CzeSL-MAN – a corpus of non-native speakers' Czech with manual annotation

The CzeSL-MAN corpus includes manually annotated transcriptions of essays written by non-native speakers of Czech, a subset of texts included in the CzeSL-SGT corpus http://dx.doi.org/10.13140/RG.2.1.1906.2487.

The annotation includes corrections of the original text (manual), error types (manual and automatic), and morphosyntactic categories and lemmas for the corrected text (automatic). For details see §2. Most texts are equipped with metadata about the author, the text and the annotation process. See §3 for details. The corpus is available for download from the LINDAT data repository (www.lindat.cz/), see §4.

For more about the *CzeSL* learner corpus and *AKCES*, the umbrella project of acquisition corpora, see http://utkl.ff.cuni.cz/learncorp/ and http://akces.ff.cuni.cz/. The sites include bibliography lists. For more recent papers, see, e.g., Rosen et al. (2014), Štindlová et al. (2013), Jelínek et al. (2012).

1 Choice of texts

The corpus includes transcripts of essays of non-native speakers of Czech, written in 2009–2013, the total of 645 texts written by native speakers of 32 different languages, including 298 doubly annotated texts. The texts contain 128 thousand word tokens, including 59 thousand doubly annotated tokens.

For the number of texts authored by students according to their first language and proficiency level in Czech see the tables below (IE = non-Slavic Indo-European, nIE = non-Indo-European, S = Slavic, ? = unknown).

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

Table 1: Texts by language group and proficiency level

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

Table 2: Doubly annotated texts by language group and proficiency level

The texts are 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 are replaced by `QQQ`. Unreadable characters or words are transcribed as `XXX`.

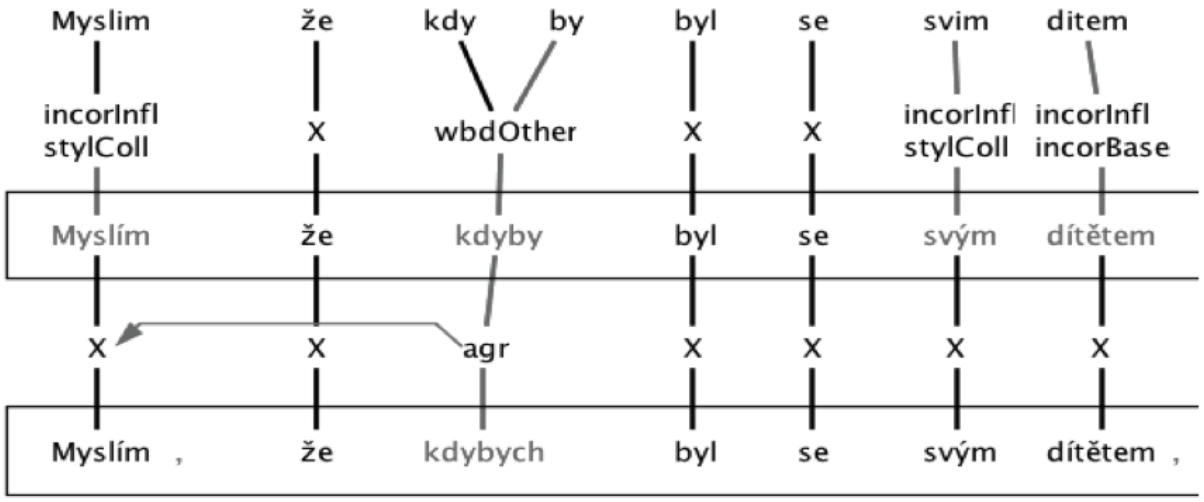

Figure 1: Example of the three-tier error annotation scheme

2 Annotation

The annotation scheme consists of three interconnected tiers – see Fig. 1, glossed in (1):

- Tier 0 – anonymised transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings)
- Tier 1 – forms incorrect in isolation are fixed. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole
- Tier 2 – handles all other types of errors (valency, agreement, word order, etc.)

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

Errors in individual word forms, treated at Tier 1, include misspellings (also diacritics and capitalisation), misplaced word boundaries but also errors in inflectional and derivational morphology and unknown stems – made-up or foreign words. The result of the correction is the closest correct form, which can be further modified at Tier 2 according to context, e.g. due to an error in agreement or semantic incompatibility. See Table (1) for a list of errors manually annotated at Tier 1. The last three error types (*stylColl*, *stylOther* and *problem*) are used also at Tier 2.

The rule of "correct forms only" at Tier 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.

Corrections at Tier 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 and which is referred to in the annotation. Table 4 gives a list of error types manually annotated at Tier 2. The automatically identified errors include word order errors and subtypes of the analytical forms error *vbx*.

The correspondences between successively emended forms are explicitly expressed. Nodes at neighbouring tiers are usually linked 1:1, but words can be joined (*kdy by* as in Fig. 1) or split, deleted or added. These relations can interlink any number of potentially non-contiguous words across the neighbouring tiers. The type of error is specified as a label at the link connecting the incorrect form at a lower tier with its corrected form at a higher tier.

Manual annotation is supported by the purpose-built annotation tool *feat*[1].

Corrected forms are tagged with morphosyntactic categories and lemmas using standard tools. Each word is assigned a lemma and a tag from a standard morphological tagset Hajič (2004). Instead of a fully disambiguated

[1] See http://purl.org/net/feat.

Error type	Description	Example
incorInfl	incorrect inflection	*pracovají* v továrně; bydlím s *matkoj*
incorBase	incorrect word base	lidé jsou moc *mérný*; musíš to *posvětlit*
fwFab	non-emendable, made-up word	pokud nechceš slyšet *smášky*
fwNC	foreign word	váza je na *Tisch*; jsem v *truong*
flex	supplementary flag used with fwFab and 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	holka s *hnědými očimi*
problem	supplementary label for problematic cases	

Table 3: Manually assigned errors at Tier 1

Error type	Description	Example
agr	violated agreement rules	to jsou *hezké* chlapci; Jana *čtu*
dep	error in valency	bojí se *pes*; otázka *čas*
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
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	pošta je *nejvíc blízko*
sec	secondary error (supplementary flag)	stará se o *našich holčičkách*
stylColl	colloquial expression	viděli jsme *hezký* holky
stylOther	bookish, dialectal, slang, hyper-correct expression	zvedl se mi *kufr*
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	disrupted construction	*kratka jakost vyborné ženy*
problem	supplementary label for problematic cases	

Table 4: Manually assigned errors at Tier 2

tag and lemma, T1 is tagged using potentially ambiguous morphological analysis of isolated forms in combination with the tag and lemma assigned at T2 as follows:

- If the forms at both tiers are identical, the tag and lemma assigned at T2 is used.
- If the forms are different, but their lemmas are identical, then that lemma and the appropriate tags are used. For example, if the T1 form is *má* 'has' or 'my' and the T2 form is *mou* 'my', we assign *má* the lemma *můj* 'my'.
- If the T1 form's lemma is different from the lemma at T2, the T1 form 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 with the lemma *můj* 'my'.

The Czech morphological tagset is described at http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html or http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/docc0pos.pdf.

The automatically assigned 'formal' errors complement manual tags as an additional dimension of annotation. 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 correct *h* is incorrectly devoiced, thus the *h/ch* error is annotated as *formVcd1*.[2]

The formal T1 error tags express the way in which a T1 form differs from the original incorrect T0 form. Table 5 provides examples of some currently handled automatically assigned errors on T1, see http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1_en.html for a full list. Some errors affect only spelling with no change

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

Error type	Error description	Example
Cap0	capitalisation: incor. lower case	*evropě/Evropě*; *štědrý/Štědrý*
Cap1	capitalisation: 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 palatalisation (*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 5: Examples of automatically assigned errors at Tier 1

in pronunciation (capitalisation, diacritics 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).

3 Metadata

Metadata are available for most of the texts (e.g., in four documents, the first language of the author is unknown): 15 items about the author of the text and 15 items about the text itself. For a list of all attributes and values in Czech and English see http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html. The content of the individual items is explained below in §3.2 and §3.1. Missing items are represented as empty elements. Some attributes may include multiple values, delimited by vertical bar ("|"). The items are included in the `*.meta.xml` files.

3.1 Data about the task

- `t_id` – identification; a single value: character string, e.g. `TOU_H305_442`
- `t_date` – date of the text collection; a single value: date in the format `YYYY-MM-DD`
- `t_medium` – medium of the text; one of the values:
  - `manuscript`; `pc`
- `t_limit_minutes` – time limit for writing the text in minutes; one of the values:
  - `10`; `15`; `20`; `30`; `40`; `45`; `60`; `other`; `none`
- `t_aid` – permitted aid; one or more of the values:
  - `ano`; `dictionary`; `textbook`; `other`; `none`
- `t_exam` – was the text written as a part of an exam?; one or more of the values:
  - `yes`; `interim`; `final`; `n/a`
- `t_limit_words` – size limit in the assignment; one of the values:
  - `20`; `20-`; `25`; `30`; `35-`; `40`; `40-`; `50`; `50-`; `60`; `60-`; `70`; `70-`; `80`; `80-`; `90`; `90-`; `100`; `100-`; `120`; `120-`; `150`; `150-`; `170`; `180`; `200`; `200-`
- `t_title` – title of the essay; one or more values: character string, e.g. `Událost, která změnila můj život`

- t_topic_type – type of the topic; one of the values:
  - general; specific
- t_activity – activity before writing the text; one of the values:
  - exercise; discussion; visual; vocabulary; other; none
- t_topic_assigned – topic specified in the assignment; one of the values:
  - multiple choice; specified; free; other
- t_genre_assigned – genre specified in the assignment; one of the values:
  - free; specified
- t_genre_predominant – genre predominant in the resulting text; one of the values:
  - informative; descriptive; argumentative; narrative
- t_words_count – actual number of words; a single value: integer
- t_words_range – range of the actual number of words; one of the values:
  - -50; 100-149; 150-199; 200-; 50-99

3.2 Data about the author of the text (the student)

- s_id – identification; a single value: character string, e.g. TOU_H305
- s_sex – sex; one of the values:
  - m – male
  - f – female
- s_age – age; a single value: integer
- s_age_cat – age category; one of the values:
  - 6-11; 12-15; 16-
- s_L1 – first language; one of the values: two-character code according to the standard ISO 639-1, e.g. sq (Albanian); or three-character code ISO 639-3 if necessary, e.g. xal (Kalmyk) or bem (Bemba), see http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes and http://en.wikipedia.org/wiki/ISO_639-3.
- s_L1_group – language group according to the first language; one of the values:
  - IE – Indo-European non-Slavic
  - nIE – non-Indo-European
  - S – Slavic
- s_other_langs – knowledge of other languages; one or more of the values:
  ISO code (see s_L1)
- s_cz_CEF – proficiency in Czech at the time of writing; one of the values:
  - A1; A1+; A2; A2+; B1; B2; C1; C2
- s_cz_in_family – knowledge of Czech in the family; one or more of the values:
  - mother; father; partner; sibling; 3 (3 family members); other; nobody
- s_years_in_CzR – length of stay in the Czech Republic in years; one of the values:
  - -1; 1; -2; 2-
- s_study_cz – past or present study; one or more of the values:

  - `1to1` – individual tutoring
  - `paid`
  - `TY` – self-study
  - `university`
  - `foreign`
  - `primary-secondary`
  - `other`
- `s_study_cz_months` – length of study of Czech in months; one of the values:
  - `-3`; `3-6`; `6-12`; `12-24`; `24-36`; `36-48`; `48-60`; `60-`
- `s_study_cz_hrs_week` – intensity of study of Czech in hours per week; one of the values:
  - `-3`; `5-15`; `15-`
- `s_textbook` – textbook used in the past or present by the student; one or more of the values:
  - `BC` – Basic Czech
  - `CC` – Communicative Czech
  - `CE` – Čeština pro ekonomy
  - `CMC` – Chcete mluvit česky?
  - `CpC` – Čeština pro cizince
  - `ECE` – Easy Czech Elementary
  - `NCSS` – New Czech Step by Step
  - `other`
- `s_bilingual` – bilingual; one of the values:
  - `yes`; `no`

3.3 Data about the annotation

At the moment, only two items are available: the ID of the annotator and the supervisor.

4 Format of the texts

The annotation1 and annotation2 folders contain two parallel annotations of the same set of documents. Not all documents were annotated twice, therefore the annotation2 folder contains a proper subset of the documents in annotation1 folder. Each document consist of several related files:

- *.jpg - scan of the handwritten original (not part of the distribution, for privacy reasons)
- *.html - transcription of the handwritten original (anonymized)
- *.meta.xml - metainformation about the document, its author and annotation
- *.w.xml - tokenized text (T0)
- *.a.xml - T1 annotation of the text (roughly addressing word-level errors, e.g., spelling, word-boundaries)
- *.b.xml - T2 annotation of the text (roughly addressing contextual errors, e.g., agreement or wordorder)

5 Acknowledgment

The work was supported in 2009–2012 from the European Structural Funds grant *Innovation in the Education of Czech as a Second Language*, reg. no. CZ.1.07/2.2.00/07.0119 and *PRVOUK*, the research funding programme at Charles University: *P10 – Linguistics, Acquisition and Development of Linguistic and Communicative Competence in Selected Communities of the Czech Republic* and from the project *Czech National Corpus*, supported by the Ministry of Education of the Czech Republic as a part of the *Projects of Large Infrastructures for Science, Research and Innovations* (2012-2015, project no. LM2011023).

References

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.

Jelínek, T., Štindlová, B., Rosen, A., & Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, 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.

Rosen, A., Hana, J., Štindlová, B., & 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.

Štindlová, B., Škodová, S., Hana, J., & Rosen, A. (2013). A learner corpus of Czech: current state and future directions. In S. Granger, G. Gilquin, and F. Meunier, editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve. Presses Universitaires de Louvain.

Alexandr Rosen, 7 October 2015