CzeSL-man v1 searchable – a corpus of non-native Czech with manual error annotation in a simplified tiered scheme

CzeSL-man v1 searchable is a corpus that includes transcripts of essays written by non-native speakers of Czech. It is a manually annotated subset of texts included in the *CzeSL-SGT* corpus, corresponding to a subset of texts written by foreign learners included in *CzeSL-man v0*. Content-wise, *CzeSL-man v1 searchable* is identical with *CzeSL-man v1 downloadable*.

The manual error annotation is a simplified version of the two-tier (2T) annotation scheme designed for the *CzeSL* project (see §2 below). The annotation includes corrections of the source text – the target hypothesis (manual), error types (manual and automatic), morphosyntactic categories and lemmas for the corrected text (automatic), and dependency syntactic structure and functions for the corrected text (automatic). Most texts are equipped with metadata about the author and the text.

This corpus is available for on-line searching using *KonText*, the search interface of the Czech National Corpus.[1] The corpus differs from both *CzeSL-man v0* and *CzeSL-man v1 downloadable* in two aspects: (i) there are no texts with alternative error annotation: each text is annotated by a single annotator (just one version of each doubly annotated text is included), and (ii) the two-tier annotation scheme is radically modified to fit the token-based setup of the search tool. Apart from that, the content and metadata are identical to *CzeSL-man v1 downloadable* and the search options to those of *CzeSL-SGT*.

For more about the *CzeSL* learner corpus project, including an overview of all releases of the *CzeSL* learner corpus with links to the search or download options, see http://utkl.ff.cuni.cz/learncorp/ and Rosen et al. (2020).

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. The texts contain 128 thousand word tokens.[2]

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

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`.

[1] https://kontext.korpus.cz/first_form?corpname=czesl-man

[2] The number of tokens in the corpus reported by the *KonText* tool is slightly lower (124 thousand), because what is counted are tokens in the corrected version of the corpus.

	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

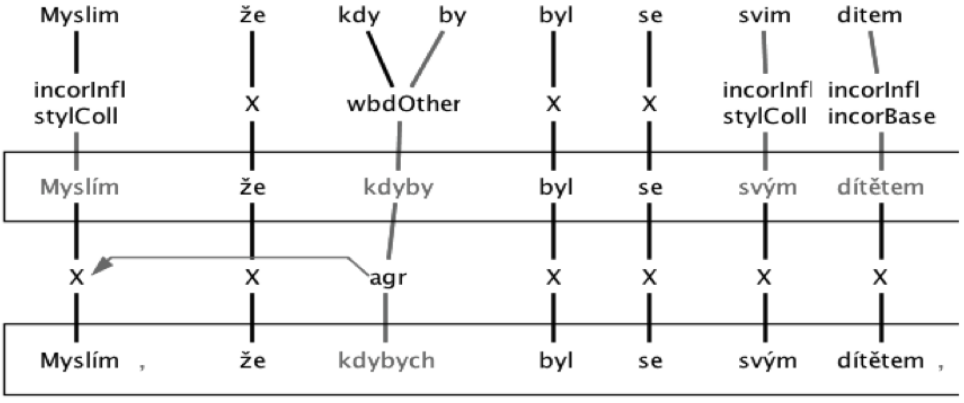

Figure 1: Example of the two-tier error annotation scheme

2 Annotation

2.1 The two-tier annotation scheme

The original 2T 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. This tier is not represented in *CzeSL-man v1 searchable*.
- 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, . . . '

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. In this corpus, corrections at Tier 1 are incorporated in the Tier 2 corrections. Table 2 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*. Correspondences between successively corrected forms are explicitly expressed. Forms 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.

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 *nemocných dětech*
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 2: Manually assigned errors at Tier 2

2.2 The two-tier scheme simplified

The main feature in the annotation of this release is the reversal of the source text and its annotation. The target hypothesis at T2, the corrected text, is assumed to be the basis for the annotation. The tokens of this corpus represent the words at T2. The original text is added as annotation of the T2 tokens. Each token of the corrected text receives its corresponding T0 form and a T2 error label as attributes. This annotation discards any T1 corrections and error tags, and simplifies other than 1:1 links between tokens at T0 and T2.

2.3 Linguistic annotation

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). 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 radical simplification of the two-tier error annotation scheme allowed for parsing the T2 target hypothesis in a way similar to some other Czech corpora searchable in *KonText*, such as *SYN2015*. For a list of syntax-related attributes assigned to each token see Table 4. For a more detailed description see https://wiki.korpus.cz/doku.php/en:pojmy:syntakticka_analyza.

2.4 The annotation in corpus searches

See Table 3 for a list of attributes representing the basic error and morphosyntactic annotation in this release. The syntax-related attributes are listed in Table 4. As in *CzeSL-SGT*, dynamic attributes derived from the morphosyntactic tags can be used in queries and visualization of the results, see Table 5.

`err`	T2 error tag of `word`, if any
`word0`	T0 form (the source)
`lemma0`	Lemma of `word0`; same as `word0` if the form is not recognized
`tag0`	Morphological tag of `word0`; if the form is not recognized: `X@----------`
`word`	T2 corrected form; same as `word0` if determined as correct
`lemma`	Lemma of `word`
`tag`	Morphological tag of `word`

Table 3: Token attributes used in *KonText* for the morphosyntactic and error annotation of *CzeSL-man v1 searchable*

proc	Disambiguation step responsible for the analysis
afun	Syntactic function
parent	Relative pointer to parent
eparent	Relative pointer to effective parent
prep	Preposition as parent
p_tag	Parent tag
p_lemma	Parent lemma
p_afun	Syntactic function of the parent
ep_tag	Effective parent tag
ep_lemma	Effective parent lemma
ep_afun	Syntactic function of effective parent
lc	Lowercase T2 word
lemma_lc	Lowercase T2 lemma
p_k	Parent category (POS)
p_c	Parent case

Table 4: Syntax-related attributes used in *KonText* for *CzeSL-man v1 searchable*

k0, k	Word class (position 1 of the tag)
s0, s	Detailed word class (position 2 of the tag)
g0, g	Gender (position 3 of the tag)
n0, n	Number (position 4 of the tag)
c0, c	Case (position 5 of the tag)
p0, p	Person (position 8 of the tag)

Table 5: Dynamic attributes used in *KonText* for *CzeSL-man v1 searchable*

3 Metadata

Metadata are represented as in *CzeSL-SGT* and are in English even in *KonText*. There are 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.

4 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, 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.

Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S., & Štindlová, B. (2020). *Compiling and annotating a learner corpus for a morphologically rich language – CzeSL, a corpus of non-native Czech*. Karolinum, Charles University Press, Praha.

Alexandr Rosen, 9 November 2020