The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-45510-5 27

CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered

Ondřej Bojar(X), Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Charles University in Prague, Prague, Czech Republic
{bojar,odusek,kocmi,libovicky,mnovak,popel, sudarikov,varis}@ufal.mff.cuni.cz

Abstract. We present a new release of the Czech-English parallel corpus CzEng. CzEng 1.6 consists of about 0.5 billion words ("gigaword") in each language. The corpus is equipped with automatic annotation at a deep syntactic level of representation and alternatively in Universal Dependencies. Additionally, we release the complete annotation pipeline as a virtual machine in the Docker virtualization toolkit.

Keywords: Parallel corpus · Automatic annotation · Machine translation

1 Introduction

We present the new release of our Czech-English parallel corpus with rich auto- matic annotation, CzEng 1.6. The version number is aligned with the year of the release, 2016.

CzEng 1.6 is the fifth release of the corpus and serves as a replacement for the previous version, CzEng 1.0 [1]. The parallel corpus CzEng was successfully used in multiple NLP experiments, most notably in the WMT shared translation tasks since 2010, see [2] through [3].[1] A pre-release of CzEng 1.6 has been already released and this year's WMT shared task is based on it.

CzEng releases are freely available for research and educational purposes and restricted versions of CzEng have been separately licensed for commercial use.

The main aim of the current release is to update and enlarge the collection of sources and to provide CzEng users with all tools needed to replicate the automatic annotation on other data.

2 CzEng 1.6 Data

CzEng 1.6 primarily uses the same data sources as the previous versions. Most of the sources grow in time and some can be better exploited. Table 1 summarizes the number of parallel sentences and surface (a-layer) and deep-syntactic (t-layer) nodes from each source for both languages. The a-layer nodes correspond to words

[1] http://www.statmt.org/wmt10 through http://www.statmt.org/wmt15.

Table 1. CzEng 1.6 data size.

Section	Sentence Pairs	Czech		English	
		a-layer	t-layer	a-layer	t-layer
Subtitles	39.44 M	286.70 M	211.49 M	325.20 M	208.78 M
EU legislation	10.18 M	296.19 M	219.79 M	324.11 M	200.47 M
Fiction	6.06 M	80.65 M	57.68 M	89.37 M	54.20 M
Parallel web pages	2.35 M	37.08 M	28.07 M	41.26 M	26.45 M
Technical	2.00 M	12.92 M	10.10 M	13.82 M	9.65 M
Medical	1.53 M	22.30 M	16.67 M	23.08 M	15.29 M
PDFs from web	0.64 M	9.64 M	7.51 M	10.32 M	6.61 M
News	0.26 M	5.65 M	4.20 M	6.22 M	3.93 M
Navajo	35.29 k	501.01 k	371.70 k	566.33 k	352.23 k
Tweets	0.52 k	9.55 k	6.97 k	10.19 k	6.78 k
Total	62.49 M	751.65 M	555.88 M	833.96 M	525.73 M

and punctuation symbols, with English sentences being by about 10 % longer due to articles and other auxiliaries.

CzEng 1.6 is shuffled at the level of sequences of not more than 15 consecutive sentences. The original texts cannot be reconstructed but some inter-sentential relations are retained for automatic processing of coreference (Sect. 3.3). Only sentences aligned 1-1 and passing the threshold of 0.3 of our filtering pipeline were included, leading to (indicated) gaps in the sequences. The filtering pipeline reduces the overall corpus size from 75 M sentence pairs to the 62 M reported in Table 1.

We prefer to de-duplicate each source at the level of documents, where avail- able, which necessarily leads to duplicated sentences. Comparing the overall size with the previous release (de-duplicated in the same manner), we see a substan- tial growth in size: 62 M vs. 15 M sentence pairs.

The largest portions of CzEng 1.6 come from movie subtitles, European leg- islation and fiction, as it was the case in the past. In this release, we also attempt to improve the coverage of the medical domain. In the following, we list changes since the last release for specific data sources:

European legislation was previously based on DGT-Acquis[2] in one of its pre- liminary versions, spanning the years 2004–2010 of the Official Journal of the European Union. Since there is no recent update of the corpus, we now use the search facility of EUR-Lex[3] to get access to more recent documents. The added benefit is that we can also obtain other documents in the col- lection, not only issues of the Official Journal. Particularly interesting are the Summaries of EU legislation,[4] which are written in less formal style and intended for general audience.

[2] https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis.
[3] http://eur-lex.europa.eu/.
[4] http://eur-lex.europa.eu/browse/summaries.html.

Movie subtitles are available in the OPUS corpus [4] and since the previous CzEng release, the collection has been significantly extended with Open- Subtitles.[5] Very recently, OPUS released yet another update[6] but it did not make it in time to be included in CzEng 1.6.

Subtitles of educational videos can be obtained from other sources and represent a rather different genre than movie subtitles. Not only are the topics slightly different (and mostly, there is one clear topic for each video), but the register is different: the sentences are longer and there are nearly no dialogues. CzEng 1.6 includes translated subtitles coming from Khan Academy[7] and TED talks.[8]

Medical domain is of special interest of several European research and innovation projects. We try to extend CzEng in this direction by specifically crawling some parallel health-related web sites using Bitextor [5] and also by re-crawling EMEA (European Medicines Agency)[9] corpus because its OPUS version[10] suffers from tokenization issues (e.g., decimal numbers split) and it is probably smaller than what can be currently obtained from the database.

3 Rich Annotation in CzEng 1.6

As in previous versions, CzEng is automatically annotated in the multi-purpose NLP processing framework Treex [6][11] based on the theory of Functional Generative Description [7]. The core of the platform is available on CPAN[12] and the various NLP models get downloaded automatically.

Treex integrates many processing tools including morphological taggers, lemmatizers, named entity recognizers, dependency and constituency parsers, co- reference resolvers, and dictionaries of various kinds. Many of these tools are well-known third-party solutions, such as McDonald's MST parser [8]; Treex wraps them into a unified shape. The heart of the integration is the Treex data format, where each of the processing modules (called "blocks" in Treex termi- nology) modifies the common data. Complete NLP applications such as a dia- logue system or a transfer-based MT system are implemented using sequences of processing blocks, called "scenarios".

Figure 1 illustrates the core annotation of CzEng. The left-hand trees rep- resent the Czech (top) and English (bottom) sentences at the surface-syntactic layer of representation (analytical, a-tree), and include morphological analysis (shown in teal). The dashed links between the trees show one of the automatic

[5] http://www.opensubtitles.org/.
[6] http://opus.lingfil.uu.se/OpenSubtitles2016.php.
[7] http://www.khanacademy.org/ and http://www.khanovaskola.cz/.
[8] http://www.ted.com/.
[9] http://www.ema.europa.eu/.
[10] http://opus.lingfil.uu.se/EMEA.php.
[11] http://ufal.mff.cuni.cz/treex, a web demo is available at http://lindat.mff.cuni.cz/ services/treex-web/run.
[12] https://metacpan.org/release/Treex-Core.

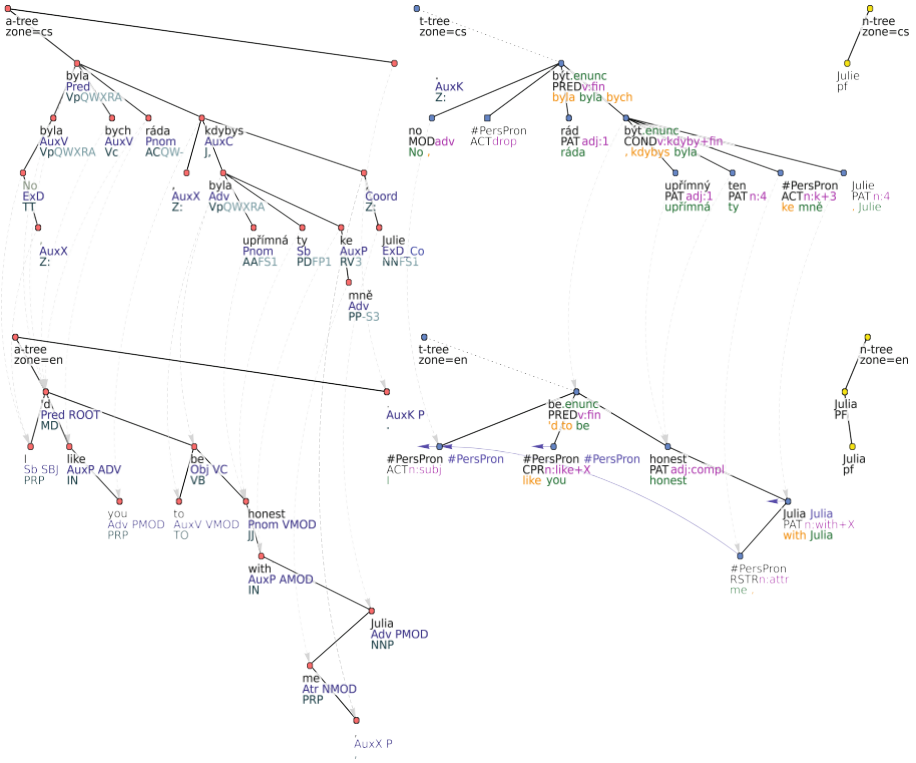

Fig. 1. One sentence pair from CzEng with the core parts of the automatic annotation.

word alignments provided in the data. The right-hand pair of trees are the deep-syntactic (tectogrammatical, t-tree) analyses and include again cross-language alignment. The blue arrows are co-reference links. The rightmost tree (n-tree) captures named entities in the sentence, as annotated by NameTag [9].

For the purposes of CzEng 1.6, we introduced several improvements into the pipeline, mostly on the deep-syntactic layer. They concern semantic role labeling (Sect. 3.1), word sense disambiguation (Sect. 3.2), and co-reference (Sect. 3.3).

3.1 Semantic Role Labels – Functors

The t-tree annotation involves assigning to each node its semantic role, *functor* [10], similar to PropBank [11] labels (shown in capitals in t-trees in Fig. 1). Func- tor assignment in CzEng 1.0 was handled by a logistic regression classifier [12] based on the LibLINEAR package [13]. Using Prague Dependency Treebank 2.5 [10] for Czech and Prague Czech-English Dependency Treebank (PCEDT) 2.0 [14] for English, we trained a new linear classifier using the VowpalWabbit library [15] and features based on syntactic and topological context. Automatic analy- sis with projected gold-standard labels is used for training to make the method

more robust. We achieve about 2 % accuracy gain in comparison to the previous method; classification accuracy on the evaluation sections of the treebanks used is 80.16 % and 78.12 % for Czech and English, respectively.

3.2 Verbal Word Sense Disambiguation

CzEng 1.6 includes word sense disambiguation in verbs [16], providing links to senses in valency lexicons for Czech and English. The setup used in parallel analysis exploits information from both languages (lemmas of aligned nodes) and the CzEngVallex bilingual valency lexicon [17] to gain better annotation accuracy.

3.3 Coreference Resolution

All the documents also contain annotation of coreference. In Czech, this is performed by the same Treex internal resolvers that were used in annotating CzEng 1.0. It covers coreference of pronouns and zeros. For English, coreference anno- tation has been provided mostly by BART 2.0 [18, 19]. BART is a modular end- to-end toolkit for coreference resolution. It covers almost all kinds of anaphoric expressions including nominal phrases. Only relative pronouns must be processed by the Treex internal resolver. To smooth down the processing pipeline, we set limits on BART's time and memory usage, which may cause that some doc- uments are excluded from coreference annotation. However, it happens only in around 1 % of CzEng documents. In addition, anaphoricity detection by the NADA tool [20] is applied to instances of the pronoun *it*. For an instance declared as non-anaphoric, a possible coreferential link assigned by BART is deleted.

Furthermore, coreferential expressions are exploited to improve word alignment. We use the approach presented in [21]. It is based on a supervised model trained on 1,000 sentences from PCEDT 2.0 [14] with manual annotation of word alignment for coreferential expressions. The only difference is that we do the analysis of PCEDT completely automatically in order to obtain the features distributed similarly to CzEng. Using this approach the alignment quality rises from 78 % to 85 % and from 71 % to 85 % for English and Czech coreferential expressions, respectively.

4 Analysis Dockered

While the whole Treex platform is in principle freely available and a lot of effort is invested in lowering the entry barrier, it is still not very easy to get the pipeline running, especially with all the processing blocks utilized in CzEng.

If a part of CzEng rich annotation is used as training data for an NLP task, CzEng users naturally want to apply such models to new sentences. Indeed, we have received several requests to analyze some data with the same pipeline as CzEng was annotated.

With the current release, we want to remove this unfortunate obstacle, and we release CzEng 1.6 with the complete monolingual analysis pipeline (for both Czech and English) wrapped as a Docker[13] container. Docker is a software bun- dle designed to provide a standardized environment for software development and execution through container virtualization. Docker's standardized contain- ers allow us to make Treex installation automatic, even without the knowledge of the user's physical machine environment. This should make it very easy to replicate the analysis on most operating systems.

An important added benefit is that the whole processing pipeline will be frozen in the version used for CzEng. This strong replicability is very hard to achieve even with current solid versioning, because some of the processing blocks depend on models that cannot be included in the repository for space or licensing reasons. Dockering CzEng analysis will thus help also ourselves.

We release two Treex-related Docker images: *ufal/treex*,[14] which creates a container with the latest release of Treex, and *ufal/czeng16*,[15] which contains
Treex frozen on the revision that was used to process CzEng 1.6. Both images are aimed at simplifying the process of Treex installation; the latter providing the means to easily reproduce the CzEng 1.6 monolingual analysis scenario.

The Dockerfile that is used to build the *ufal/czeng16* image simply specifies all the dependencies that have to be installed to run the Treex modules correctly, then it clones the Treex repository from GitHub[16] and configures the necessary environment variables. It also downloads and installs dependencies that are not available in the repository (mainly Morče tagger and NADA tool).

To run the analysis pipeline, you just need to pull the CzEng 1.6 Docker image from the Docker Hub repository and follow the instructions in the Readme file. The pipeline is able to process data as a filter (read standard input and write into standard output) as well as process multiple input files and save the results into all CzEng 1.6 export formats.

5 Availability

CzEng 1.6 is freely available for educational and research non-commercial uses and can be downloaded from the following website:

http://ufal.mff.cuni.cz/czeng

It is available in the following file formats; the first three are identical with the previous release, the last one is new.

Plain text format is very simple and has only four tab-separated columns on each line: sentence pair ID, filter score, Czech and English sentence.

[13] https://www.docker.com/.
[14] https://hub.docker.com/r/ufal/treex/.
[15] https://hub.docker.com/r/ufal/czeng16/.
[16] https://github.com/ufal/treex.

Treex XML format is the format used internally in the Treex NLP toolkit. It can be viewed and searched with the TrEd tool.[17]

Export format is a line-oriented format that contains most of the annotation. It uses one sentence per line, with tab-separated data columns (see the website and [1] for details).

CoNLL-U is a new text-based format with one token per line, introduced within the Universal Dependencies project[18] and further enriched with word alignment and multilingual support within the Udapi project.[19] The a-trees were automatically converted to the Universal Dependencies style, the t-trees are missing in CoNLL-U.

6 Conclusion

We introduced a new release of the Czech-English parallel corpus CzEng, ver- sion 1.6. We hope that the new release will follow the success and popularity of the previous version, CzEng 1.0.

CzEng 1.6 is enlarged, contains a slightly improved and extended linguis- tic annotation, and the whole annotation pipeline is now available for simple installation using the Docker tool. This makes it much easier to annotate other data than what is already provided in the corpus, which has been one of the major drawbacks of the previous CzEng release. We hope that by wrapping the analysis pipeline as a Docker container, we make an important step to making the annotation widely usable.

Acknowledgement. We would like to thank to Christian Buck for providing us with raw crawled PDFs. This project received funding from the European Union's Hori- zon 2020 research and innovation programme under grant agreement 645452 (QT21) and 644402 (HimL), and also from FP7-ICT-2013-10-610516 (QTLeap), GA15-10472S
(Manyla), GA16-05394S, GAUK 338915, and SVV 260 333. This work has been using language resources and tools developed and/or stored and/or distributed by the LIN- DAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. Bojar, O., Žabokrtský, Z., Dušek, O., Galušščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., Tamchyna, A.: The joy of parallelism with CzEng
   1.0. In: LREC, Istanbul, Turkey, pp. 3921–3928 (2012)
2. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Joint WMT and MetricsMATR, pp. 17–53 (2010)

[17] http://ufal.mff.cuni.cz/tred.
[18] http://universaldependencies.org/.
[19] http://udapi.github.io/.

3. Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., Turchi, M.: Findings of the 2015 workshop on statistical machine translation. In: WMT, Lisboa, Portugal, pp. 1–46 (2015)
4. Tiedemann, J.: News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In: RANLP, pp. 237–248 (2009)
5. Esplà-Gomis, M., Forcada, M.L.: Combining Content-based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites With Bitextor. Prague Bulletin of Mathematical Linguistics, vol. 93. Charles University (2010)
6. Popel, M., Žabokrtský, Z.: TectoMT: modular NLP framework. In: Loftsson, H.,
Rögnvaldsson, E., Helgadóttir, S. (eds.) IceTAL 2010. LNCS, vol. 6233, pp. 293–
304. Springer, Heidelberg (2010)
7. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Academia/Reidel Publishing Company, Prague (1986)
8. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: Non-projective dependency pars- ing using spanning tree algorithms. In: HLT/EMNLP, pp. 523–530 (2005)
9. Straková, J., Straka, M., Hajič, J.: Open-source tools for morphology, lemmatiza- tion, POS tagging and named entity recognition. In: Proceedings of ACL: System Demonstrations, Baltimore, Maryland, pp. 13–18. ACL (2014)
10. Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J.,
Žabokrtský, Z.: Prague dependency treebank 2.5 a revisited version of PDT 2.0. In: Coling, pp. 231–246 (2012)
11. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**, 71–106 (2005)
12. Mareek, D., Duek, O., Rosa, R.: Progress report on translation with deep genera- tion. Project FAUST deliverable D5.5 (2012)
13. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. JMLR **9**, 1871–1874 (2008)
14. Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J.,
Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z.: Announcing Prague Czech-English Dependency Treebank 2.0. In: LREC, Istanbul, Turkey (2012)
15. Langford, J., Li, L., Strehl, A.: Vowpal Wabbit online learning project. Technical report (2007)
16. Dušek, O., Fučíková, E., Hajič, J., Popel, M., Šindlerová, J., Urešová, Z.: Using
parallel texts and lexicons for verbal word sense disambiguation. In: Depling, Upp- sala, Sweden, pp. 82–90 (2015)
17. Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., Šindlerová, J.: Bilingual English-
Czech valency lexicon linked to a parallel corpus. In: LAW IX, Denver, Colorado, pp. 124–128 (2015)
18. Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A.: BART: a modular toolkit for coreference resolution. In: ACL-HLT, pp. 9–12 (2008)
19. Uryupina, O., Moschitti, A., Poesio, M.: BART goes multilingual: the

UniTN/Es- sex submission to the CoNLL-2012 shared task. In: EMNLP-CoNLL (2012)
20. Bergsma, S., Yarowsky, D.: NADA: a robust system for non-referential pronoun detection. In: Hendrickx, I., Lalitha Devi, S., Branco, A., Mitkov, R. (eds.) DAARC 2011. LNCS, vol. 7099, pp. 12–23. Springer, Heidelberg (2011)
21. Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., Mírovský, J.: Coreference in Prague Czech-English dependency treebank. In: LREC (2016)