CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation

Alexandr Rosen

May 26, 2014

The CzeSL-SGT corpus (***Cze**ch as a **S**econd **L**anguage with **S**pelling, **G**rammar and **T**ags*) includes transcriptions of essays written by non-native speakers of Czech, extending the "foreign" (ciz) part of the CzeSL-plain corpus (http://wiki.korpus.cz/doku.php/cnk:czesl-plain) by texts collected in 2013. For more details see §1.

Word forms are tagged by word class, morphological categories and base forms (lemmas). Some forms are corrected and the resulting texts are tagged again. Original and corrected forms are compared and error labels are assigned. The annotation is assigned automatically, which necessarily results in some inaccuracy and error rate. For details see §2.

Most texts are equipped with metadata about the author and the text. See §3 for details.

The corpus is available either for on-line searching using the search interface of the Czech National Corpus (http://korpus.cz), see §3.3, or for download as a whole from the LINDAT data repository (http://www.lindat.cz), see §3.4.

For more about the *CzeSL* learner corpus and *AKCES*, the more general project of aquisition 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

- Transcripts of essays of non-native speakers of Czech, written in 2009–2013
- Extends the **ciz** part of the CzeSL-plain corpus (http://ucnk.ff.cuni.cz/czesl-plain.php) by texts written in 2013
- 8,617 texts by 1,965 different authors with 54 different first languages
- 111 thousand sentences, 958 thousand words, 1 148 thousand tokens
- Without transcription markup, with the final author's version restored

2 Annotation

Each token is labelled by the following attributes:

- `word` – original word form
- `lemma` – lemma of `word`; same as `word` if the form is not recognized
- `tag` – morphological tag of `word`; if the form is not recognized: `X@----------`
- `word1` – corrected form; same as `word` if determined as correct
- `lemma1` – lemma of `word1`
- `tag1` – morphological tag of `word1`
- `gs` – information on whether the error was determined as a spelling (`S`) or grammar (`G`) error; for grammar errors, `word` is mostly recognized
- `err` – error type, determined by comparing `word` and `word1` http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1_en.html.

word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@----------	Ten	ten	PDYS1----------	S	Quant1
pes	pes	NNMS1-----A----	pes	pes	NNMS1-----A----		
míluje	míluje	X@----------	miluje	milovat	VB-S---3P-AA---	S	Quant1
svécho	svécho	X@----------	svého	svůj	P8MS4----------	S	Voiced
kamarada	kamarada	X@----------	kamaráda	kamarád	NNMS4-----A----	S	Quant0
-	-	Z:----------	-	-	Z:----------		
člověka	člověk	NNMS2-----A----	člověka	člověk	NNMS4-----A----		
.	.	Z:----------	.	.	Z:----------		

Table 1: Annotation of a sample sentence

In addition to the attributes listed above, the search interface of the Czech National Corpus offers "dynamic" attributes, derived from some positions of `tag` and `tag1`. They can be used in queries to specify values of morphological categories without regular expressions, to stipulate identity of these values in two or more forms to require grammatical concord or to compare values of a category for `word` and `word1`. These attributes are available for the following categories of the original and the corrected form:

- `k`, `k1` – word class (position 1 of the tag)
- `s`, `s1` – detailed word class (position 2 of the tag)
- `g`, `g1` – gender (position 3 of the tag)
- `n`, `n1` – number (position 4 of the tag)
- `c`, `c1` – case (position 5 of the tag)
- `p`, `p1` – person (position 8 of the tag)

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.

3 Metadata

Metadata are available for the new and for most of the old texts: 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 numbers of documents, listed according to specific attribute values, are given here: http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta_en.html. The content of the individual items is explained below in §3.1 and §3.2.

Some or even all items may be missing for some texts: identification of the author is present in 96.7% texts, the first language in 96.3% textů. Missing items are represented as empty values. Some attributes may include multiple values, delimited by vertical bar ("|").

Metadata are available in Czech and English. The Czech National Corpus site offers the Czech version, while the LINDAT data repository offers the entire corpus using their English version.

All the items are attributes of the `doc` element.

3.1 Data about the author of the text (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

3.2 Data about the text

- 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

3.3 Searching the corpus

The corpus can be searched from the unified search interface of the Czech National Corpus (korpus.cz). The czesl-sgt corpus is one of SYNCHRONIC WRITTEN CORPORA, in the category SPECIALIZED. With the QUERY TYPE set to BASIC and no other specifications, a string entered in the QUERY field returns sentences where the form occurs in the original, uncorrected text. For more advanced queries, including references to tags, lemmas, error types, corrected forms and metalanguage attributes, the QUERY TYPE should be set to CQL and/or the settings in SPECIFY QUERY ACCORDING TO THE META-INFORMATION modified. For general help on using CQL see http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying.

3.4 Format of the texts

This part is relevant for whole texts, available from the LINDAT repository.

Data about the text and sentences are represented as XML entities with corresponding attributes. The text itself is represented in tab-delimited columns, in the order shown in Table 1. A sample text is shown below:

```
<doc t_id="UJA2_PH_003" t_date="2010-12-21" t_medium="manuscript" t_limit_minutes="45" t_aid="none"
t_exam="yes|interim" t_limit_words="25" t_title="E-mail kamarádce/kamarádovi" t_topic_type="general"
t_activity="" t_topic_assigned="specified" t_genre_assigned="specified"
t_genre_predominant="informative" t_words_count="30" t_words_range="-50" s_id="UJA2_PH" s_sex="m"
s_age="17" s_age_cat="16-" s_L1="vi" s_L1_group="nIE" s_other_langs="" s_cz_SER="A1"
s_cz_in_family="" s_years_in_CzR="" s_study_cz="university"
s_study_cz_mesice="" s_study_cz_hrs_week="15-" s_textbook="NCSS" s_bilingual="no">
<s id="1">
mám      mít     VB-S---1P-AA--- mám     mít     VB-S---1P-AA---
dobře    dobře   Dg-------1A---- dobře   dobře   Dg-------1A----
.        .       Z:------------- .       .       Z:-------------
</s>
<s id="2">
V        v       RR--4---------- V       v       RR--4----------
neděli   neděle  NNFS4-----A---- neděli  neděle  NNFS4-----A----
dival    dival   X@------------- díval   dívat   VpYS---XR-AA--- S       Quant0
jsem     být     VB-S---1P-AA--- jsem    být     VB-S---1P-AA---
se       se      P7-X4---------- se      se      P7-X4----------
na       na      RR--6---------- na      na      RR--6----------
televizi televize        NNFS6-----A---- televizi        televize        NNFS6-----A----
a        a       J^------------- a       a       J^-------------
ukliděl  ukliděl X@------------- uklízel uklízet VpYS---XR-AA--- S       Quant0|Caron1
jsem     být     VB-S---1P-AA--- jsem    být     VB-S---1P-AA---
.        .       Z:------------- .       .       Z:-------------
</s>
<s id="3">
Ano      ano     TT------------- Ano     ano     TT-------------
přijdu   přijít  VB-S---1P-AA--- přijdu  přijít  VB-S---1P-AA---
se       se      P7-X4---------- se      se      P7-X4----------
tebe     ty      PP-S2--2------- tebe    ty      PP-S2--2-------
do       do      RR--2---------- do      do      RR--2----------
kina     kino    NNNS2-----A---- kina    kino    NNNS2-----A----
a        a       J^------------- a       a       J^-------------
taky     taky    Db------------- taky    taky    Db-------------
mám      mít     VB-S---1P-AA--- mám     mít     VB-S---1P-AA---
čas      čas     NNIS4-----A---- čas     čas     NNIS4-----A----
jen      jen     TT------------- jen     jen     TT-------------
večer    večer   Db------------- večer   večer   Db-------------
,        ,       Z:------------- ,       ,       Z:-------------
večer    večer   Db------------- večer   večer   Db-------------
půjdeme  jít     VB-P---1F-AA--- půjdeme jít     VB-P---1F-AA---
do       do      RR--2---------- do      do      RR--2----------
kina     kino    NNNS2-----A---- kina    kino    NNNS2-----A----
.        .       Z:------------- .       .       Z:-------------
</s>
<s id="4">
Tvoje    tvůj    PSHS1-S2------- Tvoje   tvůj    PSHS1-S2-------
kamarád  kamarád NNMS1-----A---- kamarád kamarád NNMS1-----A----
.        .       Z:------------- .       .       Z:-------------
</s>
</doc>
```

4 Acknowledgment

There would not be any corpus like this without the efforts of many students of Czech, and, indeed, without the devoted collectors and transcribers of texts. Collection and transcription of new data and metadata, including extensive search of metadata for the original texts, was performed and coordinated by Kateřina Lundáková with a significant help by Dagmar Toufarová.

We are grateful to everyone who helped to build the tools needed for tagging and lemmatization of Czech (Hajič (2001), Votrubec (2006)), to Michal Richter and Milan Straka for the spelling and grammar checker (Richter (2010), Richter et al. (2012)), and to Tomáš Jelínek for the tool assigning error lables (Jelínek et al. (2012)).

Many thanks are due to Pavel Procházka, whose generous help was crucial for the final technical stages, and Hana Skoumalová, who was the patient advisor for all technical problems.

The lion's share of merit gets Karel Šebesta as the instigator and leader of the whole enterprise. However, the result would not occur without the support and motivation of Barbora Štindlová, Svatava Škodová and Jirka Hana.

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 with the published CZESL-PLAIN corpus as a main result; new aquisitions, transcriptions, the provision of metadata and other work related to CZESL-SGT from *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. (2001). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Praha. 334 pp.

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.

Richter, M. (2010). *Pokročilý korektor češtiny [An Advanced Spell Checker of Czech]*. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague.

Richter, M., Straňák, P., & Rosen, A. (2012). Korektor – a system for contextual spell-checking and diacritics completion. In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.

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.

Votrubec, J. (2006). Morphological tagging based on averaged perceptron. In *WDS'06 Proceedings of Contributed Papers*, pages 191–195, Praha, Czechia. Matfyzpress, Charles University.

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.