

# CzeSL – an error tagged corpus of Czech as a second language

*Barbora Štindlová, Alexandr Rosen, Jirka Hana and Svatava Škodová*

**Abstract:** Using an error-annotated learner corpus as the basis, the goal of this paper is two-fold: (i) to evaluate the practicality of the annotation scheme by computing inter-annotator agreement on a non-trivial sample of data, and (ii) to find out whether the application of automated linguistic annotation tools (taggers, spell checkers and grammar checkers) on the learner text is viable as a substitute for manual annotation.

**Keywords:** learner corpus, error annotation, second language acquisition

## 1. Introduction

Texts produced by non-native speakers are a precious source of information about the acquisition of a language by the learners and about second language acquisition in general. Collections of such texts – learner corpora – can be annotated in a way similar to other corpora with morphosyntactic categories or syntactic structure. However, their most interesting aspect is examples of deviant use, which can be corrected and assigned a tag specifying the type of error. Annotation of this kind is a challenging task, even more so for a language such as Czech, with its rich inflection, derivation, agreement, and a largely information-structure-driven constituent order.

The present work is based on a project aimed at building a learner corpus with errors manually corrected and labelled within a three-level annotation scheme. Manual annotation is supplemented by morphosyntactic tags assigned to the hand-corrected input by a tagger, and by additional error tags, whenever they can be derived automatically. Options to provide corrections and error annotations in a fully automatic way are being investigated.

The paper is organized as follows: first, Section 2 presents our learner corpus project in more detail; then Section 3 describes our annotation scheme and its evaluation by computing inter-annotator agreement using data from a trial annotation; finally, Section 4 presents the results of experiments in applying two taggers and a spell checker to the uncorrected text.

## 2. Learner corpus of Czech (CzeSL)

The learner corpus of Czech as a Second Language (CzeSL)<sup>1</sup> is built as a part of the Acquisition Corpora of Czech (AKCES), a research programme pursued at

---

1 The corpus is one of the tasks of the project *Innovation of Education in the Field of Czech as a Second Language* (project no. CZ.1.07/2.2.00/07.0259), a part of the operational programme *Education for Competiveness*, funded by the European Structural Funds

Charles University in Prague since 2005 (Šebesta 2010). Intended to reach the size of up to two million running words, the corpus is focused on the language of non-native speakers of Czech, consisting of three main groups: (1) speakers of Slavic languages, (2) speakers of distant non-Indo-European languages, (3) speakers of other Indo-European languages. A separate subcorpus, built along slightly different guidelines, covers the language of Czech pupils with Romani background (Bedřichová et al. 2011).

Although written texts prevail, each subcorpus has its oral part. A large portion of the written parts consists of short essays, collected as manuscripts and transcribed into an electronic format, preserving specific features of handwritten texts, such as self-corrections or emoticons (Štindlová 2011, 106). Manuscripts are used for their availability and also because the authors cannot check them easily by automatic proofreading tools. The rest of the written texts are Bachelors' and Masters' theses, written in Czech by non-native students.

The data being collected cover all language levels according to the Common European Framework of Reference for Languages (CEFR), from real beginners (A1 level) to advanced learners (level B2 and higher), with a balanced mix of levels as much as possible. While most other learner corpora include texts elicited only as a part of written or oral examination, we use texts produced during all range of situations throughout the language-learning process, such as homework and texts produced during the class.

Each text is equipped with metadata records, some of them relate to the respondent (including sociological data about the learner, such as age, gender, first language, proficiency level in Czech, knowledge of other languages, duration and conditions of language acquisition), while others specify the character of the text and circumstances of its production (availability of reference tools, type of elicitation, temporal and size restrictions etc.).

The finished corpus will be used in the education of teachers of Czech as a foreign language, as a source of knowledge about interlanguage and second language acquisition, and ultimately as a source of data for the compilation of teaching materials and optimization of the learning process (Štindlová 2011).

### 3. The annotation scheme

The annotation scheme for CzeSL was designed to reflect the goals of the project and the specifics of Czech.<sup>2</sup> Rather than focusing on a narrow domain of learner language as the annotation target (such as spelling or lexical errors), the

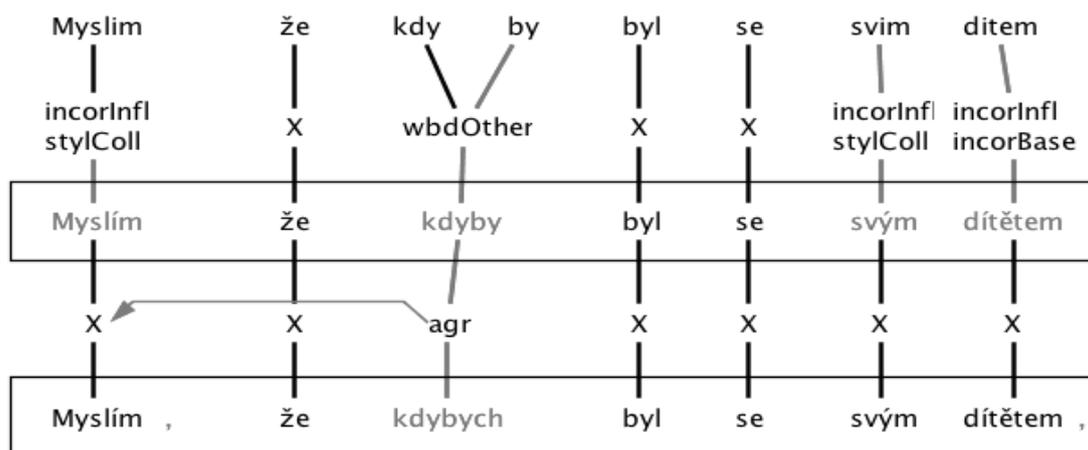
---

(ESF) and the Czech government. The annotation tool was also partially funded by grant no. P406/10/P328 of the Grant Agency of the Czech Republic. The project team consists of partners from the Technical University in Liberec and Charles University in Prague.

2 See Hana et al. (2010) and Štindlová et al. (in press) for a more detailed description of the annotation scheme.

corpus is intended as open to as many research goals as possible using a multi-level annotation scheme, supporting successive emendation. As a compromise between several theoretically motivated levels and practical concerns about the process of annotation, the scheme offers two levels of annotation. This enables the annotators to register anomalies in isolated forms separately from the annotation of context-based phenomena but saves them from difficult theoretical dilemmas.

Levels of annotation are represented as a graph consisting of a set of inter-linked parallel paths, where a path is a sequence of word forms corresponding to a sentence at a given level – see Fig. 1, glossed in (1). Each word in the input text is represented at every level, unless it is split, joined (as *kdy by* in Fig. 1), deleted or added by the annotator. Whenever a word form is emended, the type of error can be specified as a label at the link connecting the incorrect form at a lower level with its emended form at a higher level (such as *incorInfl* or *incorBase* for morphological errors in inflectional endings and stems, *stylColl* as a stylistic marker, *wbdOther* as a word boundary error, and *agr* as an error in agreement). These labelled relations can inter-link any number of potentially non-contiguous words across the neighbouring levels. Multiple words can thus be identified as a single segment, while any of the participating word forms can retain their 1:1 links with their counterparts at other levels.



**Figure 1.** Example of the three-level error annotation scheme

- (1) *Myslím, že kdybych byl se svým dítětem, ...*  
 think<sub>1stSg</sub> that if<sub>1stSg</sub> was<sub>masc</sub> with my child,  
 ‘I think that if I were with my child, ...’

At Level 0 (L0), the level of the transcribed input, the words represent the original strings of graphemes. At L1, the level of orthographical and morphological emendation, only individual forms are treated. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. A

formally correct form may be corrected if the author clearly misspelled the form she intended to use, creating an unintended homograph. All other types of errors (such as errors in agreement) are emended at L2.

Manual annotation is supported by the purpose-built annotation tool *feat* (<http://ufal.mff.cuni.cz/~hana/feat.html>) and followed by automatic post-processing, providing additional information about lemma, POS and morphological categories for correct and emended forms, with error types not assigned manually (by comparing the original and corrected strings), and with formal error description: type of a spelling alternation, missing/redundant expression, inappropriate word order. In the future, we plan to automatically tag errors in verb prefixes, inflectional endings, spelling, palatalization, metathesis, etc. Table 1 shows the number of different error tags. Options to perform a fully automatic annotation are investigated in Section 4 below.

**Table 1.** Manually and automatically assigned error tags at L1 and L2

Error tags	L1 only	L2 only	L1 and L2	Total
Manual	8	11	3	22
Automatic	1	6	0	7
Total	9	17	3	29

### 3.1. Evaluation of the error mark-up

There is no widely accepted metric evaluating the consistency of annotation of learner corpora. In the current annotation practice of non-native speakers' corpora, it is common to have ill-formed texts tagged by a single annotator, despite problems in reliability and evaluation. A general shift towards multiple annotation of learner corpora is imminent – cf. Tetreault and Chodorow (2008), Rozovskaya and Roth (2010), Meurers (2011).

### 3.2. Inter-annotator agreement (IAA)

The manual annotation of CzeSL was evaluated using the metric  $\kappa$  (kappa, Cohen 1960), widely accepted as the standard measure of inter-annotator agreement, especially for tagged corpora. It is calculated as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where  $P(A)$  is the observed agreement among the annotators, and  $P(E)$  is the expected agreement, i.e.,  $P(E)$  is the probability that the coders agree by chance. The values of  $\kappa$  are within the interval  $[-1, 1]$ , where  $\kappa = 1$  means perfect agreement,  $\kappa = 0$  agreement equal to chance, and  $\kappa = -1$  “perfect” disagreement.

In the evaluation, we use the scale proposed by Rietveld and van Hout (1993), where the values 0.21–0.40 indicate fair agreement, 0.41–0.60 moderate

agreement, and 0.61–0.80 substantial agreement. Values below 0.20 indicate slight agreement, whereas those above 0.81 are almost perfect.

### 3.3. Sample data

The data for the annotation were selected from a database compiled for CzeSL. The sample consists of 74 texts totalling 9848 tokens. Most of them were written by native speakers of Russian; the texts are classified according to the CEFRL scale as A2 or B1.

#### 3.3.1. Method

The sample was annotated by 14 annotators. They were split into two groups: Annotators A and Annotators B. Each group annotated the whole sample independently. On average each annotator processed 1475 words in 11 texts.

On L1, the annotators chose from 8 tags for morphological, orthographical and word-boundary errors, and “non-Czech” expressions. On L2, syntactic, morphosyntactic, lexical and stylistic errors were captured by 15 tags. Stylistically marked expressions could be assigned additional tags at both levels.<sup>3</sup> The results of inter-annotator agreement for the domain categories (*incor*, *wbd*, *fw* and *styl*)<sup>4</sup> were summed up, without distinguishing the particular error subtypes.

#### 3.3.2. Results

Table 2 below summarizes the distribution of selected error tags. The column “Only A” shows counts for each tag used by annotators in group A but not by those in group B. Similarly for the next column. The following column shows cases when both groups agreed. The last column gives the agreement measure  $\kappa$ .

**Table 2.** Inter-annotator agreement on selected tags

Tag	Type of error	Only A	Only B	A & B	$\kappa$
<i>incorSum</i>	<i>incorStem+incorInfl</i>	168	130	894	0.84
<i>incorStem</i>	Incorrect stem	167	165	559	0.61
<i>incorInfl</i>	Incorrect inflection	173	130	250	0.75
<i>wbdSum</i>	Incorrect word boundary	14	21	45	0.72
<i>fwSum</i>	<i>fw+fwFab+fwNc</i>	25	17	18	0.46
<i>fw</i>	“Non-Czech” expression	4	6	1	0.17
<i>fwFab</i>	Author’ coinage	23	13	3	0.14

3 The numbers of tags given here correspond to a slightly outdated taxonomy and differ from the current state, presented in Section 3.

4 The error taxonomy is hierarchical – error types are partitioned into domains, which are further divided into more specific subcategories, tagged manually or automatically. For example, the domain of complex verb form errors on L2 can be further specified as errors in analytical verb forms (*cvf*), modal verbs (*mod*), verbo-nominal predicates, passive or resultative form (*vnp*).

<i>fwNc</i>	Foreign/unidentified form	10	9	3	0.24
<i>stylColl L1</i>	Colloquial style at L1	10	2	2	0.25
<i>agr</i>	Agreement violation	82	99	110	0.54
<i>dep</i>	Errors in expressing syntactic dependency	99	118	87	0.43
<i>neg</i>	Incorrectly expressed negation	11	9	9	0.47
<i>stylColl L2</i>	Colloquial style at L2	14	14	10	0.42
<i>lex</i>	Lexical or phraseology error	107	131	74	0.37
<i>rflx</i>	Incorrect reflexive expression	6	11	3	0.26
<i>use</i>	Improper use of tense, aspect etc.	60	74	19	0.21
<i>vbx</i>	Ill-formed complex verb forms	20	9	3	0.17
<i>ref</i>	Incorrect pronominal references	14	17	3	0.16
<i>sec</i>	Secondary (consequent) error	45	18	4	0.11

The table shows that on L1 the annotators tend to agree in the domain categories *incorSum*, *wbdSum* and *fwSum*, i.e., for incorrect morphology, for improper word boundaries and for “non-Czech” expressions in general ( $\kappa > 0.8$ ,  $\kappa > 0.6$ , and  $\kappa > 0.4$ , respectively). IAA was lower ( $\kappa < 0.4$ ) for categories with a fuzzy interpretation, where a target hypothesis is difficult to establish, such as subcategories of *fw*, used to tag attempts to coin a new Czech lexeme (*fwFab*), or foreign/unidentified strings of words (*fwNc*). Even the choice between the two subcategories was problematic (accounting for 26% of the total number of cases where the two annotators differed in the use of these two tags). This is true especially in cases where an annotator is not able identify the origin of the lexeme.

On L2 the annotators agree ( $\kappa > 0.4$ ) in agreement errors (*agr*) and errors in expressing syntactic dependency (*dep*), and also in the well-defined category of errors in negation (*neg*). However, pronominal references (*ref*), secondary (consequent) errors (*sec*) and surprisingly also analytical verb forms and complex predicates (*vbx*), show a very low level of IAA, even though they are identifiable by formal linguistic criteria. In all these three cases, the distribution of tags and the annotators’ feedback suggest that the annotation manual fails to provide enough guidance in distinguishing between the error types *ref* vs. *agr* and *ref* vs. *dep* (in either case the disagreement represents 19% of all the inconsistent uses of the tag *ref*), and at the same time does not specify the formal aspect of using this tag. The use of tags for lexical and usage errors is highly dependent on the annotator’s judgment, and the results are low as expected.

### 3.4. Error tags depend on emendation

Analysis of the tagged data shows that the disagreement in using error tags is not necessarily caused by an annotator’s fault, but could rather be dependent on the choice of the emended form, both on the current and the preceding level. For ex-

ample, from the 181 cases of different use of the tag *agr*, 70 cases (39%) have a different L2 emendation. See (2) for an example.

(2)	R0:		<i>a kdyz stratil manzel</i>
	R2:	annotator A	<i>a když ztratí<sub>agr</sub> manžela<sub>dep</sub></i> 'and when she loses her husband'
		annotator B	<i>a když se ztratil manžel</i> 'and when the husband got lost'

From the remaining 111 disagreements in the use of the tag *agr*, 28 cases (15%) differ in the emendation already on L1.

In all these cases, tagging is correct vis-a-vis the selected emendation. Currently, we investigate the impact of emendation on error annotation at the individual levels, but we can already support the requirement of explicit target interpretation in the annotation scheme (Lüdeling, 2008). The scheme can thus be verified by the calculation of IAA in the distribution of the tags, depending on the final hypothesis (cf, i.a., Meurers, 2011).

### 3.5. Outline of the possible causes of the annotators disagreement

We can identify the following causes of the annotators' disagreements:

1. Invalid or imprecise annotation scheme: Generally, the annotators' disagreement can be caused by the annotation scheme itself: if it includes invalid tags or misses some necessary tags, or if the definition of a tag misleads the annotator. In the case of trial tagging of a sample of CzeSL data, it was problematic in several points, such as poorly distinguished subtypes of word boundary error (*wbd*), fuzzy definition of the error in pronominal reference (*ref*), also in contrast to the *agr* and *dep* types, or an imprecise boundary between the wrong choice of verbal tense (*use*) and the error in the analytical verb form (*vbX*).

2. Insufficient screening and training of the annotators: The level of screening and training process has a significant effect on the IAA rate. Higher IAA was demonstrated for annotators exposed to extensive and detailed pre-annotation training. It would be interesting to test what kind of impact the annotators' exposure to Czech as a foreign language has on the consistency of their annotation.

3. Different target hypotheses: Some annotations require a considerable amount of interpretation, while each annotator can have his/her own interpretation because of age, gender, education, etc. Moreover, in the case of multilevel annotation, annotators can differ also on intermediate levels, even though their target hypothesis might be identical. However, the annotation scheme of CzeSL, supporting emendation on both levels, makes the reasons for the possible disagreements explicit.

#### 4. Automatic processing of learner texts

Despite the benefits of annotators' insight and judgment, manual annotation is tedious and costly. On the other hand, automatic tools are more error-prone and cannot produce the sort of sophisticated annotation envisaged in the present project. Aware of these pros and cons, we are still interested in how far we could get without manual annotation. Due to the lack of methods targeting learner texts, we confronted some 'native Czech' tools (two taggers and a spell checker) with ill-formed input.

The two taggers are based on different concepts: *Morče* (Votrubec, 2006) uses a morphological analyser, preferring lexical and morphological diagnostics over syntactic context, while *TnT* (Brants, 2000) has the opposite strategy while using lexicon extracted from training data. Both taggers were trained on the same tagset and include a method to handle unknown words. Because of the different strategies the taggers use to tag correct input, they respond differently to various types of deviations. A mutual comparison of their results is thus as interesting as their evaluation against a golden standard, which – in the case of ill-formed input – is a difficult concept anyway.

Identifying all errors would involve comparing manual annotations at L2 form-by-form with the original text at L0. In the current absence of such data, we used data obtained from the easier task of comparing L0 to L1, where all erroneous forms are emended to a closest correct version, disregarding context.

Table 3 presents data extracted from a sample of 93 texts including 12,681 word tokens, with 1,323 tokens (8.9%) identified as ill-formed by the morphological analyser. The two taggers agreed on the same tag in 405 cases, i.e. in 28.8% of the total of ill-formed tokens, and disagreed in 918 cases (71.2%). The figures are additionally split by 12 morphological categories constituting the tag. Column 1 (L0m x L0t) shows in which categories the two taggers disagree at L0 for the 918 tokens, where their tags do not match at least in one category. Agreement is significantly lower between categories largely determined by syntactic context (POS, Gender, Number, Case) as opposed to those determined lexically. Columns 2 (L0m x L1) and 3 (L0t x L1) show agreement rates of tags assigned by *Morče* and *TnT*, respectively, to all tokens at L0<sup>5</sup> in comparison with tags assigned by *Morče* to the corresponding tokens at L1.<sup>6</sup> *Morče* shows better results overall and in most categories. Columns 4 and 5 show agreement rates for an ill-formed subset of the sample used in Columns 2 and 3. Interest-

---

5 The size of the sample is smaller than in the previous comparison at L0 only due to a more demanding procedure to obtain the data at L1.

6 The reason why *Morče* was used to tag L1 is because it is currently the best tagger of Czech and we were only interested in the cross-tagger comparison on the ill-formed input at L0.

ingly, *TnT* shows significantly better results, except in the categories of Person and Tense.

**Table 3.** Tags on L0 and L1 – percentages of agreement

	L0m x L0t	L0m x L1	L0t x L1	L0m x L1	L0t x L1
<b>No. of tokens</b>	918	2589	2589	314	314
<b>Entire tag</b>	0	84.1	79.0	19.1	26.1
<b>POS</b>	39.2	89.6	88.7	43.9	52.5
<b>SubPOS</b>	37.1	89.2	87.9	42.0	49.7
<b>Gender</b>	23.9	88.8	88.2	36.0	46.5
<b>Number</b>	36.9	91.1	91.2	49.0	63.1
<b>Case</b>	31.2	89.0	86.5	43.0	51.3
<b>Possessive Gender</b>	98.6	99.8	99.9	98.4	99.7
<b>Possessive Number</b>	99.5	99.8	99.7	99.0	99.7
<b>Person</b>	68.1	96.3	94.2	81.8	76.1
<b>Tense</b>	70.6	96.7	95.3	83.1	77.4
<b>Grade</b>	78.3	96.4	96.9	75.2	81.5
<b>Negation</b>	74.4	95.3	93.8	73.9	74.2
<b>Voice</b>	70.6	96.7	95.5	83.1	78.7

The difference between the two taggers is also reflected in the share of different POS categories assigned to ill-formed words. Table 4 shows that *Morče* has a more even distribution, but strongly disprefers all verbal categories.

**Table 4.** Numbers of tags assigned to ill-formed words

POS	<i>Morče</i>	<i>TnT</i>	POS	<i>Morče</i>	<i>TnT</i>
adjective	158	94	particle	8	–
adverb	118	21	finite verb	32	129
gradable adverb	31	11	infinitive	7	41
Noun	499	441	l-participle	10	119
preposition	10	–	passive participle	1	29

To sum up, the comparison of the two taggers confirms the assumption that the differences in their strategies will have a significant effect on the interpretation of faulty forms. A more general observation concerns the comparison of the success rate of the two taggers on the ill-formed input: *TnT* loses ground in a context with many errors but outperforms *Morče* on faulty forms, while *Morče* strongly disprefers verbs and works better in general.

Next, a spell checker was used to test whether automatic emendation is possible. In fact, *Korektor* (Richter, 2010) has some functionalities of a grammar checker, using a sophisticated combination of lexicon, morphology and a syntax

model. Moreover, the same core can be used as a spell checker, proofreader and diacritics assigner.

The tool was tested on a subset of the 918 ill-formed words in the sample, namely on those where the annotators came up with an identical emendation – there were 786 such forms, assumed as “truth”. *Korektor* was used in three different scenarios. The diacritics assigner mode was right in 552 cases (70.2%), the proofreader mode in 639 cases (80.5%), and the diacritics assigner followed by proofreader in 644 cases (81.9%). Even though L1 data are not the optimal benchmark, this tool seems to be useful as an aid to the annotator, or even as a way to obtain large quantities of annotated (emended) texts at the cost of a higher error rate.

## 5. Conclusion

It is no simple task to design an annotation scheme for a learner corpus and to maintain consistency in the annotated texts, both in a way that would reflect most demands of the corpus users. One of the main reasons is that annotating learner texts tends to be a highly specific enterprise, and even seemingly similar projects do not offer enough guidance – solutions are often too specific to a language or to the project concept and user requirements. On the other hand, annotation itself is quite rewarding due to the plentiful feedback about all aspects of the task and, of course, about the learners’ interlanguage.

More specifically, our experience shows the rules for tagging morphosyntactic errors (labelled in the annotated texts as *incorStem*, *incorInfl*, *agr* and *dep*) are relatively easy to formalize and it is thus possible to obtain a high inter-annotator agreement for such errors. However, we were unable to obtain a similarly robust annotation of semantic errors, which are much more dependent on subjective judgement. It is even unclear whether it is desirable to aim to standardize their annotation.

The pilot study, where two POS taggers and a spell checker were applied to ill-formed input, confirmed the viability of a fully automatic annotation as an alternative to manual annotation, when the demand for large data is higher than concerns about the error rate. It remains to be seen to what extent the comparison of results of multiple taggers, based on different tagging strategies, can lead to usable interpretations of faulty forms.

## Acknowledgments

The authors wish to express thanks to other members of the project team, namely Karel Šebesta, Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and Hana Skoumalová.

## References

- Bedřichová, Z., Šormová, K. and K. Šebesta (2011). ROMI – první rozsáhlá databanka romského etnolektu češtiny [ROMI – the First Large Database of Romani ethnolect of Czech]. *Lidé města*. 2011, n. 13/1. Available at: <http://lidemesta.cz/index.php?id=769>
- Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, vol. 20, no. 1, s. 37–46.
- Díaz-Negrillo, A., Meurers, D., Valera, S. and H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*. 36, 139–154.
- Hana, J., Rosen, A., Škodová, S. and B. Štindlová (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala: Association for Computational Linguistics.
- Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Grommes, P. and Walter, M. (eds.) *Fortgeschrittene Lerner-varietäten*. Niemeyer, Tübingen, pp. 119–140.
- Meurers, D. (2011). On Automatically Analyzing Learner Language: Interpreting Form and Meaning in Context. Invited talk at *Colloquium of the Research Center for English and Applied Linguistics (RCEAL)*, 8.2.2011. University of Cambridge. Available at: <http://www.sfs.uni-tuebingen.de/~dm/presentations.html>
- Richter, M. (2010). *Pokročilý korektor češtiny [An Advanced Spell Checker of Czech]*. Master's Thesis. Charles University, Praha, Faculty of Mathematics and Physics.
- Rietveld, T. and R. Van Hout (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.
- Rozovskaya, A. and D. Roth (2010). Annotating ESL Errors: Challenges and Rewards. In *Proceedings of NAACL'10 Workshop on Innovative Use of NLP for Building Educational Applications* University of Illinois at Urbana–Champ, Available at: <http://www.cs.rochester.edu/~tetreaul/naacl-bea5.html>
- Šebesta, K. Korpusy češtiny a osvojování jazyka [Corpora of Czech and Language Acquisition]. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*. 2010, vol. 1, n.2, pp. 11–34.
- Štindlová, B. (2011). *Evaluaace chybové anotace v žákovském korpusu češtiny [Evaluation of Error Mark-Up in a Learner Corpus of Czech]*. Dissertation. Charles University, Praha, Faculty of Arts.
- Štindlová, B., Škodová, S., Rosen, A. and J. Hana (in press). Annotating foreign learners' Czech. In *Proceedings of FDSL 8.5*, Brno, 2010.
- Tetreault, J. and M. Chodorow (2008). Native Judgements of Non-Native Usage: Experiments in Preposition Error Detection. In *COLING Workshop on Human Judgements in Computational Linguistics*. Manchester. Available at: <http://portal.acm.org/citation.cfm?id=1611633>
- Votrubec, J. (2006). Morphological Tagging Based on Averaged Perceptron. In *WDS'06 Proceedings of Contributed Papers*, Matfyzpress, Charles University, Praha, pp. 191–195.