Annotating foreign learners' Czech

Barbora Štindlová and Svatava Škodová (Technical University, Liberec)
Alexandr Rosen and Jirka Hana (Charles University, FF & MFF, Prague)

1. Introduction

One of the challenges of contemporary corpus linguistics is the compilation and annotation of corpora consisting of texts produced by non-native speakers. In addition to morphosyntactic tagging and lemmatisation, such texts can be annotated by information relevant to the specific nonstandard use. Cases of deviant language use can be corrected and identified by a tag specifying the type of the error. Because of the properties of Czech, namely rich inflection, derivation, agreement, and a largely information-structure-driven constituent order, it is not straightforward to design an annotation scheme satisfying all requirements on the description of errors produced by non-native learners. Our proposal aims at an optimal solution that is still realistic given the annotation costs and the demands of the corpus users.

After an overview of issues related to learner corpora in §2 and a brief introduction to the project of a learner corpus of Czech in §3 we present the issues of annotation in §4 and the concept of our annotation scheme in §5, followed by a description of the annotation process in §6.

2. Learner corpus

A learner corpus, also called interlanguage or L2 corpus, is a computerised textual database of language as produced by foreign/second language (L2) learners (Leech 1998: xiv). It is a very powerful resource in the research of second language acquisition (SLA) and foreign language teaching (FLT). It serves as a repository of authentic data about a specific variety of natural language (Granger 2003), namely the *learner language,* in the context of SLA and FLT (Selinker 1972) often called *interlanguage* (IL).

Learner corpora allow to compare non-native and native speakers' language, or to compare interlanguage varieties. They can be studied on the background of national corpora, which helps to track various deviations from standard usage in the language of non-native speakers, such as frequency patterns – cases of overuse or underuse – or *foreign-soundingness* as compared with the language of native speakers. Recent studies have focused primarily on the frequency of use of separate language elements (e.g. Ringbom 1998), collocations

and prefabs (e.g. Nesselhauf 2005), lexical analysis and phrasal use (e.g. Altenberg & Tapper 1998), etc.

An error-tagged corpus can be subjected to *computer-aided error analysis* (CEA), which is not restricted to errors seen as a deficiency, but understood as a means to explore the target language and to test hypotheses about the functioning of L2 grammar. CEA also helps to observe meaningful use of non-standard structures of IL. Recent studies focus on lexical errors (Leńko-Szymańska 2004), wrong use of verbal tenses (Granger 1999) or phrasal verbs (Waibel 2008).

Learner corpora can be classified according to several criteria:

- Target language (TL): Most learner corpora cover the language of learners of English as a second or foreign language (ESL or EFL). The number of learner corpora for other languages is smaller but increasing.
- Proficiency in TL: Some gather texts of students at the same level, other include various levels. Most corpora focus on advanced students.
- Medium: Learner corpora can capture written or spoken texts. The latter are much harder to compile, thus less common.
- First (native) language (L1): The data can come from learners with the same L1 or with various L1s.
- Annotation: Most corpora provide at least one *target hypothesis* – a corrected (*emended*) version for each identified faulty expression. Additionally, the type of the error can be captured. The original and/or the emended text can also be annotated as standard texts – by POS, morphosyntactic categories, syntactic functions and structure.

Table 1 gives a brief summary of existing learner corpora.

3. CzeSL – a learner corpus of Czech

The corpus of Czech texts produced by non-native learners[1] (see also Hana et al. 2010, Štindlová 2011, Štindlová et al. 2011) was conceived as a part of a larger project AKCES, consisting of several acquisition corpora of Czech (Šebesta 2011). With its planned size of 2 million words, CzeSL (Czech as a Second

[1] We wish to thank other members of the project team, namely Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and Hana Skoumalová for their numerous stimulating ideas, acute insight and important feedback. We are especially grateful to Karel Šebesta, for all of the above and for initiating and guiding this enterprise.

The Project (CZ.1.07/2.2.00/07.0259) is a part of the operational programme Education for Competiveness, funded by the European Structural Funds (ESF) and the budget of the Czech Republic. The grant holder is the Technical University in Liberec in partnership with the Charles University in Prague and the Association of Teachers of Czech as a Foreign Language. The development of the annotation tool *feat* was partially supported by the GACR grant P406/10/P328.

Language) will become one of the largest learner corpora for languages other than English.

Corpus	**Size (in thous. of words)**	**First language**	**Target language**	**Proficiency level**	**Medium**	**Error annotation**
ICLE – International Corpus of Learner English	3,000	26	English	advanced	written	yes (1/4)
CLC – Cambridge Learner Corpus	35,000	130	English	all levels	written	yes (part)
LINDSEI – Louvain International Database of Spoken English	800	11	English	advanced	spoken	yes (part)
PELCRA – Polish Learner English Corpus	500	Polish	English	all levels	written	yes (part)
USE – Uppsala Student English Corpus	1,200	Swedish	English	advanced	written	no
HKUST – Hong Kong Univ. of Science and Technology Corpus of Learner English	25,000	Chinese	English	advanced	written	yes (200 th. words)
CHUNGDAHM – Chungdahm English Learner Corp.	131,000	Korean	English	all levels	written	yes (6.6 mil. words)
JEFLL – Japanese EFL Learner Corpus	700	Japanese	English	beginners	written	yes (part)
MELD – Montclair Electronic Language Learners' Database	1,000	16	English	advanced	written	no
MICASE – Michigan Corpus of Academic Spoken English	1,800	various	English	advanced	spoken	no
NICT JLE – NICT Japanese Learner English	2,000	Japanese	English	all levels	spoken	yes (part)
FALKO – Fehlerannotiertes Lernerkorpus	300	5	German	advanced	written	yes
FRIDA – French Interlanguage Database	200	various	French	intermediate	spoken	yes (2/3)
FLLOC – French Learner Language Oral Corpora	2,000	English	French	all levels	spoken	no
PiKUST – Poskusni korpus usvajanja slovenščine kot tujega jezika	40	18	Slovene	advanced	written	yes
ASU – ASU Corpus	500	various	Norwegian	advanced	written	no
TUFS – TUFS Learners' Corpus: Japanese	600 chars	various	Japanese	all levels	written	no (planned)

Table 1: Some currently available learner corpora (Štindlová 2011:63)

The corpus consists of the following subcorpora, distinguished by the native language of the authors of the texts:

- Slavic: mostly Russian or other East Slavic, followed by Polish. Other Slavic languages are only marginally represented.
- Other Indo-European: German, followed by French, English, Spanish etc.
- Non-Indo-European: a varied mix, with a slight majority of Vietnamese.
- Romani: text produced by the Czech Roma. It might be difficult to decide what their L1 is, yet the students often exhibit many traits typical for the process of acquisition of Czech as a second language.
- Czech: native speakers – pupils of elementary schools, included for comparison of Czech as L1 and L2.

In other aspects relevant for the use of learner corpora the corpus is designed to cover as much real data as possible:

- Although written texts prevail, each subcorpus has an oral part. Written texts are collected as manuscripts and transcribed according to a set of instructions (Štindlová 2011, 106ff) to preserve most of the information, including corrections made by the author.
- The corpus is based on texts covering all language levels according to the Common European Framework of Reference for Languages, from real beginners (A1 level) to advanced learners (level B2 and higher). Most learner corpora focus on one or two levels of proficiency, usually intermediate or advanced. However, there is no ambition to achieve a balanced mix of levels – due to the circumstances of the collection process most texts are of level B, with lower levels under-represented.
- The texts are elicited during various situations; they are not restricted to parts of written or oral examination as in most other learner corpora.
- Each text is equipped with background information, including sociological data about the learner (age, gender, L1, proficiency level, other languages, duration and conditions of language acquisition) and the specifics of the text and related circumstances (availability of reference tools, type of elicitation, temporal and size restrictions).

We expect the corpus to be used primarily in teaching. The corpus can:

- Provide data for the analysis of non-native speakers' competence in Czech. Such analysis can serve as a basis for improving the teaching process through a focus on actual problems students make. It will also help to tailor instructions and teaching materials to specific groups of learners (e.g., groups with different native languages or groups of different ages).

- Become a source of examples for particular phenomena or of complete authentic texts that can be used both in the classroom, in the production of teaching materials and in the instruction of future teachers of Czech as a second language.
- Be used to teach future teachers to identify, describe and explain particular error types.

From the very beginning of the project, the language data are used in language analysis in seminars on Czech as a second language at the Technical University of Liberec and at Charles University in Prague.

4. Annotation of learner corpora

In the context of second/foreign language acquisition, the learners' language is seen as an independent system, which should be analyzed in its entirety, with incorrect structures as an important part. Texts produced by non-native speakers can be annotated in two different ways:

1. Linguistic mark-up (e.g. part-of-speech tagging, morphological or syntactic annotation, lemmatization etc.). In most learner corpora, at least some parts are POS-tagged by tools originally developed for the analysis of the national language, cf. e.g. van Rooy & Schäfer (2003).
2. Error annotation, cf. e.g. Díaz-Negrillo & Fernández-Domínguez (2006).

Despite the time-consuming manual effort involved, the number of error-annotated learner corpora is growing. However, the level, extent and concept of error annotation differ widely. While 45% of learner corpora are error-annotated, only 7% use a comprehensive error taxonomy. The rest of error-annotated corpora include error tags in relation to a specific research goal, such as deficiencies in pronunciation (ISLE) or syntactic issues (CEDEL2), cf. Štindlová (2011:74).

4.1. Implicit error capture – reconstruction

The so-called reconstruction approach identifies errors only by their emendation. The advantage of this approach is the absence of an error classification scheme (Fitzpatrick & Seegmiller 2004) – the annotator does not need to learn any classification rules, which speeds up the annotation task and avoids misclassification. However, reconstruction without error labelling does not describe the error or substantiate the correction, which may obscure the annotation. Moreover,

without morphosyntactic mark up such a corpus is not easy to analyze by quantitative or statistical methods.

4.2. Explicit error capture – error classification

During the annotation process, errors are identified and categorized according to a fixed error taxonomy. Every error taxonomy reflects its theoretical background and may introduce a bias, but in comparison with the previous approach, corpora with explicitly marked errors are easier to search and analyze statistically. Error-tagged corpora may use the following taxonomies:

i. Linguistically-based taxonomies, with a varying degree of detail, ranging from very general categories (labelling an error in morphology, lexicon, syntax) to quite specific labels (auxilliary, passive, negation).
ii. Taxonomies based on (i) can be combined in multi-dimensional schemes – an error domain (grammar, lexicon, style) is complemented by error category (agglutination, diacritics, inflexion, gender) and word class (POS).
iii. Taxonomies based on a formal classification of superficial alternations of the source text, such as missing, redundant, faulty or incorrectly ordered element.

5. Annotation scheme

Our error annotation is primarily concerned with the acceptability of the grammatical and lexical aspects of the learner's language in a narrow sense, evaluated with respect to Standard Czech. However, we anticipate that future projects would annotate the corpus with less formal properties of speech, such as the degree of achievement of a communicative goal.

5.1. Annotation scheme as a compromise

Building an error-annotated learner corpus of Czech is a unique enterprise. In comparison with Czech, languages of the existing annotated learner corpora have simpler morphology and/or a more fixed word order. Therefore, many of the problems we have encountered were new and have not been addressed before.[2] Moreover, although the annotation scheme should be sufficiently informative and extensible, it should also be manageable and easily applicable, i.e. not too extensive. The resulting scheme and the error typology is a compromise

[2] To the best of our knowledge, there is only one learner corpus built for a Slavic language – PiKUST (Stritar 2009) – see Table 1. However, it is of a modest size of 35,000 words, and its error annotation is adopted from a Norwegian project ASK.

between the limitations of the annotation process and our research goals. Some of the issues involved, such as interference, interpretation, word order or style, do not have straightforward solutions:

Interference: Being no experts in L2 acquisition, the annotators cannot be expected to spot cases of linguistic interference of L1 or some other language known to the learner. Thus a sentence such as *Tokio je pěkný hrad* 'Tokio is a nice castle' is grammatically correct, but its author, a native speaker of Russian, was misled by 'false friends' and assumed *hrad* 'castle' as the Czech equivalent of Russian *gorod* 'town, city'.

Interpretation: For some types of errors, the problem is to define the limits of interpretation. The clause *kdyby citila na tebe zlobna* is grammatically incorrect, yet roughly understandable as 'if she felt angry at you'. In such cases the task of the annotator is interpretation rather than correction. The clause can be rewritten as *kdyby se na tebe cítila rozzlobená* 'if she felt angry at you', or *kdyby se na tebe zlobila* 'if she were angry at you'; the former being less natural but closer to the original. It is difficult to provide clear guidelines.

Word order: Czech constituent order reflects information structure. It may be hard to decide (even in a context) whether an error is present. The sentence *Rádio je taky na skříni* 'A radio is also on the wardrobe' suggests that there are at least two radios in the room, although the more likely interpretation is that among other things which happen to sit on the wardrobe, there is also a radio. The latter interpretation requires a different word order: *Na skříni je taky rádio*.

Style: Students often use colloquial expressions, usually without being aware of their status and the appropriate context for their use.[3] Even though these expressions might be grammatical, we emend them with their standard counterparts under the rationale that the intention of the student was to use a register that is perceived as unmarked.

5.2. Multi-level annotation

The optimal error annotation strategy is determined both by the goals and resources of the project and by the type of the language. A single-level scheme could be used for a specific narrowly defined purpose, such as the investigation of morphological properties of the learner language. However, given our goals, to apply the single-level scheme would be problematic. First of all, our corpus should be open to multiple research goals. Thus a restricted set of linguistic phenomena or a single level of analysis is not satisfactory. As a result, it is necessary to register successive emendations and to maintain links between the original and the emended form even when the word order changes or in cases of

[3] Diglossia is another important trait of Czech: its written form often differs from the spoken language.

dropped or added expressions. Another reason is the need to annotate errors spanning multiple forms, often in discontinuous positions.

In the ideal case, the annotator should be free to use an arbitrary number of levels to suit the needs of successive emendations, choosing from a set of linguistically motivated levels or introduce annotation tiers ad hoc. On the other hand, the annotator should not be burdened with theoretical dilemmas and the result should be as consistent as possible, which somewhat disqualifies a scheme using a flexible number of tiers. This is why we adopted a compromise solution with two levels of annotation, distinguished by formal but linguistically founded criteria to make the annotator's decisions easy.

Level 0 is the transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings). At Level 1, only forms wrong in isolation are treated. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. All other types of errors (valency, agreement, word order, etc.) are handled at Level 2.

5.3. Formalism

Annotated learner corpora sometimes use data formats and tools developed originally for annotating speech. Such environments allow for an arbitrary segmentation of the input and multilevel annotation of segments (Schmidt 2009). Typically, the annotator edits a table with columns corresponding to words and rows to levels of annotation. A cell can be split or more cells merged to allow for annotating smaller or larger segments. This way, phenomena such as agreement or word order can be emended and tagged (Lüdeling et al. 2005).

However, the tabular format is not quite suitable for languages with free word order and rich inflection, where a single form may be wrong in several domains at once: typography, orthography, morphosyntax, lexicon, word order. When cells in a table are split or merged, it may be difficult to keep track of links between successively emended forms. This is why we adopted a scheme where correspondences between successively emended forms are expressed explicitly.

Our annotation scheme has the shape of a graph consisting of three interconnected parallel levels, representing the original text (Level 0) and two levels of annotation (Level 1 and Level 2) – see Fig. 1. Usually every word in the input text corresponds to a node at every level. Nodes at neighbouring levels are usually linked 1:1, but words can be joined or split, deleted or added. Even discontinuous word sequences can be related across two neighbouring levels. In general, there is no restriction on the number of nodes participating in a single relation across neighbouring levels.

Each node may be assigned information in addition to the form of the word, such as lemma, morphosyntactic category and syntactic function. When-

ever the original form (or multiple forms) is emended, the links beween levels can be labelled by the error type.

Some error types, such as a wrong form due to violated rules of agreement or valency, may be complented by simple syntagmatic annotation, linking the error label with a different form, determining the correct version and further explaining the reason of the error. E.g. the subject or another form exhibiting the same agreement categories is the target of this type of link in case of a faulty finite verb form such as *jsme* in Fig. 1.

Corrections of morphosyntactic errors often result in secondary errors, as in *dívá se na americkém filmu* 'watches an American$_{loc}$ film$_{loc}$'. The adjective *americkém* correctly agrees with the head noun, but when the noun's case is corrected to accusative the case of the adjective must be corrected as well. Then multiple references are made: to the verb (or the preposition) as the case assigner for the noun, and to the noun as the source of agreement for the adjective, while the error of the form of the adjective is the result of another correction and it is marked as such.

An error can often be identified only in relation to a target hypothesis, while more than one such hypothesis may be available. So far, annotation using multiple target hypotheses exists as a theoretical possibility to be implemented in further stages of the project.

5.4. Error types

A typical learner of Czech makes errors all along the hierarchy of theoretically motivated linguistic levels, starting from the level of graphemics up to the level of pragmatics. For practical reasons we emend the input conservatively to arrive at a coherent and well-formed result, without any ambition to produce a stylistically optimal solution, refraining from too loose interpretation. Where a part of the input is not comprehensible, it is marked as such and left without emendation.

The taxonomy of errors is based on linguistic categories combined with a formal description of errors. Whenever possible, the error type is determined by comparing the original and the emended forms and/or by using results of a tagger and lemmatizer (see §6.4). So far, emendation is a manual task, although options of using an automatic spelling and grammar checker are investigated.

5.4.1. Errors at Level 1

Errors in individual word forms, treated at Level 1, include misspellings (also diacritics and capitalisation), misplaced word boundaries but also errors in inflectional and derivational morphology and unknown stems – fabricated or foreign words. Except for misspellings, all these errors are annotated manually. The result of emendation is the closest correct form, which can be further modi-

fied at Level 2 according to context, e.g. due to an error in agreement or semantic incompatibility of the lexeme. See Table 2 for a list of errors manually annotated at Level 1. The last three error types (*stylColl*, *stylOther* a *problem*) are used also at Level 2.

Error type	Description	Example
incorInfl	incorrect inflection	***pracovají*** *v továrně*; *bydlím s* ***matkoj***
incorBase	incorrect word base	*byla velká* ***tema***; *lidé jsou moc* ***mérný***; *musíš to* ***posvětlit***
fwFab	non-emendable, „fabricated“ word	*pokud nechceš slyšet* ***smášky***
fwNC	foreign word	*váza je na* ***Tisch***; *jsem v* ***truong***
flex	supplementary flag used with *fwFab* a *fwNC* marking the presence of inflection	*jdu do* ***shopa***
wbdPre	prefix separated by a space or preposition without space	*musím to* ***při pravit***; ***veškole***
wbdComp	wrongly separated compound	***český anglický*** *slovník*
wbdOther	other word boundary error	***mocdobře***; ***atak***; ***kdy koli***
stylColl	colloquial form	***dobrej*** *film*
stylOther	bookish, dialectal, slang, hyper-correct expression	*holka s hnědými* ***očimi***
problem	supplementary label for problematic cases	

Table 2: Errors at Level 1

The rule of “correct forms only” at Level 1 has a few exceptions: a faulty form is retained if no correct form could be used in the context or if the annotator cannot decipher the author’s intention. On the other hand, a correct form may be replaced by another correct form if the author clearly misspelled the latter, creating an unintended homograph with another form.

5.4.2. Errors at Level 2

Emendations at Level 2 concern errors in agreement, valency, analytical forms, pronominal reference, negative concord, the choice of aspect, tense, lexical item or idiom, and also in word order. For the agreement, valency, analytical forms, pronominal reference and negative concord cases, there is usually a correct form, which determines some properties (morphological categories) of the faulty form. Table 3 gives a list of error types manually annotated at Level 2. The automatically identified errors include word order errors and subtypes of the analytical forms error *vbx*.

Error type	Description	Example
agr	violated agreement rules	*to jsou **hezké** chlapci*; *Jana **čtu***
dep	unsatisfied valency requirements (complements), wrong forms of modifiers	*bojí se **pes***; *otázka **čas***; *mám plán **pracuju***
ref	error in pronominal reference	*dal jsem to jemu i **jejího** bratrovi*
vbx	error in analytical verb form or compound predicate	*musíš **přijdeš***; *kluci **jsou** běhali*; *začal **pracuje***
rflx	error in reflexive expression	*dívá na televizi*; *Pavel **si** raduje*
neg	error in negation	*žádný to **ví***; *půjdu **ne** do školy*
lex	error in lexicon or phraseology	*jsem **ruská***; *dopadlo to **přírodně***
use	error in the use of a grammar category	*včera **bude sněžit***; *pošta je **nejvíc blízko***; ***v** vodě*
sec	secondary error (supplementary flag)	*stará se o **našich** holčičkách*
stylColl	colloquial expression	*viděli jsme **hezký** holky*
stylOther	bookish, dialectal, slang, hyper-correct expression	*rozbil se mi **hadr***
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	disrupted construction	*kratka jakost vyborné ženy*
problem	supplementary label for problematic cases	

Table 3: Errors at Level 2

5.4.3. Example

The annotation scheme is illustrated in Fig. 1, using an authentic sentence (1), split in two parts for space reasons.[4]

(1)

Bojal	*jsme*	*se*	*že*	*ona*	*se*	*ne*	*bude*	*libila*	*slavnou*	*prahu,*
*feared$_{sg}$	AUX$_{pl}$	REFL	that	she	REFL	not	will	*like	famous	Prague
Bál	*jsem*	*se,*	*že*	*se*	*jí*	*nebude*		*líbit*	*slavná*	*Praha*

'I was afraid she would not like the famous (city of) Prague'

proto	*to*	*bylo*	*velmí*	*vadí*	*pro*	*mně*
therefore	it	was	*very	resent	for	me
protože	*to*	*by*	*mi*	*velmi*	*vadilo*	

'because I would be very unhappy about it.'

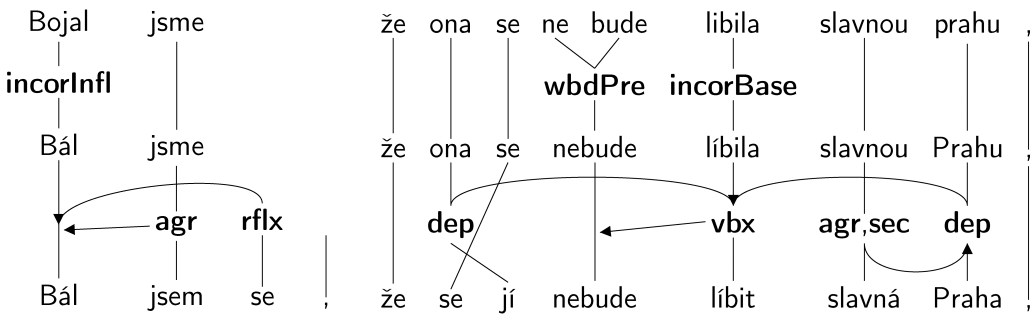
The three parallel strings of word forms represent the three levels, with links for corresponding forms. Most emendations are labelled with an error type. In the first part of the sentence two forms are labelled at Level 1 as errors in word base (*bojal – bál*, *libila – líbila*). The rest of Level 1 errors are purely orthographic: the negative particle *ne* is joined with the verb and the placename initial charac-

[4] The asterisked forms in the glosses mark forms that are incorrrect in any context. The line following the glosses gives the emended version of the sentence.

ter is capitalized (the error label is assigned automatically). At Level 2 another form is emended as an error in agreement (*jsme – jsem*) with reference to a form exhibiting the correct morphological category of singular number (*bál*). The missing reflexive particle is inserted with reference to the inherently reflexive verb and the comma is inserted without any label, because this type of error is identified automatically. The reflexive particle *se* is misplaced as a second position clitic (the label of a word-order error is assigned automatically). The pronoun *ona* – 'she' in the nominative case – is governed by the form *líbit se*, and should bear the dative case: *jí*, with reference to the head verb, which has changed its finite form *líbila* into the infinitive, because it is now a part of the analytical future tense, identified by the error type *vbx* and a link to the future auxiliary. The original accusative case of *Praha* is changed into nominative, again with a reference to the governing verb. The form of the adjective *slavnou* must be modified accordingly with an additional label *sec* as a secondary error.

In the second half of the sentence, there is only one Level 1 error in diacritics, but quite a few errors at Level 2. *Proto* 'therefore' is changed to *protože* 'because' as a lexical error. However, the main issue is the two finite verbal forms *bylo 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.

The prepositional phrase *pro mně* 'for me' is another complex issue. Its proper form is *pro mě* (homonymous with *pro mně*, but with 'me' bearing 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 second position clitic, following the conditional auxiliary (also a clitic) in the clitic cluster. The change from PP to the bare dative pronoun and the reordering are both properly represented, including the pointer to the head verb. What is missing is an explicit annotation of the faulty case of the prepositional complement, which is lost during the Level 1 – Level 2 transition; it is a price for a simpler annotation scheme with fewer levels.



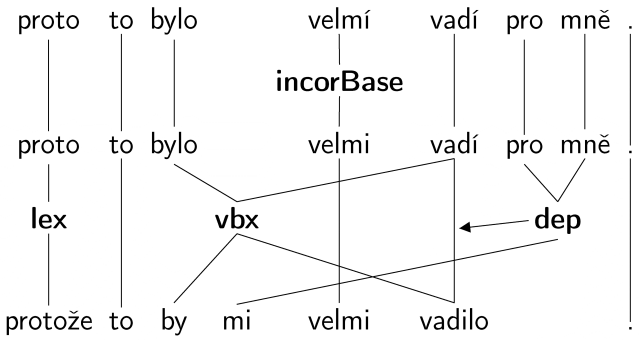


Figure 1: Annotation of a sample sentence

6. Annotation process

The whole annotation process proceeds as follows:

1. A handwritten document is transcribed using off-the-shelf tools (e.g. Open Office Writer or Microsoft Word). This means the transcribers can use a tool they are familiar with and no technical training is required. A set of codes is used to capture the author's corrections and other properties of the manuscript.
2. The transcript is converted into the annotation format, where Level 0 roughly corresponds to the tokenized transcript and Level 1 is set as equal to Level 0 by default. Both are encoded as PML (an XML-based format for structural linguistic annotation, see Pajas & Štěpánek 2006).
3. The annotator manually corrects the document and provides some information about errors using our annotation tool *feat*.
4. Error information that can be inferred automatically is added.

The manual portion of annotation is supported by the purpose-built annotation tool *feat* (http://ufal.mff.cuni.cz/~hana/feat.html). The annotator corrects the text on appropriate levels, modifies relations between elements (all relations are 1:1 by default) and annotates relations with error tags as needed. Manual annotation is followed by automatic post-processing, providing the corpus with additional information:

1. Level 1: lemma, POS and morphological categories for the individual forms (potentially ambiguous)
2. Level 2: lemma, POS and morphological categories (disambiguated)
3. Level 1: error types not assigned manually (by comparing the original and corrected strings), with the exception of lexical errors that involve lemma changes (e.g. **kadeřnička – kadeřnice* 'hair-dresser')
4. Level 2: morphosyntactic errors caused by violated agreement or valency (by comparing morphosyntactic tags at Level 1 and Level 2)

5. Formal error description: type of a spelling alternation, missing/redundant expression, wrong order

In the future, we plan to automatically tag errors in verb prefixes, inflectional endings, spelling, palatalisation, metathesis, etc.

7. Conclusion

Error annotation is quite resource-intensive but the result is very useful. Depending on the annotation scheme, the corpus user has access to detailed error statistics, which is difficult to obtain otherwise and which provides a reliable picture of the learners' interlanguage. This helps to adapt teaching methods and learning materials by identifying the most frequent error categories in accordance with the learner's proficiency level or L1 background.

The annotation process brings plentiful feedback, reflected in the annotation manual, training sessions and discussions in the web forum. The feedback has already helped to improve instructions to deal with thorny issues such as the uncertainty about the author's intended meaning, the inference errors, the proper amount of interference with the original, or the occurrence of colloquial language. In all of this, we need to make sure that annotators at least make some effort to handle similar phenomena in the same way.

Automatic tools could also be used in a pre-processing step to assist annotators, or for a fully automatic annotation of larger volumes of texts, which cannot be processed in a manual way due to limited resources. Some pilot studies have already been made, such as those applying different POS tagging methods to the original text, giving different results for faulty forms. By comparing these results a hypothesis about the error type might be proposed (Díaz-Negrillo et al. 2010). Another option is the employment of an automatic spelling and grammar checker to propose emended forms. Additionally, both the original and the emended versions of the text could be assigned syntactic functions and structure by a parser, possibly making use of some syntactic hints provided by reference links present with some error types (concord and valency).

References

Altenberg, Bengt & Marie Tapper. 1998. The use of adverbial connectors in advanced Swedish learner's written English. In S. Granger, *Learner English on Computer*. London: Longman. 80–93.

Carletta, Jean. 1996, Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*. 22, 249–254.

Corder, Pitt. 1981. *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

Díaz-Negrillo, Ana & Jesús Fernández-Domínguez. 2006. Error Tagging Systems for Learner Corpora. *Resla.* 19, 83–102.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera & Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*. 36, 139–154.

Fitzpatrick Eileen & Steve Seegmiller. 2004. The Montclair electronic language database project. In U. Connor & T. A. Upton*, Applied Corpus Linguistics: A Multidimensional Perspective.* Rodopi. 223–238.

Granger, Sylviane. 2003. Error–tagged Learner Corpora and CALL: A Promising Synergy. *CALICO journal.* 20, 465–480.

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

Leech, Geoffrey. 1998. Preface. In S. Granger, *Learner English on Computer*. London: Addison Wesley Longman. xiv–xx.

Leńko-Szymańska, Agnieszka. 2004. Demonstratives as anaphora markers in advanced learners' English. In G. Aston, S. Bernardini & D. Stewart, *Corpora and Language Learners.* Amsterdam: John Benjamins. 89–107.

Lüdeling, Anke, Maik Walter, Emil Kroymann & Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005.* Birmingham.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus.* Amsterdam: John Benjamins.

Pajas, Petr & Jan Štěpánek. 2006. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information*. Genoa, Italy: ELRA.

Ringbom, Håkan. 1998. Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger, *Learner English on Computer*. Harlow: Longman. 41–52.

Selinker, Larry. 1972. Interlanguage. *IRAL*, 10, 209–231.

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

Stritar, Mojca. 2009. Slovene as a Foreign Language: The Pilot Learner Corpus Perspective. *Slovenski jezik – Slovene Linguistic Studies.* 7, 135–152.

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

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

Štindlová Barbora, Svatava Škodová, Jirka Hana & Alexandr Rosen. 2011. CzeSL – an error tagged corpus of Czech as a second language. PALC 2011 – Practical Applications in Language and Computers, Lódż 13–15 April 2011. Selection of papers will be published in the publishing house Peter Lang in the series Łódź Studies in Language.

Van Rooy, Bertus & Lande Schäfer. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In D. Archer, R. Rayson, A. Wilson & T. McEnery, *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28–31 March 2003*. Lancaster: UCREL, Lancaster University. 835–844.

Waibel, Birgit. 2008. *Phrasal verbs. German and Italian learners of English compared*. Saarbrücken: VDM.