ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 80–87
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 A. Rosen

ITAT

Building and using corpora of non-native Czech

Alexandr Rosen

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

1 Introduction

Investigating language acquisition by non-native learners helps to understand important linguistic issues and develop teaching methods, better suited both to the specific target language and to the learner. These tasks can now be based on empirical evidence from learner corpora.

A learner corpus consists of language produced by language learners, typically learners of a second or foreign language (L2). Such corpora may be equipped with morphological and syntactic annotation, together with the detection, correction and categorization of non-standard linguistic phenomena.

The tasks of designing, compiling, annotating and presenting such corpora are often very much unlike those routinely applied to standard corpora. There may be no standard or obvious solutions: the approach to the tasks is often seen as an answer to a specific research goal rather than as a service to a wider community of researchers and practitioners. Our aim is to investigate some of the challenges, based on a learner corpus of Czech in comparison to several other learner corpora.

After an overview of learner corpora around the world in §2 and a brief presentation of several releases of a learner corpus of Czech in §3, we examine issues inherent to the process of compiling, annotating and using such corpora, including automatic identification of errors, the design and application of error taxonomy, and a user-friendly search tool, suited to a complex annotation (§4).

2 About learner corpora

Most of the existing learner corpora include English (L2) as produced by students whose native languages (L1) are varied. Most of the corpora are partially error-annotated, see Table 1 on p. .[1] The error annotation is usually inline, equivalent to XML tags, denoting the scope, correction and categorization of an error. A few corpora such as FALKO include multi-layered annotation in a tabular format, with the option of specifying multiple target hypotheses (corrections) and several error types for single word tokens or strings thereof at different levels of linguistic abstraction: orthography, morphology, syntax, lexicon, pragmatics, intelligibility.

The tabular format is also used in *MERLIN*, one of the two currently available corpora including Czech.[2] In addition to 64.5K words of Czech in CEFR levels A1–C1, the corpus includes also German and Italian. It is tagged, lemmatized, parsed and on-line searchable, with a detailed error taxonomy and the option of two target hypotheses.

3 *CzeSL* – the learner corpus of Czech as a Second Language

CzeSL is a part of an umbrella project, the Acquisition Corpora of Czech (*AKCES*), a research programme pursued since 2005 (Šebesta, 2010). In addition to *CzeSL*, *AKCES* has a written (*SKRIPT*) and spoken (*SCHOLA*) part collected from native Czech pupils, and *ROMi*, a part collected from pupils with Romani background, using the Romani ethnolect of Czech as their first language (L1). In the present paper we focus on written texts produced by non-native learners of Czech. However, most of the methods and tools can be applied to other parts of the corpus.

CzeSL is focused on native speakers of three main language groups: (1) Slavic, (2) other Indo-European, (3) non-Indo-European. The hand-written texts cover all language levels, from real beginners (A1) to advanced learners (B2, C1, C2). The texts are equipped with metadata records; some of them relate to the respondent (age, gender, first language, proficiency in Czech, knowledge of other languages, duration and conditions of language acquisition), while other specify the character of the text and circumstances of its production (availability of reference tools, type of elicitation, temporal and size restrictions etc.).

The hand-written texts were transcribed using off-the-shelf editors supporting HTML (e.g., Microsoft Word or Open Office Writer). A set of codes was used to capture variants, illegible strings, self-corrections; for details see (Štindlová, 2011b, p. 106ff). During the transcription step, the texts were 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 were replaced by `QQQ`. Unreadable characters or words were transcribed as `XXX`.

The transcripts were converted into an XML format. Some of them were corrected ('emended') and labelled

[1] For a more extensive overview see Štindlová (2011a) or an actively maintained list at https://www.uclouvain.be/en-cecl-lcworld.html.

[2] *Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context*, see http://merlin-platform.eu and Wisniewski et al. (2014); Boyd et al. (2014)

by error categories using a custom-built annotation editor, supporting a two-layered annotation format with $m : n$ links between tokens at the neighbouring tiers.[3] In a post-processing step the hand-annotated texts were tagged by tools trained on native Czech in a way similar to standard corpora, i.e. by lemmas, morphosyntactic categories, in some (currently non-public) releases of the corpus also by syntactic functions and structure. Some error annotation tasks were also done automatically: the assignment of formal error labels and even the correction step (the latter in *Czesl-SGT*, see §3.2).

There are several public releases of *CzeSL*, which differ in the depth and method of annotation, but also in the availability of metadata and size. Table 2 shows the content of available releases of *CzeSL*, including the volumes (in thousands of tokens), and the availability of annotation and metadata.[4]

3.1 Releases of *CzeSL* without metadata: *CzeSL-plain* and *CzeSL-man v. 0*

Since 2012, the transcripts of essays hand-written by non-native learners (1.3 mil. tokens) and pupils speaking the Romani ethnolect of Czech (0.4 mil. tokens) have been available together with some Bachelor and Master theses written in Czech by foreign students (0.7 mil. tokens) as the *CzeSL-plain* corpus, on-line searchable via a web-based search interface of the Czech National Corpus,[5] or as full texts under the Creative Commons license from the LINDAT repository.[6] Except for specifying the three groups above and a basic structural mark-up, this corpus does not include any metadata or annotation.

CzeSL-man v. 0 includes subsets of *CzeSL* and *ROMi*, about 330 thousand tokens. It is manually error-annotated at two levels. Texts of about 208 thousand tokens are annotated independently by two annotators. Like *CzeSL-plain*, the whole hand-annotated part is accessible online without metadata via a purpose-built search tool (*SeLaQ*);[7] for more about the manual annotation and the annotation process see Hana et al. (2014).

The manual annotation scheme in *CzeSL* is based on a two-stage annotation design, reflecting the distinction roughly between errors in orthography and morphemics on the one hand and all other error types on the other. Tokens in the original transcript are linked with their counterparts at the two successive levels by edges, possibly labelled with the type of error – see Figure 1 on p. . A syntactic error label may be linked by a pointer to a word token, specifying an agreement, valency or referential relation.[8] The level of transcribed input (Tier 0) is followed by the level of orthographical and morphemic corrections (Tier 1), where only forms incorrect in any context are treated. Errors at Tier 1 are mainly non-word errors while those at Tier 2 are real-word and grammatical errors. However, a faulty form that happens to be spelled as a form which would be correct in a different context, is still corrected at Tier 1. The result at Tier 1 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 Tier 2, representing a grammatically correct, though stylistically not necessarily optimal target hypothesis.[9] Manual annotation is complemented by morphosyntactic tags and lemmas at Tier 2, ambiguously specified tags and lemmas at Tier 1, and automatically identified formal errors.[10] Splitting, joining and reordering words, together with the pointers may make the picture rather complex, as in an authentic sentence in Figure 1 on p. .

The three tiers are represented as parallel strings of word forms with links for corresponding forms. Tier 0 is glossed for readability; forms marked by asterisks are incorrect in any context.

Errors corrected at Tier 1 include incorrect inflection (**incorInfl**), word boundaries (**wbdPre**), and stems (**incorBase**). Errors in punctuation (the missing comma), capitalization (*prahu*) or word order (*se* in the *that*-clause at Tier 2) are tagged automatically in a post-processing step.

Tier 2 captures the rest of errors. Some error labels are linked to a token which makes the reason for the correction explicit. This includes errors in agreement (**agr**), government or valency in a broad sense (**dep**), complex verb forms (**vbx**) or reflexive particles (**rflx**). For example, *ona* in the nominative case is governed by the form *líbit se*, and should be in the dative case: *jí*. The label **dep** has an arrow pointing to the governor *líbit*. There is also a simple **lex**ical correction: *Proto* 'therefore' is changed to *protože* 'because'.

However, the main issue are the two finite verbs *bylo* and *vadí*. The most likely intention of the author is best expressed by the conditional mood. The two non-contiguous forms are replaced by the conditional auxiliary and the content verb participle in one step using a 2:2 relation. Another complex issue is the prepositional phrase *pro mně* 'for me'. Its proper form is *pro mě* (homonymous with *pro mně*, but with 'me' in accusative instead of dative), or *pro mne*. The accusative case is required by the preposition *pro*. However, the head verb requires that this complement bears bare dative – *mi*. Additionally, this form is a

[3] https://bitbucket.org/jhana/feat

[4] Some texts in *CzeSL-man v.0* are doubly annotated. The texts annotated by an additional annotator are included in the *CzeSL-man v.0, a2* part. See http://utkl.ff.cuni.cz/learncorp/ for links and more details.

[5] https://kontext.korpus.cz

[6] http://lindat.mff.cuni.cz

[7] http://chomsky.ruk.cuni.cz:5125

[8] This scheme is already a compromise between a linear annotation and an open multi-layered format, but a compromise preserving links between split, joined and re-ordered tokens, corrected in two stages simultaneously, something not obviously supported in the multilayered tabular format mentioned above in §2.

[9] See Hana et al. (2010) and Rosen et al. (2014) for more details.

[10] See Jelínek et al. (2012) for details, including a list of formal error types. The last column of Table 3 shows examples of the formal error labels.

clitic, following the conditional auxiliary.

The correction *slavnou*$_{accusative}$ →*slavná*$_{nominative}$ is due to the correction of the case of the head noun. Such corrections receive an additional label as **sec**ondary errors.

3.2 The automatically anotated *CzeSL-SGT*

The 'real' *CzeSL*, i.e. the corpus consisting of essays written only by non-native learners (1.1 mil. tokens), is available with automatic annotation as *CzeSL-SGT*,[11] extending the "foreign" part of the *CzeSL-plain* corpus by texts collected in 2013. This was the first release of *CzeSL* including full metadata. The corpus includes 8,617 texts by 1,965 different authors with 54 different first languages. The original transcription markup is discarded in this corpus, while the final author's version is restored. The corpus is available again either for on-line searching using the search interface of the Czech National Corpus or for download from the LINDAT data repository.[12]

Word forms are tagged by word class, morphological categories and base forms (lemmas). Some forms are corrected by *Korektor*, a context-sensitive spelling/grammar checker,[13] and the resulting texts are tagged again. Original and corrected forms are compared and error labels are assigned. *Korektor* detected and corrected 13.24% incorrect forms, 10.33% labelled 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 were tagged and lemmatized, and "formal error tags," based on the comparison of the uncorrected and corrected forms, were assigned.[14] The share of non-words detected by the tagger is slightly lower – 9.23% (the tagger uses a larger lexicon).

Automatic correction is a crucial annotation step. The tool is concerned mainly with errors in orthography and morphemics, and handles some errors in morphosyntax, including real-word errors (i.e. errors that produce a word which seems to be correct out of context), as long as they are detectable locally, within a reasonably small window of *n*-grams. Corrections are limited to single words, targetting a single character or a very small number of characters by insertion, omission, substitution, transposition, addition, deletion or substitution of a diacritic. Errors that involve joining or splitting of word tokens or word-order errors of any type are not handled at the moment.

The performance of *Korektor* was evaluated first in Štindlová et al. (2012) with about 20% error rate on the set of non-words, and later in Ramasamy et al. (2015). In an optimal setting of the model, the best results achieved in terms of F1 score were 95.4% for error detection and 91.0% for error correction. In a manual analysis of 3000 tokens, about 23% of the tokens included either a form error at Tier 1 (62%), a grammar error at Tier 2 (27%), or an accumulated error at both tiers (11%). Form errors were detected with a success rate of 89%. For grammar errors (real-word errors) the detection rate was much lower, about 15.5%. The detection of accumulated errors was similar to form errors (89%).

After all the automatic annotation steps are finished, 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; for grammar errors, `word` is mostly recognized
- `err` – error type, determined by comparing `word` and `word1`.

Table 3 on p. shows the use of the annotation in a simple sentence (1).[15]

(1) T**é**n pes m**í**luje své**ch**o kamar**a**da – člověka.
that dog loves self's friend – man
'That dog loves its friend – the 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`. Dynamic attributes 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)

[11] ***Cze**ch as a **S**econd **L**anguage with **S**pelling, **G**rammar and **T**ags*

[12] http://hdl.handle.net/11234/1-162

[13] See Richter et al. (2012). The tool is available from the LINDAT repository (https://lindat.mff.cuni.cz) under the FreeBSD license.

[14] See Jelínek et al. (2012).

[15] The example comes from a CzeSL-SGT text, written by a 17 years old student, with Russian as L1 and B2 as the proficiency level in Czech (document ID ttt_G1_434).

- p, p1 – person (position 8 of the tag)

They are meant especially for CQL queries[16] including a "global condition". As in standard corpora, such queries target two or more word tokens with an arbitrary but equal value of an attribute such as case to express grammatical agreement and similar morphosyntactic phenomena (2).

(2) 1:[] 2:[] & 1.c = 2.c

In a learner corpus, such queries make sense even for a single word token, e.g. for expressing identical or distinct values of the morphological case of the original form and of its corrected version (3).[17]

(3) 1:[] & 1.c != 1.c1

In a learner corpus, metadata about the author of the text are at least as important as all other types of annotation. For the number of texts authored by students according to their first language and the CEFR proficiency level in Czech see Table 4 below. The language group abbreviations read as follows: IE = non-Slavic Indo-European, nIE = non-Indo-European, S = Slavic.

	S	IE	nIE	unknown	Σ
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
Σ	5642	898	1747	330	8617

Table 4: Number of texts by language group and proficiency level in *CzeSL-SGT*

3.3 *CzeSL-man v. 1*

CzeSL-man v. 1 is a collection of manually annotated transcripts of essays of non-native speakers of Czech, written in 2009–2013, the total of 645 texts, including 298 doubly annotated texts. The texts contain 128 thousand word tokens, including 59 thousand doubly annotated tokens; for a comparison with *CzeSL-SGT* see Table 5.

Tables 6 and 7 show the number of texts for each combination of CEFR level and language group in *CzeSL-man v. 1*.

[16] See https://www.sketchengine.co.uk/corpus-querying/

[17] Unfortunately, queries including global conditions on dynamic attributes do not produce expected results in the present version of the *Manatee* search engine.

	CzeSL-SGT	*CzeSL-man v. 1*
Texts	8,600	645
Sentences	111K	11K
Words	958K	104K
Tokens	1,148K	128K
Different authors	1,965	262
Different L1s	54	32
Proficiency levels	A1–C2	A1–C1
Women/Men	5:3	3:2
Words per text	100–200	100–200

Table 5: *CzeSL-man v. 1* and *CzeSL-SGT* compared

	S	IE	nIE	unknown	Σ
A1	49	6	4		59
A1+			3		3
A2	18	26	67		111
A2+	81	9	59		149
B1	123	26	30		179
B2	102	11	15		128
C1	10		2		12
unknown				4	4
Σ	383	78	180	4	645

Table 6: Number of texts by language group and proficiency level in *CzeSL-man v. 1*

In addition to the number of tokens for the same category, Table 8 shows also the frequency of errors of the **dep** type, i.e. valency errors in the broad sense, including errors in the number of complements and adjuncts or errors in their morphosyntactic expression. The rather frequent error type shows a considerable and expected decrease in higher proficiency levels

CzeSL-man v. 1 is about to be released soon for download in the LINDAT repository and for on-line searching in https://kontext.korpus.cz. Some solutions to the problem of using a feature-rich corpus search engine, which is still not suited to the two-level annotation scheme of *CzeSL-man*, are presented in 4.

4 Some issues and lessons learnt

Several points can be made about some of the *CzeSL* releases, reflecting issues involved in the design, compilation and presentation of learner corpora.

We start with *CzeSL-plain* and its hand-annotated part *CzeSL-man v. 0*: (i) Both corpora include some *ROMi* texts, actually produced by native speakers of a *dialect* of Czech, rather than by non-native speakers of Czech. This is due to the original strategy of grouping texts by the way they are processed. This has been changed in later releases, where texts produced by non-native and native learners (the latter including speakers of the Romani ethnolect of Czech) are parts of distinct corpora. (ii) Neither

	S	IE	nIE	Σ
A1	37	2	1	40
A1+			3	3
A2	5	23	47	75
A2+	21	6	49	76
B1	20	23	28	71
B2	7	11	12	30
C1	1		2	3
Σ	91	65	142	298

Table 7: Number of doubly annotated texts by language group and proficiency level in *CzeSL-man v. 1*

	A1	**A2**	**B1**	**B2**	**C1**	Σ
IE	227	7,336	5,311	2,340	0	15,214
dep	13	361	118	28	0	520
%dep	5.73%	4.92%	2.22%	1.20%		3.42%
nIE	439	17,640	7,606	4,219	760	30,664
dep	13	715	237	116	7	1,088
%dep	2.96%	4.05%	3.12%	2.75%	0.92%	3.55%
S	6,434	16,939	27,226	22,173	4,761	77,533
dep	225	470	652	443	17	1,807
%dep	3.50%	2.77%	2.39%	2.00%	0.36%	2.33%
Σ	7,100	41,915	40,143	28,732	5,521	123,411
dep	251	1,546	1,007	587	24	3,415
%dep	3.54%	3.69%	2.51%	2.04%	0.43%	2.77%

Table 8: Number of tokens and valency errors by language group and proficiency level in CzeSL-man v. 1

CzeSL-plain nor *CzeSL-man v. 0* includes the full set of metadata, which were not available in the appropriate form and content at the time the two corpora were prepared and released. In *CzeSL-plain*, the texts are categorized into three groups: as essays, written either by non-native learners, or by speakers of the Roma ethnolect of Czech, and as theses written by non-native students. In *CzeSL-man v. 0* there is no distiction available. (iii) Due to the uncertainty abouth the optimal way of representing the complex two-level manual annotation, the *SeLaQ* tool cannot display the two-level annotation format in a graphical format.

There is a strong demand for *CzeSL-man* to become available for on-line searches at the Czech National Corpus portal, even if some of the properties and information present in the corpus may get lost in the conversion to the format used by the corpus search tool, based on the single-level annotation of a string of tokens. However, the converted format might still retain enough annotation to be attractive and useful for most tasks. Instead of assigning the error-related annotation to word tokens, which makes the option to annotate strings of tokens, or even discontinuous strings very difficult, errors and corrections can be treated as structural annotation, i.e. similarly to the markup for paragraphs, sentences, phrases or text chunks. Even the splitting and joining of words and word order corrections can then be expressed.

The *Manatee* corpus search engine, used in the Czech National Corpus, and its (*No*)*Sketch Engine* front end actually include support for learner corpora,[18]. The in-line annotation can even have embedded structures, which may be used at least for some cases of multi-layered annotation. Making *CzeSL-man* with most of the annotation available this way thus seems a real prospect.

4.1 Corpus design and planning

The target corpus may be intended for a group of users with specific research or practical needs, or for a wide audience of language acquisition experts, researchers or practitioners. In any case the goals should be realistic in order to avoid a mission ending before the goals are achieved.

4.2 Text acquisition

Some balance or at least representative proportions of text and learner categories are necessary or at least useful. Tables 4–7 show an opposite, opportunistic approach, driven by practical constraints, often justified by the unavailablity of texts of a specific category.

4.3 Transcription

To avoid the need of cleaning transcripts with improperly used mark-up, an editing tool including strict format controls is preferable to a free-text editor.

4.4 Annotation scheme and searching

A scheme ideally suited to the data may turn into a problem later, if the consequences for the annotation process and the use of the corpus are not foreseen. Standard concordancers may require substantial tweaking of the data, while a custom-built tool may lack features of the tools developed for a long time. At the same time, most users of this type of corpora definitely need a friendly interface.

5 Conclusion

We have presented several releases of a learner corpus of Czech, available for on-line queries and under the Creative Commons license as full texts.

In order to reach its goals and become useful, a learner corpus project should be conceived carefully, considering many factors. By way of an example, we have shown some pitfalls in the process of building and presenting such a corpus.

The methods and tools developed within this project are not tied to the specific use and we hope they will be found useful in other projects.

[18] See https://www.sketchengine.co.uk/learner-corpus-functionality/

Acknowledgements

The corpus could never be built without many other members of the *CzeSL* team. For the work reported here the author is grateful especially to Barbora Štindlová, Jirka Hana and Tomáš Jelínek. The author's thanks are also due to two anonymous reviewers who helped to improve the paper, and to the Grant Agency of the Czech Republic, which currently provides financial support for *Non-native Czech from the Theoretical and Computational Perspective* (project ID 16-10185S).

References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hana, J., Rosen, A., Škodová, S., and Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden. Association for Computational Linguistics.

Hana, J., Rosen, A., Štindlová, B., and Štěpánek, J. (2014). Building a learner corpus. *Language Resources and Evaluation*, 48(4):741–752.

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

Ramasamy, L., Rosen, A., and Straňák, P. (2015). Improvements to Korektor: A case study with native and non-native Czech. In Yaghob, J., editor, *ITAT 2015: Information technologies – Applications and Theory / SloNLP 2015*, pages 73–80, Prague. Charles University in Prague.

Richter, M., Straňák, P., and Rosen, A. (2012). Korektor – a system for contextual spell-checking and diacritics completion. In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.

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

Wisniewski, K., Woldt, C., Schöne, K., Abel, A., Blaschitz, V., Štindlová, B., and Vodičková, K. (2014). The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language. Technical report. Available online http://merlin-platform.eu/.

Šebesta, K. (2010). Korpusy češtiny a osvojování jazyka [Corpora of Czech and language acquistion]. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*, 1:11–34.

Štindlová, B. (2011a). Evaluace chybové anotace navržené pro žákovský korpus češtiny. *SALi*, 2(2):37–60.

Štindlová, B. (2011b). *Evaluace chybové anotace v žákovském korpusu češtiny [Evaluation of Error Mark-Up in a Learner Corpus of Czech]*. PhD thesis, Charles University, Faculty of Arts, Prague.

Štindlová, B., Rosen, A., Hana, J., and Škodová, S. (2012). CzeSL – an error tagged corpus of Czech as a second language. In Pęzik, P., editor, *Corpus Data across Languages and Disciplines*, volume 28 of *Łódź Studies in Language*, pages 21–32, Frankfurt am Main. Peter Lang.

Corpus	Size (MW)	L1	L2	Level	Medium	Annotation
ICLE	3	26	en	advanced	written	part
CLC	35	130	en	all	written	part
LINDSEI	0.8	11	en	advanced	spoken	part
PELCRA	0.5	pl	en	all	written	part
USE	1.2	sv	en	advanced	written	no
HKUST	25	zh	en	advanced	written	part
CHUNGDAHM	131	ko	en	all	written	part
JEFLL	0.7	jp	en	beginners	written	part
MELD	1	16	en	advanced	written	no
MICASE	1.8	various	en	advanced	spoken	no
NICT JLE	2	jp	en	all	spoken	part
RusLTC	1.5	ru	en	advanced	written	no
FALKO	0.3	5	de	advanced	written	part
FRIDA	0.2	various	fr	med-adv	spoken	part
FLLOC	2	en	fr	all	spoken	no
PiKUST	0.04	18	sl	advanced	written	yes
ASU	0.5	various	no	advanced	written	no
TUFS	0.6 Mchars	various	jp	all	written	no

Table 1: A list of learner corpora around the world

	Non-native		Ethnolect	TOTAL	Annotation	Metadata
	Essays	Theses				
CzeSL-plain	1315	732	428	2475	no	no
CzeSL-SGT	1147			1147	auto	yes
CzeSL-man v.0, a1	134		192	326	manual	no
CzeSL-man v.0, a2	59		149	208	manual	no
CzeSL-man v.1	134			134	manual	yes

Table 2: Available releases of *CzeSL*

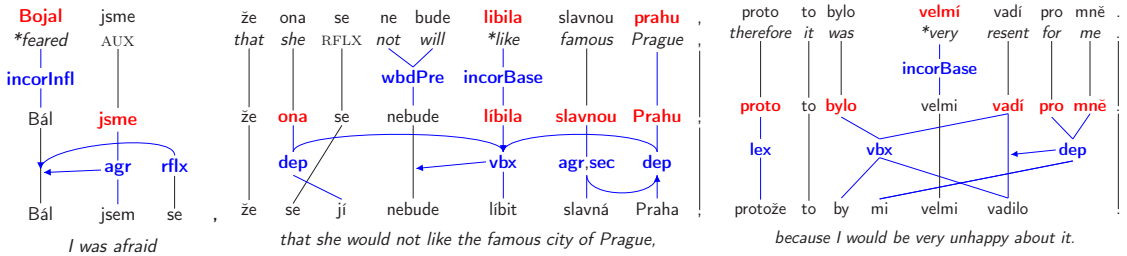

Figure 1: Two-level manual annotation of a sentence in *CzeSL*, the English glosses are added

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

Table 3: Annotation of a sample sentence in *CzeSL-SGT*