	other single extra character	*opratrně → opatrně*; *zrdcátko → zrcátko*	3157	1.96
Caron1	error in diacritics: extra caron	*břečel → brečel*; *bratřem → bratrem*	2661	1.65
Unspec	error in the middle of the word	*provudkyně → průvodkyně*; *kreréնu → kterému*	2504	1.55

Y0	*i* instead of correct *y*	*pražskích → pražských*; *vipije → vypije*	2384	1.48
Y1	*y* instead of correct *i*	*hlavným → hlavním*; *líbyl → íbil*	2179	1.35
MissChar	missing character	*zaímavou → zajímavou*; *bohaství → bohatství*	1805	1.12
Voiced	voicing: other errors	*pěžky → pěšky*; *hodili → chodili*	1783	1.10

Table 10: The 12 most frequent error types detected in CzeSL-SGT

As Table 11 shows, broader error categories are represented in CzeSL-SGT in proportions similar to those in hand-annotated CzeSL-MAN. This is a comforting result – there is no evaluation of the error labels assignment at the moment. The differences in some categories (omission) may also be due to the heterogeneity of texts in CzeSL-MAN, namely to the high share of Roma ethnolect texts.

General error type	CzeSL-SGT	CzeSL-MAN
Insertion	3.76	3.52
Omission	1.39	9.20
Substitution	31.30	37.67
Transposition	0.16	0.19
Missing diacritic	50.19	40.40
Addition of diacritic	12.69	8.60
Wrong diacritic	0.51	0.43

Table 11: Percentages of error types detected automatically in CzeSL-SGT and manually in Czesl-MAN

In addition to the attributes listed above, the search interface of the Czech National Corpus offers ‘dynamic’ attributes, derived from some positions of *tag* and *tag1*. They can be used in queries to specify values of morphological categories without regular expressions, to stipulate identity of these values in two or more forms to require grammatical concord or to compare values of a category for *word* and *word1*. These attributes are available for the following categories of the original and the corrected form:

- *k*, *k1* – word class (position 1 of the tag)
- *s*, *s1* – detailed word class (position 2 of the tag)
- *g*, *g1* – gender (position 3 of the tag)
- *n*, *n1* – number (position 4 of the tag)
- *c*, *c1* – case (position 5 of the tag)
- *p*, *p1* – person (position 8 of the tag)

4.3 Using the corpus

The corpus can be searched from the unified search interface of the Czech National Corpus (https://kontext.korpus.cz). CzeSL-SGT is one of “Synchronic written corpora”, in the category “specialized”. With the “Query Type” set to “Basic” and no other specifications, a string entered in the “Query” field returns sentences where the form or lemma occurs in the original, uncorrected text. For more advanced queries, including references to tags, lemmas, error types, corrected forms and metalanguage attributes, the "Query Type" should be set to "CQL" and/or the settings in "Specify query according to the meta-information" modified.[12]

In addition to query types available in other types of corpora, dynamic attributes support some other interesting options. A CQL query in (3) returns nouns, adjectives and pronouns recognized as such in the original, detected as grammatically incorrect, preserving the word class in the corrected form but having a different case.

(3) 1:[k="[NAP]" & gs="G"] & 1.k=1.k1 & 1.c!=1.c1

The corpus is also available for download from the LINDAT data repository (http://hdl.handle.net/11234/1-162. The corpus is currently in release 2. Some bugs present in the original release have been fixed and the whole corpus is now a single XML document with each text as a "div" element. See Figure 1 for an extract from a sample text with the annotation, including metadata in the header.[13]

```
<div t_id="UJA2_PH_003" t_date="2010-12-21" t_medium="manuscript" t_limit_minutes="45" t_aid="none" t_exam="yes|interim"
t_limit_words="25" t_title="E-mail kamarádce/kamarádovi" t_topic_type="general" t_activity="" t_topic_assigned="specified"
t_genre_assigned="specified" t_genre_predominant="informative" t_words_count="30" t_words_range="-50" s_id="UJA2_PH" s_sex="m"
s_age="17" s_age_cat="16" s_L1="vi" s_L1_group="nIE" s_other_langs="" s_cz_CEF="A1" s_cz_in_family="" s_years_in_CzR=""
s_study_cz="university" s_study_cz_months="" s_study_cz_hrs_week="15" s_textbook="NCSS" s_bilingual="no">
<s id="1">
<word lemma="mít" tag="VB-S---1P-AA" word1="mám" lemma1="mít" tag1="VB-S---1P-AA" gs="" err="">mám</word>
<word lemma="dobře" tag="Dg-------1A" word1="dobře" lemma1="dobře" tag1="Dg-------1A" gs="" err="">dobře</word>
<word lemma="." tag="Z:" word1="." lemma1="." tag1="Z:" gs="" err="">.</word>
</s>
<s id="2">
<word lemma="v" tag="RR--4" word1="V" lemma1="v" tag1="RR--4" gs="" err="">V</word>
<word lemma="neděle" tag="NNFS4" word1="neděli" lemma1="neděle" tag1="NNFS4" gs="" err="">neděli</word>
<word lemma="dival" tag="X@" word1="díval" lemma1="dívat" tag1="VpYS---XR-AA" gs="S" err="Quant0">dival</word>
<word lemma="být" tag="VB-S---1P-AA" word1="jsem" lemma1="být" tag1="VB-S---1P-AA" gs="" err="">jsem</word>
<word lemma="se" tag="P7-X4" word1="se" lemma1="se" tag1="P7-X4" gs="" err="">se</word>
<word lemma="na" tag="RR--6" word1="na" lemma1="na" tag1="RR--6" gs="" err="">na</word>
<word lemma="televize" tag="NNFS6" word1="televizi" lemma1="televize" tag1="NNFS6" gs="" err="">televizi</word>
<word lemma="a" tag="J^" word1="a" lemma1="a" tag1="J^" gs="" err="">a</word>
<word lemma="uklizěl" tag="X@" word1="uklízel" lemma1="uklízet" tag1="VpYS---XR-AA" gs="S" err="Quant0|Caron1">uklizěl</word>
<word lemma="být" tag="VB-S---1P-AA" word1="jsem" lemma1="být" tag1="VB-S---1P-AA" gs="" err="">jsem</word>
<word lemma="." tag="Z:" word1="." lemma1="." tag1="Z:" gs="" err="">.</word>
</s>
[...]
</div>
```

Figure 1: A sample annotated text in the XML format

5 Evaluating the automatic annotation

The error annotation can be evaluated using the CzeSL-MAN, the existing manually annotated subset of the corpus – the manual annotation includes one or two target hypothesis about an incorrect form and one or more error labels. So far, only the proposed corrected forms were evaluated.

[12] For general help on using CQL see http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying.

[13] The metadata attributes about the text are prefixed by "t_", while those about the student by "s_". In the annotation of "word" elements, insignificant tag suffixes are not shown for space reasons.

In Rosen et al. (2014) we report on results that *Korektor* achieved in an experiment based on a pilot corpus consisting of 67 CzeSL-MAN texts (9.4K tokens), including 786 unrecognized tokens, where two annotators agreed on the same corrected form. The language and the error models, trained on native texts, were the same as those used for annotating the present version of CzeSL-SGT.

The comparison of *Korektor*'s output with either of the two annotation levels of CzeSL-MAN is not quite fair: only non-words are corrected at level 1, while level 2 includes errors in syntax, word order and style, mostly well beyond the current reach of *Korektor*. Still, for level 1 precision was 74% and recall 71%. For level 2, the precision dropped to 60% and recall to 45%. These results were considered sufficiently high to justify the use of *Korektor* in the annotation of CzeSL-SGT.

Ramasamy et al. (2015) experiment with different setups of language and error models. The best results were comparable or better – see Table 12 ("Pilot corpus" for the previous results, "CzeSL-MAN" for the new results).[14] They were achieved by using models trained on native texts for the entire CzeSL-MAN test set. The authors report in detail on an easier task of error detection: in a sample of 3K most frequent tokens identified by an annotator as incorrect, more than 89% non-words (form errors) were detected. On the other hand, the result for real-word (grammar) errors was only 15.5%. Interestingly, for the combination of the two error types (as in **zajímavy → zajímavý → zajímavé* 'interesting'), the best detection result was also 89%.

	Pilot corpus		CzeSL-MAN	
	Level 1	Level 2	Level 1	Level 2
Precision	74%	60%	73%	78%
Recall	71%	45%	80%	62%

Table 12: Evaluation of the automatic error correction

6 Discussion and perspectives

A reliable correction tool is the key to a successful automatic error annotation. There are at least two obvious paths to a more successful result: (i) better training data especially for the error model, which should consist only of Czech texts produced by foreigners and thus be more in line with the content of CzeSL-SGT, and (ii) extending the tool to handle errors spanning word boundaries, including splitting/joining and word order errors. |Other options include parameterizing the model according to a specific type of learner Czech (by the first language or proficiency level), or experimenting with the tool design, perhaps in combination with a machine translation approach.

The absence of automatic methods and tools targeting non-native language is not caused only by the computational complexity of the task and the absence of data resources, e.g. for machine learning applications. There is a more fundamental issue of largely missing concepts and schemes to describe non-standard linguistic phenomena. As a separate research track, we develop categories for annotating non-standard word forms, which can replace tagging schemes used for standard language.

[14] Comparison of the two experiments should be taken with a grain of salt due to different methodology.

We are aware that some aspects of manual annotation of non-standard language cannot be substituted by an algorithm or even by a stochastic model. However, the fact that CzeSL-SGT is one of the most popular downloads from the LINDAT/CLARIN repository, together with a growing list of references to CzeSL-MAN or CzeSL-SGT (Aharodnik et al. 2013; Hudousková 2013, 2014; Štindlová 2015; Meurers 2015) may suggest that (semi-)automatic annotation is a useful help.

References

Aharodnik, K., M. Chang, A. Feldman & J. Hana. 2013. Automatic identification of learners' language background based on their writing in Czech. *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNCLP 2013)*, 1428–1436. Nagoya. https://msuweb.montclair.edu/~feldmana/publications/I13-1200.pdf.

Boyd, A. et al. 2014. The MERLIN Corpus: learner language and the CEFR. In Calzolari, N. et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/606_Paper.pdf.

Dagneaux, E. et al. 2008. *The Louvain Error Tagging Manual, Version 1.3*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Díaz-Negrillo, A., N. Ballier & P. Thompson (eds.). 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. Vol. 59. (Studies in Corpus Linguistics). Amsterdam and Philadelphia: John Benjamins.

Dickinson, M. & J. Herring. 2008. Developing online ICALL exercises for Russian. *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*, 1–9. Columbus. http://cl.indiana.edu/ md7/papers/dickinson-herring08.html.

Dickinson, M. & M. Ragheb. 2009. Dependency annotation for learner corpora. *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories TLT8*. http://cl.indiana.edu/ md7/papers/dickinson-ragheb09.pdf.

Eckert, E. 2015. *Romani in the Czech Sociolinguistic Space*. Anglo-American University Prague. http://www.aauni.edu/wp-content/uploads/2015/04/Eckert-finallc.pdf.

Granger, S. et al. 2002. *Error Tagging Manual for L2 French*. Louvain-la-Neuve: Université catholique de Louvain, Centre for English Corpus Linguistics.

Hajič, J. 1998. The Prague Dependency Treebank. In Hajičová, E. (ed.), *Issues of Valency and Meaning – Studies in Honour of Jarmila Panevová*, 106–132. Praha: Karolinum, Charles University Press.

Hajič, J. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Praha: Karolinum, Charles University Press.

Hudousková, A. 2013. The corpus CzeSL in the service of teaching Czech for foreigners – errors in the use of the pronoun *který*. In Gajdošová, K. & A. Žáková (eds.), *Proceedings of the Seventh International Conference Slovko 2013*. Lüdenscheid, Germany: RAM-Verlag.

Hudousková, A. 2014. Jmenné koncovky v češtině pro cizince – distribuce, frekvence a fonetika. První sonda. In Petkevič, V., A. Adamovičová & V. Cvrček (eds.), *Radost z jazyků. Sborník k 75. narozeninám prof. Františka Čermáka*, vol. 20.

(Studie z korpusové lingvistiky), 215–230. Praha: Nakladatelství Lidové noviny.

Jelínek, T., B. Štindlová, A. Rosen & J. Hana. 2012. Combining manual and automatic annotation of a learner Corpus. In Sojka, P., A. Horák, I. Kopeček & K. Pala (eds.), *Text, Speech and Dialogue – Proceedings of the 15th International Conference TSD 2012*. (Lecture Notes in Computer Science 7499), 127–134. Springer. http://utkl.ff.cuni.cz/ rosen/public/2012-czesl-tsd_prefinal.pdf.

Junczys-Dowmunt, M. & R. Grundkiewicz. 2014. The AMU System in the CoNLL-2014 Shared Task: grammatical error correction by data-intensive and feature-rich statistical machine translation. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 25–33. Baltimore, Maryland: Association for Computational Linguistics. http://acl2014.org/acl2014/W14-17/pdf/W14-1703.pdf.

Krivanek, J. & D. Meurers. 2014. Comparing rule-based and data-driven dependency parsing of learner language. In Kim Gerdes, E. H. & L. Wanner (eds.), *Dependency Theory*. (Frontiers in AI and Applications). Amsterdam: IOS Press.

Levy, M., F. Blin, C. B. Siskin & O. Takeuchi (eds.). 2014. *WorldCALL – International Perspectives on Computer-Assisted Language Learning*. (Routledge Studies in Computer Assisted Language Learning). Routledge.

Meurers, D. 2015. Learner corpora and natural language processing. In Sylviane Granger, G. G. & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press. http://purl.org/dm/papers/meurers-15.html.

Nagata, R., E. Whittaker & V. Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. (HLT '11) , 1210–1219. Portland, Oregon: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002472.2002625.

Ng, H. T. et al. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 1–14. Baltimore, Maryland: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W14/W14-1701.

Ng, H. T., S. M. Wu, Y. Wu, C. Hadiwinoto & J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, 1–12. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-3601.

Petkevič, V. 2006. Reliable morphological disambiguation of Czech: rule-based approach is necessary. In Šimková, M. (ed.), *Insight into the Slovak and Czech Corpus Linguistics*, 26–44. Bratislava: Veda (Publishing House of the Slovak Academy of Sciences & Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences).

Ramasamy, L., A. Rosen & P. Straňák. 2015. Improvements to Korektor: A case study with native and non-native Czech. *ITAT (Information technologies – Applications and Theory).*

Richter, M. 2010. An Advanced Spell Checker of Czech. Faculty of Mathematics and Physics, Charles University, Prague. https://redmine.ms.mff.cuni.cz/attachments/2/richter-diploma-thesis.pdf.

Richter, M., P. Straňák & A. Rosen. 2012. Korektor – a system for contextual spell-checking and diacritics completion. *Proceedings of COLING 2012*, 1019–1028. Mumbai, India: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-2099.

Rosen, A., J. Hana, B. Štindlová & A. Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation – Special Issue: Resources for language learning* 48.1, 65–92. doi:http://dx.doi.org/10.1007/s10579-013-9226-3. http://utkl.ff.cuni.cz/rosen/public/2011-czesl-lrej_prefinal.pdf.

Šebesta, K. 2012. Learner corpora and the Czech language. In Semrádová, I. (ed.), *Intercultural Inspirations for Language Education. Spaces for understanding*, 74–89. Univerzita Hradec Králové.

Shermis, M. D. & J. Burstein (eds.). 2013. *Handbook of Automated Essay Evaluation – Current Applications and New Directions*. Routledge.

Štindlová, B. 2013. *Žákovský korpus češtiny a evaluace jeho chybové anotace*. Praha: Univerzita Karlova v Praze, Filozofická fakulta.

Štindlová, B. 2015. K parcelaci gramatiky češtiny pro nerodilé mluvčí. In Švrčinová, M. & Z. Vlasáková (eds.), *Gramatika ve výuce a testování cizích jazyků* (*včetně češtiny pro cizince*), 198–209. Praha: Ústav jazykové a odborné přípravy UK.

Štindlová, B., S. Škodová, J. Hana & A. Rosen. 2013. A learner corpus of Czech: current state and future directions. In Granger, S., G. Gilquin & F. Meunier (eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. (Corpora and Language in Use – Proceedings 1). Louvain-la-Neuve: Presses Universitaires de Louvain. http://utkl.ff.cuni.cz/rosen/public/LCR2011_proceedings_Stindlova-et-al_prefinal.pdf.

Votrubec, J. 2006. Morphological tagging based on averaged perceptron. *WDS'06 Proceedings of Contributed Papers*, 191–195.