Combining Manual and Automatic Annotation of a Learner Corpus

Tomáš Jelínek[1], Barbora Štindlová[2], Alexandr Rosen[1], and Jirka Hana[3]

[1] Charles University in Prague, Faculty of Arts,
Institute of Theoretical and Computational Linguistics
[2] Technical University of Liberec, Faculty of Education,
Department of Czech Language and Literature
[3] Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Abstract. We present an approach to building a learner corpus of Czech, manually corrected and annotated with error tags using a complex grammar-based taxonomy of errors in spelling, morphology, morphosyntax, lexicon and style. This grammar-based annotation is supplemented by a formal classification of errors based on surface alternations. To supply additional information about non-standard or ill-formed expressions, we aim at a synergy of manual and automatic annotation, deriving information from the original input and from the manual annotation.

Key words: learner corpora, error annotation, Czech, morphology, syntax

1 Introduction

Texts produced by learners of a second or foreign language are a precious source of linguistic evidence for experts in language acquisition, teachers, authors of didactic tools, and students themselves. A corpus of such texts can be annotated by hand or by automatic tools in the usual ways, common in other types of corpora, i.a., by metadata, morphosyntactic categories and syntactic structure [12, 1] ??. However, the value of such texts is mainly in how they differ from the standard language. This information can be extracted by statistical comparisons with texts produced by native speakers, or by an explicit mark-up, offering hypotheses about the writer's intentions in the form of corrections and/or providing a classification of deviations from the standard. Methods and tools for dealing with Czech as a second language in this context have not been explored so far. A part of our learner corpus (about 300K tokens out of 2 million) is now manually (doubly) annotated with error tags and emended (corrected) forms, using a three-level annotation scheme with a complex grammar-based taxonomy of errors in spelling, morphology, morphosyntax, lexicon and style. This type of annotation is supplemented by a formal classification, e.g., an error in morphology can also be specified as being manifested by a missing diacritic or a wrong consonant change.

To assist the annotator and to supply additional information about deviations from the standard, we aim at a synergy of manual and automatic annotation, deriving information from the original input and from the manual annotation. Some methods interact with the annotator (e.g., a spell checker within the annotation editor marks potentially incorrect forms), or use results of manual annotation, including an automatic check for consistency and compliance with the annotation guidelines. After approval by the annotator's supervisor, some error tags are specified in more detail and more tags are added automatically. To assist the annotator even further, we experiment with methods of automatic emendation by a mildly context-sensitive spell checker and plan to use a grammar checker or a stochastic model to assign error annotation.

2 Annotating a learner corpus of Czech

Our learner corpus includes written texts[4] produced by non-native speakers of Czech at all levels of proficiency and equipped with meta-data about the authors, their learning history and the situation where the text was elicited. The authors are native speakers of Slavic, other Indo-European and some typologically distant languages, such as Chinese, Vietnamese or Arabic. A subcorpus includes texts written by pupils in primary school age with Romani background.

So far, the task of proposing a detailed methodology for teaching Czech as a foreign language has not received enough attention. Especially distinctions related to different target groups are not researched, even those most frequent and obvious between Slavic and non-Slavic students. Our project will help to change this by becoming a resource for research and design of teaching materials. At the same time, it will provide data helping to initiate and develop a systematic research of Czech as a foreign language.

2.1 Annotation scheme

Since most of the original texts are hand-written, the annotation process starts with their transcription according to detailed rules. A set of codes is used to capture the author's corrections and other properties of the manuscript.

The language of a learner of Czech may deviate from the standard in a number of aspects: spelling, morphology, morphosyntax, semantics, pragmatics or style. To cope with the multi-level options of erring in Czech and to satisfy the goals of the project, our annotation scheme answers the following requirements:

1. Preservation of the original text alongside with the emendations
2. Successive emendation
3. Ability to capture errors in single forms and discontinuous expressions
4. Syntactic relations for errors in agreement, valency, pronominal reference

[4] Spoken data are being collected but not yet transcribed or annotated.

To meet these requirements, we use a multilevel 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 annotation levels. 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.

Level 0 (L0) includes the transcribed input, where the words represent the original strings of graphemes, with some properties of the hand-written original preserved in the mark-up. Level 1 (L1) gives orthographical and morphological emendation of isolated forms; the sentence as a whole can be still incorrect. Level 2 (L2) treats all other deviations, resulting in a grammatically correct sentence. This includes errors in syntax (agreement, government), lexicon, word order, usage, style, reference, negation, or overuse/underuse. The corresponding forms at the neigbouring levels are linked, corrections are assigned error labels, and additional information such as POS tags and lemmas is added.

The whole annotation process proceeds as follows:

1. The transcript is converted into a format where L0 roughly corresponds to the tokenized transcript and L1 is set as equal to L0 by default. Both are encoded in an XML-based format [10].
2. The annotator manually corrects the document and provides some information about errors using the annotation tool *feat*.[5]
3. Automatic post-processing provides additional information about lemma, POS and morphological categories for emended forms.
4. Error information that can be inferred automatically is added.

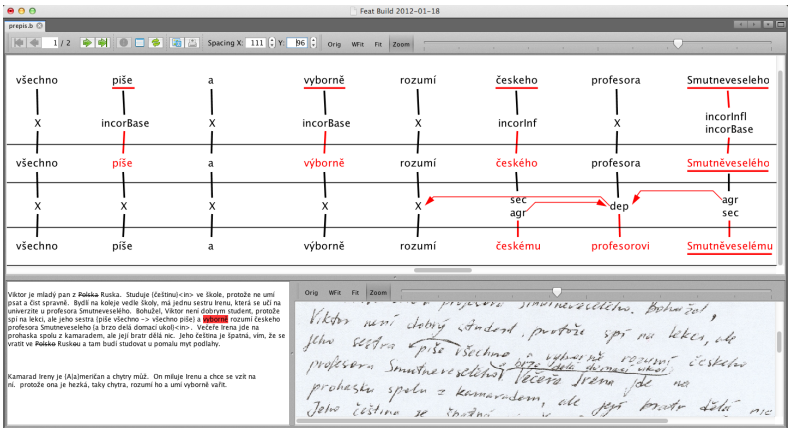

Fig. 1. A sample sentence in the *feat* annotation tool

[5] The tool *feat* (Flexible Error Annotation Tool is an environment for layered error annotation of learner corpora, see Fig. 1 [5].

2.2 Error taxonomy

Error taxonomies in learner corpora are often based on standard linguistic categories, see [2, 3, 8]. We use a similar approach, complemented by a classification of surface alternations. A single incorrect form is cross-classified as belonging to one or more types in each of the following two classes:

- grammar-based error types (spelling, morphology, word boundary, agreement, government, lexical issue, style, punctuation)
- formal error types (diacritics, capitalisation, metathesis, missing element)

For some types we identify the locus of the error, e.g., a morphological error is (manually) identified as an error in the stem or in the inflectional ending. Unlike the grammar-based types, the formal errors are more easily detectable by automatic tools. Yet the tools can detect also some of the grammar-based error types. Thus, errors can be identified in the following ways:

- manually
- automatically, by comparing the faulty and the emended forms
- automatically, by specifying a manually assigned error type in more detail, often using the word forms, their morphological tags or lemmas

Error type	Description	Example
`incorInfl`	incorrect inflection	*prac**ova**jí v továrně*; *bydlím s matko**j***
`incorBase`	incorrect word base	*lidé jsou m**é**rný*; *musíš to **po**světlit*
`fwFab`	non-emendable, „fabricated“ word	*pokud nechceš slyšet **smášky***
`fwNC`	foreign word	*váza je na **Tisch***; *jsem v **truong***
`flex`	with fwFab and fwNC: inflected	*jdu do **shopa***
`wbdPre`	word boundary: prefix or preposition	*musím to **při pravit***; ***veškole***
`wbdComp`	word boundary: compound	***český anglický** slovník*
`wbdOther`	other word boundary error	***mocdobře***; ***atak***; ***kdy koli***
`stylColl`	colloquial form	*dobr**ej** film*
`stylOther`	bookish, dialectal, hypercorrect	*holka s hnědým**i** očim**i***
`problem`	problematic cases	

Table 1. Grammar-based errors at Level 1

Grammar-based errors in individual word forms, treated at L1, include errors such as those in inflectional and derivational morphology, unknown stems (fabricated or foreign words) and misplaced word boundaries (see Table 1). All such errors are annotated manually. Emendations at L2 concern agreement, valency, analytical forms, pronominal reference, negation, the choice of aspect, tense, lexical items or idioms, and word order (see Table 2). Two or more errors may be present on one word form. Depending on the error type, two or more error tags may occur at one level or at both levels.

Error type	Description	Example
`agr`	violated agreement rules	*to jsou hez**ké** chlapci*; *Jana čt**u***
`dep`	error in valency	*bojí se p**es***; *otázka čas**[u]***
`ref`	error in pronominal reference	*dal jsem to jemu i je**jí**ho bratrovi*
`vbx`	error in analytical or compound verb form	*musíš přij**deš***; *kluci **jsou** běhali*
`rflx`	error in reflexive expression	*dívá **[se]** na televizi*; *Pavel s**i** raduje*
`neg`	error in negation	*žádný to **[ne]**ví*; ***půjdu ne** do školy*
`lex`	error in lexicon or phraseology	*jsem **r**usk**á***; *dopadlo to **přírodně***
`use`	error in the use of a grammar category	*pošta je **nejvíc blízko***
`sec`	secondary error	*stará se o naš**ich** holčičk**ách***
`stylColl`	colloquial expression	*viděli jsme hezk**ý** holky*
`stylOther`	bookish, dialectal, hypercorrect	*rozbil se mi **hadr***
`stylMark`	redundant discourse marker	*no*; *teda*; *jo*
`disr`	disrupted construction	*kratka jakost vyborné ženy*
`problem`	problematic cases	

Table 2. Grammar-based errors at L2

A doubly annotated sample (10,000 word forms) was evaluated for inter-annotator agreement to verify that the annotation scheme and taxonomy are sufficiently robust [13]. Higher agreement was found for formally well-defined categories, with satisfactory results even for those requiring subjective judgment.

3 Automatic extension of manual annotation

Manually emended and error-annotated text can be assigned additional information by automatic tools in the following three ways:

1. As far as the emended text approximates standard language, at least in grammatical correctness, a tagger/lemmatiser can be used with an error rate similar to that for standard texts [11].
2. Some manually assigned error tags were designed with the intention that they will be specified in more detail by an automatic tool.
3. Yet other tags are only assigned automatically.

3.1 Automatic addition of linguistic information

For practical reasons, especially for corpus searching, words in the corpus should be tagged with their morphological properties, including POS, case, etc. This information is added automatically.

L2 consists of correct Czech sentences only, so we can use standard tools [see, e.g., 6, 7] to assign each word a lemma and a tag from a standard morphological tagset [4]. L1 consists of correct Czech words, but they might be used with incorrect inflection, word order, etc. Therefore, using standard methods would

produce unreliable results. Instead, we combine the result of the morphological analysis with the properties of the word on L2 as follows:

- If the form is the same on both levels we use the tag/lemma from L2.
- If the forms are different, but have the same lemma (for the L1 forms suggested by the morphological analysis), then we use that lemma and the tags appropriate for it. For example, if the L1 form is *má* 'has' or 'my' and the L2 form is *mou* 'my', we assign *má* the lemma *můj* 'my'.
- If the L1 form's lemma is different from the lemma at L2, it receives all possible morphological tags. For example, *má* would be labeled both as a verb with the lemma *mít* 'to have' and as the possessive pronoun *můj* 'my'.

3.2 Automatic extension and modification of error annotation

Since the start, we assumed that some error types can be identified automatically. This is especially true for formal errors at L1, deducible by a simple comparison of the corresponding L1 and L0 forms, e.g., error in voicing or palatalization. Errors at L2 are more difficult to classify automatically, thus only a limited number of phenomena are tagged this way.

Automatic addition of formal error tags on L1 is based on the comparison of the original L0 form with the corrected L1 form. The manually assigned L1 tags cover the following three types of errors: a wrong form (`incor`), incorrect word boundaries (`wbd`), a neologism or a foreign word (`fw`). The automatically assigned errors are independent of these manual tags. For example, **chrozba/hrozba* 'threat' is manually annotated as `incorBase` (the *h/ch* error is in the stem), and **každécho/každého* 'every$_{masc.sg.gen/acc}$' as `incorInfl` (the *h/ch* error is in the *ého* ending). However, in both cases, the *h/ch* error is annotated as `formVcd1` and the correct *h* is incorrectly devoiced.[6]

The formal L1 error tags express the way in which an L1 form differs from the original incorrect L0 form. Most of these tags (such as "missing character", "switch error" or even "error in diacritics") only identify surface manifestations. However, a few error types are characterized by linguistic concepts, such as voicing assimilation or palatalization. It is the possibility of their automatic detection that puts them in the same class with the truly formal error types.

Table 3 provides examples of some currently handled automatically assigned errors on L1. Some errors affect only spelling with no change in pronunciation (capitalization, writing a wedge in *dě/tě/ně*, voicing assimilation, etc.). Other errors always affect pronunciation (vowel quantity, *e* epenthesis). Some errors might affect pronunciation in some contexts, but not others (writing *i/y*, the *c/k* substitution). We list only errors that actually occurred in real texts, using authentic examples, not every possible logical combination.

Most of the L2 error tags are assigned manually, because the variability of incorrect structures is too high to allow for a reliable automatic error tagging. Thus, only limited amount of information is added automatically:

[6] In Czech phonology, *h* and *ch* [x] act as voicing counterparts.

Error type	Error description	Example
`Cap0`	capitalization: incor. lower case	*evropě/Evropě*; *štědrý/Štědrý*
`Cap1`	capitalization: incor. upper case	*Staré/staré*; *Rodině/rodině*
`Vcd0`	voicing assimilation: incor. voiced	*stratíme/ztratíme*; *nabítku/nabídku*
`Vcd1`	voicing assimilation: incor. vcless	*zbalit/sbalit*; *nigdo/nikdo*
`VcdFin0`	word-final voicing: incor. voiceless	*kdyš/když*; *vztach/vztah*
`VcdFin1`	word-final voicing: incor. voiced	*přez/přes*; *pag/pak*
`Vcd`	voicing: other errors	*protoše/protože*; *hodili/chodili*
`Palat0`	missing palatalization (*k,g,h,ch*)	*amerikě/Americe*; *matkě/matce*
`Je0`	*je/ě*: incorrect *ě*	*ubjehlo/uběhlo*; *Nejvjetší/Největší*
`Je1`	*je/ě*: incorrect *je*	*vjeděl/věděl*; *vjeci/věci*
`Mne0`	*mě/mně*: incorrect *mě*	*zapoměla/zapomněla*
`Mne1`	*mě/mně*: incor. *mně, mňe, mňě*	*mněla/měla*; *rozumněli/rozuměli*
`ProtJ0`	protethic *j*: missing *j*	*sem/jsem*; *menoval/jmenoval*
`ProtJ1`	protethic *j*: extra *j*	*jse/se*; *jmé/mé*
`ProtV1`	protethic *v*: extra *v*	*vosm/osm*; *vopravdu/opravdu*
`EpentE0`	*e* epenthesis: missing *e*	*domček/domeček*
`EpentE1`	*e* epenthesis: extra *e*	*rozeběhl/rozběhl*; *účety/účty*

Table 3. Examples of automatically assigned errors on L1

- The reflexivity error tag (`rflx`) is added if another type of error concerns a reflexive pronoun.
- Manually assigned error tags for compound verb forms (`vbx`) are sub-divided as errors in: analytical verb forms (`cvf`), phase or modal verbs (`mod`), and copular predicates (`vnp`). The distinction uses lemmas and morphological tags.
- Tags marking deleted and inserted words are added (`odd`, `miss`).
- Word order corrections are tagged (`wo`). The annotator reorders the words as necessary, but does not tag the change. The label is assigned automatically to one or more misplaced forms using lemmas and tags on L2.

3.3 Automatic annotation checking

The system developed for automatic error tagging is also used for evaluating the quality of manual annotation, checking the result for tags that are probably missing or incorrect. For example, if an L0 form is not known by the morphological analyzer, it is likely an incorrect word which should have been emended. Also, if a word was emended and the change affects pronunciation, but no error tag was assigned, an `incorr` error tag is probably missing. This approach cannot find all problems in emendation and error annotation, but provides a good approximate measure of the quality of annotation and draws annotators' attention to potential errors.

4 Conclusion

We have discussed the schema and process of annotation of a learner corpus of Czech, showing that a combination of manual and automatic annotation can be successful. The corpus will be available soon for on-line queries, both in its error-annotated and merely transcribed parts. We also plan to explore options of (partially) automating annotation of learner corpora by presenting emended forms and error labels as suggestions to the annotator or as a raw result. The tools we consider include a context-sensitive spell checker, a grammar checker, and a stochastic model of error corrections. The tools are also tested on a corpus of transcribed speech [9]. Other plans include the use of syntactic annotation (functions and structure), and modifications and development of error taxonomy in response to users' feedback and the experience from annotating larger volumes of data.

Acknowledgments. This research was supported by the programme Education for Competiveness, funded by the European Structural Funds and the Czech government as a project no. CZ.1.07/2.2.00/07.0259. It was additionally co-funded by grant no. P406/10/P328 of the Grant Agency of the Czech Republic, and by the programme NAKI of the Czech Ministry of Culture, project no. DF11P01OVV013. We are also grateful to our colleagues for stimulating ideas and anonymous reviewers for helpful comments.

5 The References Section

Bibliography

[1] Dickinson, M.: Generating learner-like morphological errors in Russian. In: Proceedings of the 23nd International Conference on Computational Linguistics (COLING-10). Beijing (2010)

[2] Díaz-Negrillo, A., Fernández-Domínguez, J.: Error tagging systems for learner corpora. Resla 19, 83–102 (2006)

[3] Granger, S.: Error–tagged learner corpora and call: A promising synergy. CALICO journal 20, 465–480 (2003)

[4] Hajič, J.: Disambiguation of Rich Inflection: Computational Morphology of Czech. Karolinum, Charles University Press, Praha (2004)

[5] Hana, J., Rosen, A., Škodová, S., Štindlová, B.: Error-tagged Learner Corpus of Czech. In: Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV). Uppsala (2010)

[6] Jelínek, T.: Nové značkování v Českém národním korpusu [A new tagging system in the Czech National Corpus]. Naše řeč 91, 13–20 (2008)

[7] Jelínek, T., Petkevič, V.: Systém jazykového značkování korpusů současné psané češtiny [A system of linguistic markup of corpora of contemporary written Czech]. In: Petkevič, V., Rosen, A. (eds.) Korpusová lingvistika Praha 2011: 3 – Gramatika a značkování korpusů. Studie z korpusové lingvistiky, vol. 16, pp. 154–170. Ústav Českého národního korpusu, Nakladatelství Lidové noviny, Praha (2011)

[8] Lüdeling, A.: Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes, P., Walter, M. (eds.) Fortgeschrittene Lernervarietäten, p. 119–140. Niemeyer, Tübingen (2008)

[9] Nouza, J., Blavka, K., Boháč, M., Červa, P., Žďánský, J., Silovský, J., Pražák, J.: Voice technology to enable sophisticated access to historical audio archive of the Czech radio. In: Multimedia for Cultural Heritage, Communications in Computer and Information Science, vol. 247, pp. 27–38. Springer, Berlin / Heidelberg (2012)

[10] Pajas, P., Štěpánek, J.: XML-based representation of multi-layered annotation in the PDT 2.0. In: Hinrichs, R.E., Ide, N., Palmer, M., Pustejovsky, J. (eds.) Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006). pp. 40–47. Genova, Italy (2006)

[11] Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007. pp. 67–74. Association for Computational Linguistics, Praha, Czechia (2007)

[12] Van Rooy, B., Schäfer, L.: An evaluation of three POS taggers for the tagging of the Tswana Learner english corpus. In: D. Archer, R. Rayson, A.W..T.M. (ed.) Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK). p. 835–844. UCREL, Lancaster University, Lancaster (2003)

[13] Štindlová, B., Škodová, S., Hana, J., Rosen, A.: CzeSL – an error tagged corpus of Czech as a second language. In: Pęzik, P. (ed.) PALC 2011 – Practical Applications in Language and Computers, Lódż 13–15 April 2011. Łódź Studies in Language, Peter Lang (2012), to appear