Building a learner corpus

Jirka Hana, Alexandr Rosen, Barbora Štindlová, Petr Jäger

Charles University in Prague, Faculty of Mathematics and Physics & Faculty of Arts
Technical University of Liberec, Faculty of Education
jirka.hana@gmail.com, alexandr.rosen@ff.cuni.cz, barbora.stindlova@tul.cz, petrtjager@gmail.com

Abstract

The paper describes a corpus of texts produced by non-native speakers of Czech. We discuss its annotation scheme, consisting of three interlinked levels to cope with a wide range of error types present in the input. Each level corrects different types of errors; links between the levels allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified. The annotation scheme is tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results. We also explore options of application of automated linguistic annotation tools (taggers, spell checkers and grammar checkers) on the learner text to support or even substitute manual annotation.

Keywords: learner corpora, error annotation, Czech

1. Introduction

We describe the process, annotation scheme, tools and technical decisions behind the annotation of a learner corpus of Czech. The corpus is compiled from texts written by students of Czech as a second or foreign language. We discuss the whole pre-processing work-flow, starting from the transcription of hand-written texts, conversion into the annotation format, and the annotation itself. Here, we focus mainly on the computational and organizational issues, for a detailed discussion of the linguistic aspects involved see (Hana et al., 2010; Štindlová et al., 2012).

After a brief introduction to the project of the learner corpus of Czech in §2., followed by an account of reasons for building such a corpus in §3., and a sketch of our annotation scheme in §4., we present a description of the workflow in §5.

2. A learner corpus of Czech

The learner corpus of Czech as a Second Language (*CzeSL*) is built as a part of a larger project, the Acquisition Corpora of Czech (*AKCES*), a research programme pursued at Charles University in Prague since 2005 (Šebesta, 2010). In addition to *CzeSL*, *AKCES* has written (*SKRIPT*) and spoken (*SCHOLA*) parts, collected from native Czech pupils, and a part collected from pupils with Romani background (*ROMi*). Methods and tools used for collecting, transcribing, annotating and managing the texts are currently the same at least for *CzeSL* and *ROMi*, and most of them are supposed to extend to other parts of the project.

By the end of the first phase of the project in May 2012, the size of transcribed texts of *CzeSL* and *ROMi* will reach the size of 2 million word tokens, of which nearly 20% are expected to be error-annotated manually. The anonymized transcripts, supplemented by rich metadata, will be available for registered users via a standard concordancer as a part of the Czech National Corpus,[1] while a newly built search tool will be used for the error-annotated parts with their complex mark-up.

[1] http://korpus.cz

CzeSL is focused on native speakers of the following languages: (1) Slavic, (2) other Indo-European, (3) non-Indo-European. The data include mainly written texts (such as essays or exams), collected as manuscripts. They cover all language levels, from real beginners (A1) to advanced learners (B2 and higher).

Each text is equipped with metadata records, some of them relate to the respondent (such as 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.).

3. What can be done with a learner corpus

In addition to its role in the research of second language acquisition, the corpus will be used in the education of teachers of Czech as a foreign language, as a source of examples usable in the classroom and for educational tools, and will help to tailor instructions and teaching materials to specific groups of learners (Štindlová, 2011).

Despite the fact that teaching Czech as a foreign language has already become a well-established field, a proper teaching methodology is not available. Teachers often take recourse to methods and techniques used for languages such as English or German. Since Czech is a typologically different language, there are a number of phenomena that make such an approach grossly inadequate, such as much richer morphology or relatively free word order. Alternatively, teachers tend to confront their students with Czech grammar as an academic subject, making the necessary learning curve extremely steep.

A specific problem is the issue of educating children with a native language other than Czech, whose presence at Czech primary schools is a fairly recent phenomenon. Primary school teachers receive no training in teaching Czech as a foreign language, again resorting to an individual and intuitive approach.

To help answer these issues, *CzeSL* is built as a resource for educational and linguistic research and for the design of

teaching materials assisting teachers of non-native speakers (of standard Czech) at different stages of acquiring the language. At the same time, *CzeSL* should provide representative data that would help initiate and develop a systematic and comprehensive research of Czech as a foreign language (so far, there are no monographs available dealing with this topic). It should support both computer-aided error analysis and contrastive interlanguage analysis, i.e. studies of a student's interlanguage in comparison to her native language or another interlanguage (Granger, 1998).

Texts collected for *CzeSL* are already in use in the training of teachers to give them an idea about the traits of the learner language in relation to the author's L1 and proficiency. This should help them to change perspective from viewing the language as an abstract system to approaching Czech as a sum of components acquired by learners at a specific stage of the development of their *interlanguage*, i.e. a language approximating Czech at a specific point of the learning process.

More specifically, we expect the corpus to be used interactively via the user interface of a corpus manager (concordancer), or by other tools. The first scenario may be typical in the preparation of teaching materials by teachers, in independent work by their foreign students, or by students in teacher training programmes, and in research by experts for L2 analysis. This use of the corpus might include querying the data in the class. The queries may target word forms or phrases, both in their original and emended form, their lemmas, POS and error tags, their position with the sentence, the corresponding metadata, or any combination of the above.

The second scenario aims at producing aggregate values of various types, typically based on processing the whole corpus or of large subsets of it. It involves tasks such as quantitative analysis of learner language, using statistical summaries for error types in general, error types on specific lexemes, constructions and grammatical categories. The analysis may also target overuse/underuse of lexemes or phenomena as compared to native speakers' corpora, perhaps as a way to investigate L1 interference.

Moreover, we expect the corpus to be used in more sophisticated didactic applications and in machine learning to built models for automatic emendation and error annotation.

The corpus will be available via concordancer to registered users for non-profit research, or even as full texts under a research exemption clause. However, some parts of metadata, scanned documents and proper names in the texts will not be revealed.

4. Annotation scheme

Texts produced by non-native speakers can be annotated in a way similar to standard corpora, e.g., by POS tags, syntactic functions or syntactic structure, but also corrected ('emended') and labelled by error categories.

The optimal error annotation strategy is determined by the goals of the project and by the type of the language. Single-level schemes could be used, e.g., for a specific purpose or for a language without an elaborate inflection system. However, our corpus should be open to multiple research goals and handle a highly inflectional language. This is why it is

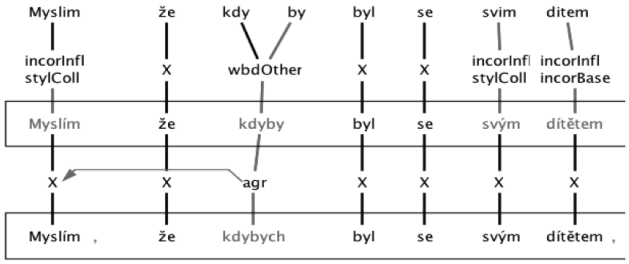

Figure 1: Example of the three-level error annotation scheme

based on a multi-level annotation scheme, allowing: (i) to register successive emendations and errors spanning multiple (potentially discontinuous) forms, and (ii) to maintain links between the original and the emended form even when the word order changes, or in cases of dropped or added expressions.

We adopted a solution with two levels of annotation, distinguished by formal but linguistically founded criteria. The scheme consists of three interconnected levels – see Fig. 1, glossed in (1):

- Level -1 – Anonymized transcript of the hand-written original in html format, encoding self-corrections etc.
- Level 0 – Tokenized text
- Level 1 – Forms wrong in isolation are corrected. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole.
- Level 2 – Handles all other types of errors (such as valency, agreement, and word order).

(1) *Myslím, že kdybych byl se svým*
think$_{SG1}$ that if$_{SG1}$ was$_{MASC.SG}$ with my
dítětem,
child,
'I think that if I were with my child, . . . '

The correspondences between successively emended forms are explicitly expressed. Nodes at neighboring levels are usually linked 1:1, but words can be joined (*kdy by* in Fig. 1), split, deleted or added. These relations can interlink any number of potentially non-contiguous words across the neighboring levels. Multiple words can thus be identified as a single unit, while any of the participating word forms can retain their 1:1 links with their counterparts at other levels. 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. Error labels used in Fig. 1 include *incorInfl* or *incorBase* for morphological errors in inflectional endings or stems, *stylColl* as a stylistic marker (here for a colloquial form), *wbdOther* as a word boundary error (other than wrongly separated prefix or preposition without a following space), and *agr* as an error in agreement. Some errors may additionally require a link pointing to a form

specifying proper agreement categories or valency requirements (such as *myslím* in our example).
The taxonomy of errors is based on linguistic categories, complemented by a classification of superficial alternations of the source text, such as the indication of a missing, redundant, faulty or incorrectly ordered element.
In addition to the form of the word, each node may be assigned information such as lemma, morphosyntactic category or syntactic function.

5. Tasks and tools in the workflow

The whole annotation process proceeds by the following steps:

1. Acquisition: The original hand-written texts are collected from schools or other sources and scanned. Data about the author and circumstances of the elicitation of the text are supplied.

2. Transcription: The scan is manually transcribed into a HTML format.

3. Proofreading: Each transcription is checked by a supervisor.

4. Conversion to PML (an XML format, see §5.2.): The transcribed HTML text is tokenized and the corresponding Level 0, encoded in PML, is generated. A default Level 1 and an empty Level 2 are created as well. The conversion includes basic checks for incorrect or suspicious transcription.

5. Annotation supervisors distribute documents to annotators using *Speed*, a purpose-built text management system (see §5.1.).

6. Error annotation: Errors in the text are manually corrected and classified using *feat*, an annotation editor designed as a part of the project (see §5.4.).

7. Each annotation is reviewed by the appropriate supervisor, who can approve it or return to the annotator for correction (with a single click of a button).

8. All texts are annotated independently by two annotators. Annotations of the same texts are checked and adjudicated.

9. Postprocessing: Error information that can be inferred automatically is added. The corrected text is automatically lemmatized and tagged with morphosyntactic information.

5.1. Text management

To coordinate work of a large project team and to control the passage of texts along the path from the scanned manuscript up to the annotated and adjudicated result, all versions of every document throughout the whole process are stored and maintained in *Speed*, a text management system developed as a part of the project.
The system distributes documents to transcribers, annotators and coordinators for processing and accepts the results, monitoring their workload and generating error-rate statistics on demand. Using this tool, coordinators can manage the team of 30 annotators efficiently, without wasting their time on administrative tasks.
User privileges are consistently applied both horizontally and vertically. Each user is assigned her views of the data and filters associated with those views. As a result, the annotator is prevented from seeing an interpretation used by a colleague. At the same time, the system is shielded from potential faults and inconsistencies within the users' local file systems.
The system is designed on top of a general workflow machine, reusable for similar applications, and it is linked with the off-line annotation tool *feat* using web services. The users receive their tasks and deliver results without leaving the environment of the application. This includes quality checking – through the same channel, the annotator may receive an inadequately annotated text for review with comments by the supervisor.

5.2. Data Format

To encode the layered annotation described above, we have designed an annotation schema in the Prague Markup Language (PML).[2] PML is a generic XML-based data format, intended for the representation of rich linguistic annotation organized into levels. In our schema, each of the higher levels contains information about words on that level, about the corrected errors and about relations to the tokens on the lower levels.
We have also considered to use a TEI format [3]. However, we found that – at least from the perspective of the present project – the PML support of stand-off (layered) annotation is superior to that of TEI, mainly in the availability of tools and libraries. This concerns tasks such as validation, structural parsing, corpus management and searching. While some of those libraries do exist for TEI, many would have to be developed.
The only established alternative supporting layered annotation is the tabular format used by *EXMARaLDA* (Schmidt, 2009; Schmidt et al., 2011). Despite its rich set of options and other tools using the format, the format has some drawbacks in a scenario involving a language with rich morphology and free word order (see, e.g. (Hana et al., 2010)). Most importantly, the correspondences between the original word form and its corrected equivalents or annotations at other levels may be lost, especially for errors in discontinuous phrases. To allow for data exchange, the editor *feat* now supports import from several formats, including *EXMARaLDA*; it also allows export limited to the features supported by the respective format.

5.3. Transcription of manuscripts

The original documents are hand-written, usually the only available option, given that their most common source are traditional language courses and exams. They are transcribed using off-the-shelf editors supporting HTML (e.g., Microsoft Word or Open Office Writer). This means that the transcribers can use a tool they are familiar with and no

[2] http://ufal.mff.cuni.cz/jazz/pml/
[3] www.tei-c.org

technical training is required. A set of codes is used to capture some properties of the manuscript (variants, illegible strings, self-corrections; see (Štindlová, 2011, p. 106ff)). Some of these encodings are supported via macros of the editor. Because the set of text formatting features used by the transcribers is limited to a few standard options, there are no issues due to potentially incorrect HTML code exported from the text editor.

5.4. Annotation

The manual portion of error annotation is supported by *feat*,[4] an annotation tool we have developed. The annotator corrects the text on appropriate levels, modifies relations between elements on adjacent levels (by default all relations are 1:1) and annotates relations with error tags as needed. The context of the annotated text is shown both as a transcribed HTML document and – optionally – as a scan of the original document. Both the editor and the data format accommodate various approaches towards the process of multi-level annotation. Some annotators prefer to annotate by paragraphs, first annotating the whole paragraph on Level 1 first and then on Level 2, while others annotate by sentences annotating a sentence on both levels before moving to the next one. The tool is written in Java on top of the Netbeans platform.[5] Figure 2 shows the tool's user interface. It automatically synchronizes with *Speed*, the text management system – the user receives (whether an annotator, supervisor or adjudicator) the assigned documents into her Inbox, processes them and moves them to Outbox.

6. Conclusion

We have described the pre-processing work-flow of a learner corpus project, including the error annotation, with a focus on the computational and organizational issues. The annotation scheme has been tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results. At the moment, the annotation is proceeding towards the goal of 1 mil. tokens.

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.

7. Acknowledgements

We wish to thank our colleagues, namely Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, Hana Skoumalová and Svatava Škodová for many stimulating ideas, and members of our team of annotators for important feedback. We are also grateful to Karel Šebesta, for all of the above and for initiating and guiding this enterprise.

Our thanks are also due to our anonymous reviewers for interesting and helpful comments.

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 Competitiveness, funded by the European Structural Funds (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.

[4] http://purl.org/net/feat/

[5] http://platform.netbeans.org/

8. References

Sylviane Granger. 1998. The computer learner corpus: a versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–19. Addison Wesley Longman, London and New York.

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden, July. Association for Computational Linguistics.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2011. New and future developments in EXMARaLDA. In *Multilingual Resources and Multilingual Applications. Proceedings of GSCL Conference 2011 Hamburg.*

Thomas Schmidt. 2009. Creating and working with spoken language corpora in EXMARaLDA. In Verena Lyding, editor, *LULCL II: Lesser Used Languages and Computer Linguistics II*, pages 151–164.

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

Barbora Štindlová, Svatava Škodová, Alexandr Rosen, and Jirka Hana. 2012. Annotating foreign learners' Czech. In Markéta Ziková and Mojmír Dočekal, editors, *Slavic Languages in Formal Grammar. Proceedings of FDSL 8.5, Brno 2010*, pages 205–219, Frankfurt am Main. Peter Lang.

Barbora Štindlová. 2011. *Evaluace chybové anotace v žákovském korpusu češtiny [Evaluation of Error Mark-Up in a Learner Corpus of Czech]*. Ph.D. thesis, Charles University, Faculty of Arts, Prague.

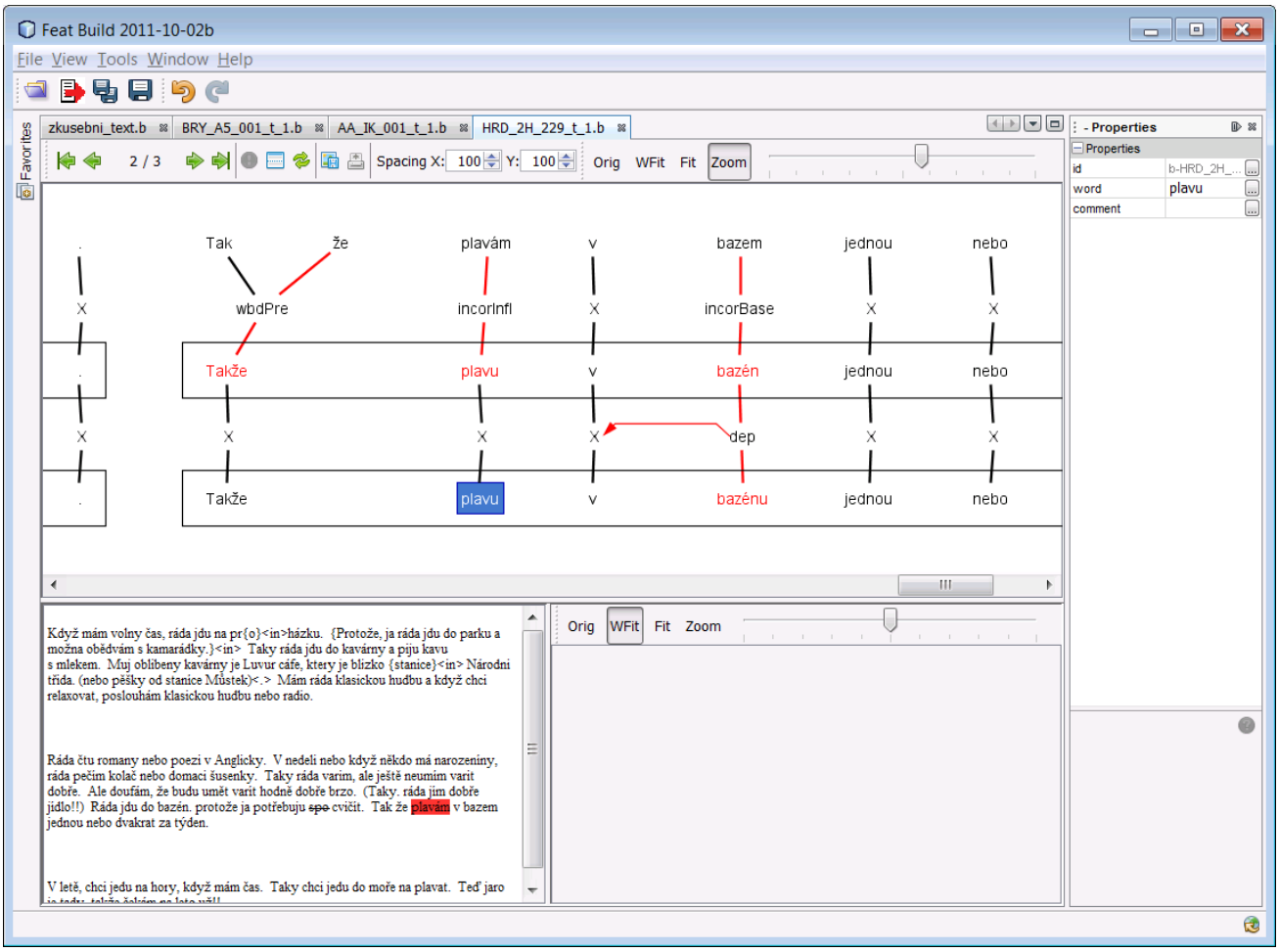

Figure 2: The user interface of *feat*