The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 63–83

CzEng 0.9

Large Parallel Treebank with Rich Annotation

Ondřej Bojar, Zdeněk Žabokrtský

Abstract

We describe our ongoing efforts in collecting a Czech-English parallel corpus CzEng. The paper provides full details on the current version 0.9 and focuses on its new features: (1) data from new sources were added, most importantly a few hundred electronically available books, technical documentation and also some parallel web pages, (2) the full corpus has been automatically annotated up to the tectogrammatical layer (surface and deep syntactic analysis), (3) sentence segmentation has been refined, and (4) several heuristic filters to improve corpus quality were implemented. In total, we provide a sentence-aligned automatic parallel treebank of about 8.0 million sentences, 93 million English and 82 million Czech words. CzEng 0.9 is freely available for non-commercial research purposes.

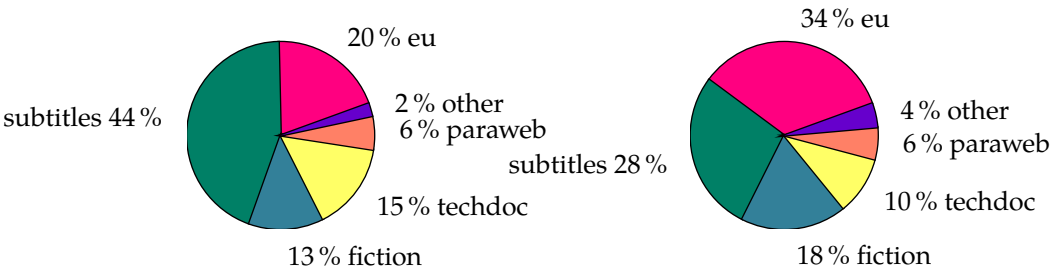
1. Introduction

Parallel corpora are essential for the training of (statistical) machine translation (MT) systems and used in other NLP tasks as well, e.g. language learning tools or terminology extraction. In the paper accompanying the previous release of CzEng (Bojar et al., 2008a), we confirmed that larger datasets usually improve the quality of MT, even if the additional data are out of the translated domain.

Some approaches to MT make use not only of large data but also of data (automatically) annotated: morphologically tagged and syntactically analyzed at a surface or a deep syntactic layer of linguistic description.

CzEng 0.9 is an extension of the previous release in both respects: we add data from several large sources like e-books and technical documentation and we use TectoMT (Žabokrtský et al., 2008) to augment the whole corpus with Czech and English automatic analyses at the morphological, analytical (surface syntactic, labelled "a-" in the sequel) and tectogrammatical (deep syntactic, labelled "t-") layers of description,

Please cite this article as: Ondřej Bojar, Zdeněk Žabokrtský, CzEng 0.9: Large Parallel Treebank with Rich Annotation. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 63–83.



(a) Sentences (b) Tokens

Figure 1. Types of parallel texts in CzEng 0.9. The depicted proportions are derived (a) from the number of included 1-1 sentence pairs, and (b) from the number of tokens (words and punctuation marks, summed for both languages).

following Functional Generative Description (Sgall, 1967; Sgall et al., 1986) and the Prague Dependency Treebank (Hajič et al., 2006).

Section 2 gives a detailed description of individual data sources included into CzEng 0.9. In Section 3, we briefly mention a general technique for fast semi-manual improvements when working with large data. The technique is then illustrated at several steps of corpus preparation, as described in Sections 4 (conversion to plain text), 5 (sentence segmentation and alignment) and 6 (automatic annotation up to the t-layer). Technical details such as sentence shuffling, the corpus structure, output file formats and corpus data size are given in Section 7 followed by the conclusion in Section 8.

2. Sources of Parallel Texts

This section gives an overview of all types of parallel text resources exploited in CzEng 0.9. The corpus is not claimed to be intentionally balanced in any sense—we simply collected as much material as possible. However, the set of covered topics is quite broad, with style ranging from formal language of laws and technical documents through prose fiction and journalistic language to colloquial language as often appearing in movies.

The proportions of the individual types of texts, which are included into CzEng 0.9, are roughly illustrated in Figure 1; detailed information is given later in Table 3. Note the difference in proportions calculated based on parallel sentences and based on words in one of the two languages.

For building CzEng 0.9, we used exclusively texts that were already publicly available in an electronic form, in most cases downloadable from the Internet. We did not do any book scanning or any other digitization activity.

2.1. Movie and Series Subtitles (`subtitles`)

Thanks to the community of movie fans, there is a huge amount of movie and series subtitles easily downloadable from several Internet subtitle archives.[1] More details about the cleaning of the data from this resource can be found in Beňa (2009), here we summarize the document alignment procedure and describe some newly implemented cleaning scripts.

As the movie/series subtitles stored in the two Internet archives were created by hundreds or thousands of contributors, one can hardly expect them to follow any strict naming conventions. First, we perform a filename normalization to represent only the following in the filename:

- the original movie/series name (from which determiners, prepositions, conjunctions and special characters were removed),
- the production year,
- the language of the subtitles (automatically detected from the file content),
- and also the series and episode numbers in the case of series.

Such normalization was reasonably reliable for de-duplicating and document-level alignment of movie subtitles, but it led to a large loss of the data in the case of series because there were too many irregularities in their original naming (or the information about episode/series number was completely missing). As mentioned in Bojar et al. (2009) an additional document-matching technique was used. Within each series, all beginning and end segments of all unpaired English subtitle files were compared with those of unpaired Czech files. The adequacy of such pairings was evaluated using a simple scoring function making use of a probabilistic translation dictionary. Then the pairs whose score was above an empirically found threshold were added into CzEng.

A number of filtering and cleaning scripts were implemented for the subtitle data, as their quality was very unstable: some authors systematically write "I'II" instead of "I'll", some others leave long passages untranslated, disregard punctuation, or disregard Czech diacritics, etc. Unless the errors were fixable with a very high reliability, we generally tend to throw out files with such a suspicious content.

Even if the subtitle data contains the highest amount of noise compared to the other sources of parallel texts, we still believe it is a valuable source because a lot of conversational phrases and colloquial language appear in them which would be difficult to find elsewhere. Moreover, the vocabulary distribution in the subtitle data probably better fit the real everyday language than e.g. European law does.

[1] CzEng 0.9 used `http://www.opensubtitles.com/` and http://www.titulky.com/.

2.2. Parallel Web Sites (paraweb)

Web sites with multilingual content can be an excellent source of parallel texts. For the most promising sites, it is worth implementing specialized crawlers and cleaners (and we do this for Project Syndicate, Section 2.6.1 and CzechNews, Section 2.6.3). However, we also wish to exploit the vast numbers of smaller sites.

Klempová et al. (2009) implement and evaluate a pipeline of tools that start with a few queries to search engines such as "lang:en česky" to obtain pages in English containing the Czech word for Czech from Google. Klempová et al. then crawl the whole web sites and use a combination of page structure and lexical content similarity to find parallel documents. In our current implementation, we apply a considerably simpler approach of aligning documents based on their URLs only.

Klempová et al. (2009) mention that it is surprisingly difficult to get large lists of candidate sites due to built-in limits on number of results available from search engines. We are grateful to Seznam[2], the largest Czech search engine, for an older version of all URLs of Czech Internet they index. We selected all domain names where the URLs contained a pattern indicating Czech or English language tag (e.g. "?lang=cs") and re-crawled the domains using our own crawler that specifically downloads only pages whose URL contains the language tag.

In addition to the selection of pages, we use the language tag also to find the document alignment. We have a short list of typical language tags for Czech and their variants for English, e.g. the above mentioned "?lang={cs,en}", implemented as regular expression substitution patterns. Given a website, we search the list of URLs of all documents of the website and apply the substitution. If the substitution can be applied and the resulting URL also exists, we promote the substitution pattern by a point. The highest scoring substitution pattern and the alignment of document it implies is then chosen for the given site.

Admittedly, we do not exploit the full potential of available parallel web pages: we require the pages to contain a language tag and the parallel version to differ only in the tag itself (and not e.g. the translation of the words in the URL). The advantage of aligning URLs only is the little computational cost and a relatively high accuracy.

2.3. Fiction (fiction)

2.3.1. E-books from Web (ebooks)

In the Internet, one can find a number of e-book archives such as Project Gutenberg[3] for English and Palmknihy[4] for Czech. We exploited such sources by downloading either e-book catalogs or directly the e-books files. Similarly to the case of subtitles

[2] http://www.seznam.cz/

[3] http://www.gutenberg.org/

[4] http://www.palmknihy.cz/

(Section 2.1), different e-book resources provided us with different metadata, so some metainformation normalization was necessary. We converted the information about the roughly 38,000 available e-books into a uniformly formatted catalog, whose entries contained

- normalized name of the book author: lowercased surname and the first letter from the first name; special rules for unifying transliteration variants (Tolstoj/Tolstoy) were applied,
- normalized book title,
- language (Czech or English),
- list of sources the book is available from.

Then the document-level alignment phase came. For each author, for whom the catalog contained at least one Czech book and at least one English book, all possible Czech-English book pairs were automatically scored. The heuristic scoring function took Czech and English titles as its input and produced a real number (weighted sum of several features) as the output. The features were based on the length similarity of the title strings, string similarity of the individual word pairs, translation probability of the individual word pairs, prefix similarity of the individual word pairs, etc. For each author, a list of Czech-English book pairs (whose score was above a certain threshold) as well as lists of remaining unpaired Czech and unpaired English books were generated. The weight and threshold values were optimized semi-automatically in several iterations, using a sample of roughly 20 authors with known book pairing.

The alignment algorithm identifies around 449 possible book pairs for 271 authors. This list was checked manually. Wrong pairs, duplicated pairs, and pairs containing poetry or dramas were excluded. 157 book pairs were confirmed as correct, and additional 102 new pairs were manually found among the unpaired books. No surprise that the simple title alignment approach did not reveal many book pairs such as in the case of Jules Verne's "Michel Strogoff" whose Czech title is "Carův kurýr" (Tsar's messenger).[5]

The e-book data, as acquired from the various archives, were stored in a highly diverse set of file formats. The need for format conversion leads to another data loss, as discussed in Section 4.

2.3.2. Kačenka corpus (`kacenka`)

Kačenka (Rambousek et al., 1997) is a Czech-English parallel corpus created by the Department of English, Faculty of Arts, Masaryk University in Brno in 1997.[6] It con-

[5] Of course, even such books could have been automatically paired supposing we already had their full texts in hand, but that was not always the case, as from some web archives it is not possible to download all books at once. That is why we performed the title-based alignment first and only then selectively downloaded the paired books.

[6] `http://www.phil.muni.cz/angl/kacenka/kachna.html`

tains texts of 12 English books and their Czech translations. The texts were manually aligned at the sentence level; this alignment has been preserved in CzEng 0.9.

All books contained in Rambousek et al. (1997) have been used when compiling CzEng 0.9. If a book pair appeared both in Rambousek et al. (1997) and in other e-book resources (Section 2.3.1), only the Rambousek et al. (1997) version was used.

2.3.3. Reader's Digest (`rd`)

Prague Czech-English Dependency Treebank (Cuřín et al., 2004) contains a parallel corpus composed of raw texts of 450 articles from the Reader's Digest, years 1993-1996, and their Czech translations.

2.4. European Union Law (`eu`)

2.4.1. JRC-Acquis (`celex`)

JRC-Acquis is a freely available parallel corpus containing European Union documents mostly of legal nature (Ralf et al., 2006).[7] It is available in 20 official EU languages. The corpus is encoded in XML, and contains roughly 8,000 millions documents per language.

We included into CzEng 0.9 all Czech-English documents pairs available in JRC-Acquis v.3.0 whose length ratio measured in characters was not too far from 1—within the interval $[1.4^{-1}; 1.4]$. If their length ratio was outside the interval, an attempt at extracting at least some parts of the documents was made: both documents were decomposed into head, body, signature and annex parts, and at least some corresponding parts were extracted if their length was inside the given interval. The motivation for this step was the following: in some cases the Czech version of the documents does only refer to the annex of the English version instead of containing the proper translation of the annex. If such document pairs are automatically sentence aligned, they might be rejected by the aligner (see Section 5.2) as a whole as they seem to be too much different, while if only their reasonably similar subparts are extracted, the chance for a successful sentence alignment grows.

2.4.2. The European Constitution proposal (`euconst`)

The European Constitution proposal from the OPUS corpus (Tiedemann and Nygaard, 2004).

[7] `http://wt.jrc.it/lt/acquis/`

2.4.3. Samples from the Official Journal of the European Union (`eujournal`)

Samples from the Official Journal of the European Union, which is a tiny collection of some rather randomly chosen issues of the the Official Journal of the European Union.

2.5. Technical documentation (`techdoc`)

2.5.1. KDE and GNOME documentation (`kde`, `gnome`)

KDE and GNOME are two most popular graphical user interface for running Linux. Both of them are open-source software projects and for both of them their Czech localizations (product translations) are available on the Internet.[8,9]

2.5.2. Microsoft glossaries (`microsoft`)

Microsoft glossaries are lists of technical terms and longer expressions and messages used e.g. in Microsoft software products. The glossaries are available for a number of languages. They are intensively used by technical translators as they constitute a rich resource of technical vocabulary. The glossaries are publicly available from the Microsoft Corporation FTP Server and its mirrors.

2.6. News texts (`news`)

2.6.1. Project Syndicate (`syndicate`)

Project Syndicate is a not-for-profit institution which currently consists of 432 newspapers in 150 countries.[10] There is a large number of newspaper articles available on its web pages, many of them existing in more language versions. Those articles that were available in English and Czech in August 2009 were used for the creation of CzEng 0.9.

2.6.2. Wall Street Journal (`wsj`)

Prague Czech-English Dependency Treebank (PCEDT, Cuřín et al. (2004)) contains English texts of Wall Street Journal articles adopted from the Penn Treebank (Marcus et al., 1993), and their Czech translations created (by human translators) specifically for the PCEDT needs.

[8] `http://www.gnome.org/projects/`

[9] `http://l10n.kde.org/`

[10] `http://www.project-syndicate.org/`

2.6.3. Czech News (`czechnews`)

The Czech news portal Aktualne.cz provides a limited selection of the news in English[11]. We implemented a custom crawler and we align the documents on the basis of links back to the Czech version available in the translated page.

2.7. User-Contributed Translations from Navajo (`navajo`)

Navajo[12] is a machine-translated Czech version of (the English content of) Wikipedia, which is a highly popular, multilingual, web-based, free-content encyclopedia. Similarly to Wikipedia, which is written and improved collaboratively by volunteers, also the content of Navajo is gradually improved by a community of volunteers who submit human-corrected translations of the individual entries. Such user-contributed Czech translations paired with their original English counterparts can be treated as a relatively reliable source of parallel texts, whose main advantage is a wide range of topics.[13] Therefore we include them into CzEng 0.9 too.

3. General Approach to Fixing Errors

Throughout the processing pipeline, we feel that the most successful correction steps are implemented using the following generic approach:[14]

1. We extend the tool in question or one of the subsequent tools to include a simple detector of suspicious positions in the corpus. We also try to automatically propose one or more possible corrections or solutions of the assumed problem.
2. We manually scan and quickly confirm or deny individual proposed solutions, e.g. by adding a prefix to each line in a text file. We carefully preserve old annotations to avoid duplicating manual effort.
3. The tool in question is extended to use the file of confirmed annotations and apply the corrections. For input with no confirmed or denied annotations, suspicious occurrences are still collected.

The main advantage of the setup is the excellent trade-off between manual labor and overall output quality. If new data are added, we can quickly add decisions for new suspicious cases. When rebuilding the whole corpus, old decisions are simply reused.

Another great advantage is the possibility to sort automatic suggestions by various criteria, such as the expected reliability (and thus little effort needed to confirm or deny a rule) or overall frequency. With time constraints on manual annotation, we can thus focus on some most important subset of the errors and leave others unsolved.

[11] `http://aktualne.centrum.cz/czechnews/`

[12] `http://www.navajo.cz/`

[13] As evaluated in Bojar et al. (2008b), about 70% of segment pairs are of reasonable quality.

[14] A similar approach proved fruitful in the pre-release corrections of PDT 2.0 (Štěpánek, 2006).

In our complex pipeline, we often take advantage of more elaborate information available in subsequent processing steps. One of the best examples is automatic suggestion of missed and superfluous sentence boundaries based on sentence alignment between Czech and English.

We used the approach in the following tasks:

- language guess based on book title, confirmed later, after the book is converted to plain text using the vocabulary of the book
- book alignment based on book titles, confirmed later by the quality of sentence-level alignment
- automatic detection and removal (upon confirmation) of page breaks and page numbers
- sentence segmentation, corrected later by sentence-level alignment

4. Handling Various Input Formats

4.1. Format Convertors

We implemented a generic wrapper of several tools to convert many file formats (pdf, doc, rtf, pdb, html, txt and also archive-like formats lit, zip and rar) to plain text encoded in UTF-8 and attempting to identify documents with malformed encoding.

Our handling of archive-like formats is rather simplistic at the moment. We make use of the archive only if the largest file in the archive clearly dominates other files and can be converted to plain text. We don't attempt to e.g. concatenate separate chapter files.

The most problematic file format in our experience is PDF. In PDFs, the content can be internally stored in various ways (including bitmap images of book pages) and e.g. Czech accented letters are prone to lose the accent or get mis-encoded. Different implementations of PDF-to-text conversions including Acrobat Reader can run in different problems on a given file. Moreover, hyphenated words and page headers are frequent and we have also found obscure cases of HTML print-outs in PDF where the printed header changed throughout the document as the timestamp in the header was changing. We attempted to solve most of the issues manually (by converting individual PDF files to txt prior to our generic convertor) but not everything has been handled due to time constraints.

4.2. Removing Page Breaks

Some of the texts include page numbers and other repetitive sequences such as page headers throughout the document. In worst cases, such header or footer appears even in the middle of a sentence. While the magnitude of the problem is not too severe (a book has a few hundred pages, so only a few hundred sentences per book can be malformed), we attempt to fix many of the cases.

We implemented a simple heuristic to identify candidates of page breaks and manually confirm them. A page break candidate, once constructed, is essentially a simple regular expression describing the prefix, the page number placeholder and the suffix that should be removed from anywhere in the document.

Our heuristic searches for all numbers in the document. Each occurrence of a number contributes to one or more candidates depending on the actual number observed and a very short character context of the number. In essence, we require the number to be not far away from the number of the last observed occurrence attributed to the candidate. For each candidate, we store all the numbers attributed to it, the prefix and suffix seen in the first occurrence and the length of a subsequence of the prefix and suffix seen in most other occurrences.

After the whole text has been processed, we sort the candidates based on the span of numbers covered by the candidate decreased by the number of gaps in the sequence and the number of duplicated entries. The most promising candidates represent the longest sequences of numbered items with the fewest errors in numbering. In most cases, these are indeed page numbers but sometimes we find footnotes or the table of contents instead. Due to the variance in book styles, we cannot assume some average number of pages so we prefer manual inspection of the list of candidates. This also allows to make sure that the suggested prefix and suffix are correct. After the manual confirmation, all occurrences of the confirmed candidate are removed from the document.

4.3. Unwrapping

Depending on the original file format and individual typesetting rules, some of the documents are hard wrapped, some indicate paragraphs by a blank line and some indicate them by indentation.

For the purposes of sentence segmentation (Section 5.1), we need somewhat normalized format to match the training data of our segmenter.

If there are more than 30% of lines longer than 90 characters, we assume the document is not hard-wrapped. For hard-wrapped documents, we check the number of blank lines in the document, and if there are more than e.g. 500, we assume they represent the paragraph boundary. With not enough blank lines, we assume the paragraphs are indicated by indentation and we insert a blank line before every indented line. Some documents do not even use indentation, so we additionally assume there is paragraph break whenever the line is shorter than 65 characters. When unwrapping individual paragraphs, we also join hyphenated words.

Some HTML documents we got are hard-wrapped using ⟨BR⟩ tags and we generally treat the ⟨BR⟩ tag as a paragraph boundary in our simple HTML stripper, so a specific rule for this case was needed.

	1-1	2-1	1-2	1-0	0-1	3-1	Others
Overall	9,860,595	688,946	495,372	331,576	316,282	87,691	167,801
subtitles	3,721,423	189,985	145,787	158,592	76,410	14,014	27,401
eu	2,382,721	312,656	155,694	64,147	99,901	55,541	90,403
techdoc	1,350,803	21,713	18,003	18,628	3,856	1,883	2,868
paraweb	1,146,999	104,264	51,046	52,441	95,343	11,434	23,657
fiction	1,070,639	55,218	119,206	34,804	37,293	4,585	22,336
news	145,763	3,733	3,778	1,891	2,902	165	831
navajo	42,247	1,377	1,858	1,073	577	69	305

Table 1. Types of aligned text segments as detected by Hunalign in the individual sources. X-Y stands for segment pairs containing X sentences in the English segment and Y sentences the Czech segment.

5. Sentence Segmentation, Alignment and De-Duplication

5.1. Sentence Segmentation

We use the trainable tokenizer introduced in Klyueva and Bojar (2008) with a few new extensions to perform sentence segmentation. The tokenizer internally performs a deterministic "rough" tokenization and deterministically inserts markers of positions where a sentence break or token join (e.g. space-delimited thousands) may happen. A maximum entropy classifier then decides where breaks or joins indeed happen based on features of surrounding tokens. We use only the sentence break information (and occasional token joins) but use the original non-tokenized format otherwise. The reason is that we wish to use TectoMT internal tokenization (Section 6 below) which should be compatible with the whole processing pipeline.

The training set for the maximum entropy classifier was further extended to contain more examples of the document types we deal with, e.g. book texts with lots of direct speech. Usually, the training set is created manually by complementing a sample plain text with the intended tokenization and segmentation. The trainable tokenizer creates training instances for the classifier by comparing the original and tokenized text. In our case, we were able to extend the training set of texts for both English and Czech semi-automatically by finding segments aligned 1-to-2 and containing a full stop somewhere around the middle of the single segment. Most of these cases were indeed errors where either the single segment should have been split, or the two corresponding segments in the other language should have been joined (e.g. at an unrecognized abbreviation). Simply adding these sentences with the correct segmentation projected from the other language improved the accuracy on our dataset.

5.2. Sentence Alignment

We use Hunalign (Varga et al., 2005) to automatically align sentences. To reduce data sparseness, we perform a rough tokenization (at this stage, the texts are only segmented and preserve original tokenization) and lowercase and restrict each token to at most first four letters. Additionally, we use a probabilistic dictionary based on GIZA++ word alignment of the previous version of CzEng, with the identical reduction of word types.

Table 1 lists alignment types seen in various source data. On average, about 82% of segments are aligned one-to-one but e.g. for the European Law texts, the percentage falls to 75%.

5.3. De-duplication

It is a common practice in corpus preparation to remove duplicated portions of data. For some types of texts, the common simple "sort | uniq" de-duplication procedure may skew the distribution of phrases unnecessarily, making e. g. a very common phrase "Yes. = Ano." occur only once in the whole corpus.

For most sections of our corpus, we completely avoid de-duplication at the level of segments and prefer de-duplication at the level of documents (e.g. e-books). For some sources, e.g. the web collection, de-duplication is inevitable because web pages from a single site usually contain large amounts of repetitive text (that is actually seldom read by humans, unlike repetitive phrases in books).

To avoid the above-mentioned distortion, we remove duplicated aligned segments of web pages using a more sensitive context-based technique: we use a sliding window of 3 consecutive lines and print the lines in the window if no such window was printed before. For instance, for the lines "a b c a b c b d b" we get "a b c b d b". The second occurrence of "a b c" got removed but the overall distribution of "b" is influenced less.

5.4. Plaintext Checks

Sentence-aligned plaintext format is suitable for performing many simple checks to filter out either mis-aligned or simply bad segments. At this stage of corpus collection, we search and remove all suspicious sentence pairs, i.a.:

- the Czech and English sentences are identical strings (usually untranslated text from a website),
- the lengths of the sentences are too different (usually due to a wrong alignment or a wrong sentence segmentation),
- there is no Czech word on the Czech side or English word on the English side[15],

[15] We use the word lists from the British National Corpus the Czech National Corpus disregarding letter case. We prefer longer words for the test: if there are some words longer than three letters, at least one of

Bad 1-1 Segments [%]		Most Frequent Errors
subtitles	4.6	Mismatching lengths (42.0%), Identical (27.3%), No English word (10.9%),
eu	33.3	Identical (39.9%), No English word (19.2%), Not enough letters (17.2%),
techdoc	10.2	Identical (37.9%), No English word (28.4%), Not enough letters (10.0%),
paraweb	59.5	Identical (61.7%), No English word (25.1%), Mismatching lengths (3.3%),
fiction	3.1	Mismatching lengths (54.9%), Suspicious char. (14.6%), Repeated character (6.1%),
news	3.8	Identical (54.1%), Suspicious char. (17.7%), No English word (9.3%),
navajo	11.9	Identical (40.9%), No English word (19.0%), Not enough letters (11.7%),

Table 2. Percentage of 1-1 sentence pairs rejected by various error-detection filters.

- there is a suspicious character (either non-printable one or an unlikely symbol) or a repeating sequence of a character.

Table 2 displays the percentage of 1-1 aligned sentences with one or more errors. The second column in the table lists the most frequent error in each of the sections.

Many of the errors can be corrected in earlier stages of corpus cleaning and we will continue to refine the cleaning process but for the time being, we prefer to remove all suspicious segments.

The overall most frequent error is "Identical", and we see that e.g. more than 36% of web data (61.7% out of 59.5% of erroneous segments) are removed due to this error. Unfortunately, many of the seemingly parallel web pages contain non-translated sections. The cleanest source is probably the ebooks section with some errors in segmentation or alignment (Mismatching lengths).

6. Sentence Annotations

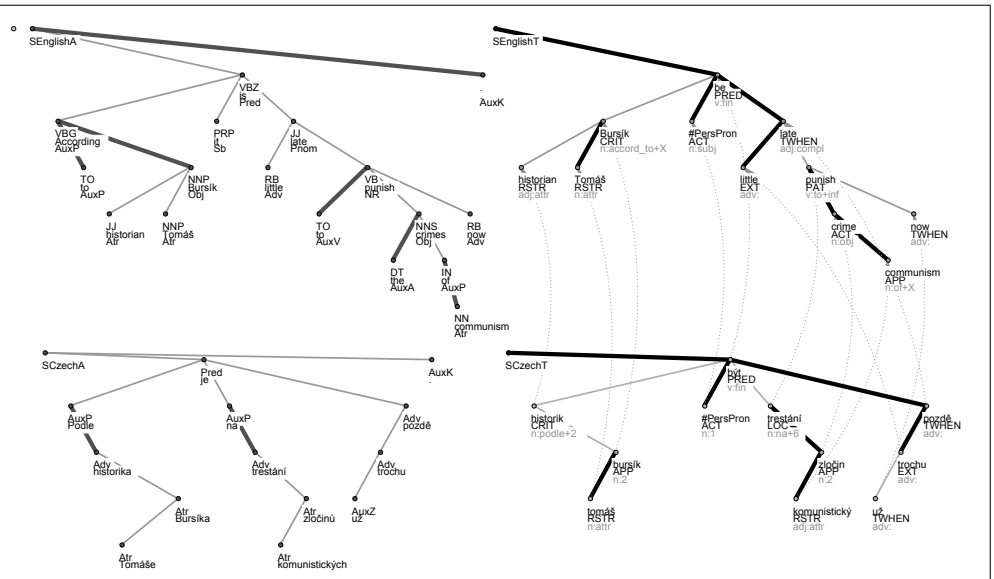
The pairs of Czech and English 1-1 aligned sentences are enriched with rich morphological and syntactic annotations. The annotation scheme is adopted (with certain modifications) from the Prague Dependency Treebank 2.0 (Hajič et al., 2006). Each sentences is provided with three layers of annotation:

- morphological layer: each token (word or punctuation mark) is labeled with its lemma and morphological tag,
- analytical layer: each sentence is represented as a surface-syntax dependency tree called analytical tree (a-tree), with nodes corresponding to tokens and edges corresponding to surface-syntax dependency relations,
- tectogrammatical layer: each sentence is represented as a deep-syntactic dependency tree called tectogrammatical tree (t-tree), in which nodes have complex structure and correspond only to autosemantic words.

In addition to the PDT 2.0 scheme, a new layer containing annotation of named entities is added.

---

them has to be confirmed in the word list. If all words contain at most three letters, we accept also shorter words for the word list check.



English: *According to historian Tomáš Bursík it is little late to punish the crimes of communism now.*
Czech: *Podle historika Tomáše Bursíka je na trestání komunistických zločinů už trochu pozdě.*

Figure 2. Simplified visualization of parallel analytical and tectogrammatical tree representations of a sample CzEng 0.9 sentence pair.

The fully automatic annotation procedure is implemented within the TectoMT framework (Žabokrtský et al., 2008). The procedure is highly similar for both languages:[16]

1. each sentence is tokenized using a simple regular expression pattern,
2. the sentence is tagged by the Morce tagger (Spoustová et al., 2007),
3. the tokens are lemmatized; this is done already in the tagging step in the case of Czech sentences, while for English a new lemmatizer was implemented in TectoMT (Popel, 2009),
4. named entities are recognized and classified; a recognizer based on Support Vector Machines described in Kravalová and Žabokrtský (2009) is used for Czech sentences, while for English sentences we use Stanford Named Entity Recognizer introduced in Finkel et al. (2005),

[16]The procedure description is highly simplified here, in fact the procedure composes of roughly sixty subsequent blocks (basic processing units in TectoMT).

5. analytical dependency tree is created by the maximum spanning-tree parser introduced in McDonald et al. (2005) (using feature pruning described in Novák and Žabokrtský, 2007),
6. a-tree nodes are labeled with analytical functions; the values are provided already by the parser on the Czech side, while on the English side the values have to be assigned subsequently (a rule-based analytical function assigner developed in Popel, 2009 is used),
7. a t-tree is created from the a-tree by merging autosemantic a-nodes with their associated auxiliary a-nodes (e.g. a noun with a preposition and a determiner node),
8. the t-tree is labeled with grammatemes,
9. grammatical coreference links are identified in the t-tree,
10. the t-tree nodes are labeled with functors by a tool developed in Klimeš (2006),
11. finally, the resulting t-trees are aligned using the tectogrammatical aligner developed in Mareček (2008).

A sample pair of resulting sentence representations (and their alignment) is shown in Figure 2.

6.1. Line-Oriented Operations

TectoMT uses a complex XML-based file format, an instance of Prague Markup Language (Pajas and Štěpánek, 2006). While the format is excellent for the rich annotation and the interoperation of TectoMT processing blocks, it brings an additional overhead for tasks performed on large sets of sentences. Quick and simple selection of sentences matching regular expressions, counting sentences or line shuffling cannot be performed with standard utilities like grep, wc or shuf, because sentences represented in XML span over multiple lines.

To facilitate the use of line-based tools on TectoMT data, we introduce a simple modification to the file format. The new file format is called "lot" (line-oriented-tmt) and stores each sentence using XML on a single line. In other words, line breaks and indentation whitespace within the XML representation of sentences are removed. To match the line-oriented approach even closer, we omit any XML header and footer sections in "lot", so every line of a "lot" file holds a sentence. Fortunately, it is not common to store any valuable information in a header section once the text has been segmented.

Both conversion to and back from "lot" is fast and can operate on an infinite stream of sentences. When converting to "lot", we use a SAX parser to read sentence after sentence, strip any line breaks and emit the sentence. To convert back from "lot", one has to simply add a proper XML header and footer and optionally reindent the file, e.g. using "xmllint -format".

7. Corpus Structure and Size

This section provides technical details on the final shape of CzEng 0.9 data.

7.1. Dividing Data into Files, Shuffling

The Czech author law[17] permits to use short citations of published works for non-commercial educational or research purposes. To avoid the possibility of reconstructing the original texts included in CzEng, we break all documents into short fragments, shuffle them and discard any explicit information that would allow to reconstruct the original ordering of the fragments.

Let us recall that CzEng contains only sentences automatically aligned 1-1. In reality, most documents are not translated sentence by sentence, and even if this were the case, the exact sentence alignment is seldom found by the automatic procedure. So the original documents are unreconstructable from what is contained in CzEng 0.9 not only because of fragmentization and shuffling, but also because of the data losses imposed by the 1-1 requirement (and also because of other losses during pair filtering).

In order to preserve the utility of CzEng for advanced NLP techniques that extend beyond sentence boundary, such as anaphora resolution, we preserve at least short sequences of sentences, if possible. Given our processing pipeline, some breaks of the continuous flow of sentences naturally happen at sentences not aligned 1-1 or filtered by one of our plaintext checks in either of the languages. We use all these breaks and add further breaks after at most 13 consecutive sentence pairs. Due to the natural breaks, there are only 4.0 sentences per block on average. We shuffle the obtained set of blocks and assign a unique identifier to each of the blocks. Finally, the blocks are concatenated to files of about 50 to 60 sentence pairs depending on the exact sizes of the blocks in the file. We use the above mentioned line-oriented approach for these operations.

For domain-specific training or domain adaptation, the block identifiers preserve the coarse data source type (subtitles, eu, paraweb, techdoc, fiction, news, navajo) but no other meta-information is available.

7.2. Dividing Data into Sections

In order to reduce the load on the file system and to simplify selection of smaller random samples of the data for e.g. debugging, we organize the final TMT files into 100 subdirectories, each containing approximately 1500 files.

We expect many researchers to use the full size of CzEng for training their systems but some may wish to reserve a portion of the data for evaluation purposes. In order to synchronize the selection of the test set, we label about 10% of the data `dtest`

[17] The law 121/2000 Sb. including amendments up to 168/2008 Sb., see §31.

(development test set) and another 10% of the data `etest` (evaluation test set). The development test set should be used for tuning of parameters and the evaluation test should be used for final evaluation only.

The directories are thus called `train00`, …, `train79`, `dtest80`, …, `dtest89`, and `etest90`, …, `etest99`.

In any case, researchers should clearly indicate which sections they used for the training and for the evaluation.

7.3. File Formats

7.3.1. CzEng in TMT Format

The main file format of CzEng 0.9 is the TectoMT file format called TMT, an instance of Prague Markup Language (Pajas and Štěpánek, 2006) based on XML. Unlike the PDT 2.0 file format, TMT allows to keep all layers of language representation in a single file. In CzEng 0.9, each TMT file is a sequence of around 50 bundles, each of them comprising morphological, analytical and tectogrammatical representations of an English sentence and of its Czech counterpart sentence, as well as their original surface string forms and their tectogrammatical alignment.

7.3.2. CzEng in Plain Text

For some applications, the rich annotation stored in TMT files is not needed or causes an unwanted bias due to our tokenization rules. Therefore, we also provide CzEng in plain text format, one sentence pair per line. The English and Czech versions of the sentence are delimited by a single tab.

We preserve the same corpus division into training and test sections. Instead of a directory `train··`, the section is stored in a single file `train··.gz`.

7.3.3. CzEng Export Format

The TMT format described above, is the only authoritative format of CzEng 0.9 rich annotation. However, to allow access to the rich annotation for researchers who do not wish to use the TectoMT framework with its API for TMT files, we provide CzEng 0.9 in a simple export format as well. Note that not all information from the original TMT files is preserved[18].

The export format represents each sentence pair on a single line consisting of the following tab-delimited columns: Sentence ID (including coarse CzEng source type), English a-layer, English t-layer, English lex.rf (i.e. links from English t-nodes to the

[18]We do not export all attributes of the nodes. We also remove any spaces delimiting thousands in numbers whereas the original TectoMT annotation pipeline represents space-delimited numbers in a single node with spaces in the attribute form and the a- and t-lemmas.

corresponding a-node bearing the lexical value), English auf.rf (i.e. links from English t-nodes to their auxiliary a-nodes), Czech a-layer, Czech t-layer, Czech lex.rf, Czech aux.rf, English-Czech t-layer alignment.

All the columns representing the dependency tree at a layer use so-called "factored" notation: each space-delimited word on the line represents one node of the tree. Individual attribute values of the node are delimiter by vertical bar "|". The order of the attributes is fixed for a given language and layer and usually can be guessed from attribute values.

The dependency structure of the tree is represented using two attributes: the "ord" stores the global linear order of the node in the tree starting from 1 and the "gov" contains the ord value of the governor of the node. The root of the tree has the gov value set to zero. The nodes of the tree are always listed in ascending order of ord and there are no gaps in the numbering.

In CzEng 0.9, we export these attributes:

- Czech and English a-layers: word form, lemma, morphological tag, ord, gov, analytical function.
- English t-layer: t-lemma, functor, ord (deepord), gov, nodetype, formeme, the grammatemes: sempos, number, negation, tense, verbmod, deontmod, indeftype, aspect, numertype, degcmp, dispmod, gender, iterativeness, person, politeness, resultative, and the attributes: is_passive, is_member, is_clause_head, is_relclause_head, val_frame.rf.
- Czech t-layer: t-lemma, functor, ord (deepord), gov, nodetype, formeme, the grammatemes: sempos, number, negation, tense, verbmod, deontmod, indeftype, aspect, numertype, degcmp, dispmod, gender, iterativeness, person, politeness, resultative, and the attributes: is_passive, is_member, is_clause_head, is_relclause_head, val_frame.rf.

All the columns representing some kind of links between two layers or languages are simple space-delimited pairs of indices. Unlike the ord and gov attributes, here the nodes are indexed starting from zero. In other words, e.g. the pair "0-1" of the Czech lex.rf indicates that the first Czech t-node (index 0) obtained its lexical value from the second (index 1) a-node, a typical situation of a noun with a preposition at the beginning of the sentence.

Some types of alignment allow 1-to-many links or possibly even many-to-many links. In these cases, some nodes are simply mentioned in the listing more than once.

Again, the same corpus division into training and test sections is preserved. Instead of a directory `train··`, the section is stored in a single file `train··.gz`.

7.4. CzEng 0.9 Size

Table 3 lists total number of sentences and Czech and English nodes at both layers of the annotation per section. The number of a-nodes can be interpreted as the number of "words" including punctuation.

		English		Czech	
Source	Sentences	a-nodes	t-nodes	a-nodes	t-nodes
eu	1,589,036	31,725,089	19,458,544	28,484,512	19,310,396
subtitles	3,549,367	26,550,305	16,615,991	22,175,284	16,675,187
fiction	1,036,952	17,045,233	10,861,341	15,031,926	11,102,760
techdoc	1,212,494	9,099,748	6,339,129	8,460,491	6,512,247
paraweb	464,522	4,946,552	3,666,149	4,750,757	3,667,297
news	140,191	3,196,303	2,019,758	2,945,777	2,220,789
navajo	37,239	612,826	385,292	539,659	405,484
Total	8,029,801	93,176,056	59,346,204	82,388,406	59,894,160

Table 3. Number of sentence pairs in CzEng 0.9 and number of nodes in their analytical and tectogrammatical tree representations. Artificial tree roots are not counted here, therefore the numbers of a-nodes given in the third and fifth column are equal to the number of tokens (words and punctuation marks) contained in the corpus.

7.5. Obtaining CzEng 0.9

CzEng 0.9 is available for non-commercial research purposes at:
`http://ufal.mff.cuni.cz/czeng/`

8. Conclusion

We have presented CzEng 0.9, a new release of our Czech-English parallel corpus, extended both in the data size and the depth of automatic annotation. Compared to previous versions, the corpus should be cleaner thanks to several automatic error detection techniques we implemented. Inevitably, many errors remain in the released corpus and we plan to further refine our filtering techniques and base them on the deep syntactic analyses and their alignment as well in future versions.

We believe that CzEng 0.9 is a unique resource for MT developers (definitely for the given pair of languages), and hope that that its availability will further boost the research in the field.

9. Acknowledgement

The work on this project was supported by the grants MSM0021620838, MŠMT ČR LC536, 1ET101120503 and FP7-ICT-2007-3-231720 (EuroMatrix Plus).

Bibliography

Beňa, Peter. Filmové titulky jako zdroj paralelních textů (movie subtitles as a source of parallel texts). Bachelor's Thesis, Faculty of Mathematics and Physics, Charles University in Prague, 2009.

Bojar, Ondřej, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech, 2008a.

Bojar, Ondřej, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech, 2008b.

Bojar, Ondřej, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25, 2004.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics ACL 2005*, pages 363–370, 2005.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. Prague Dependency Treebank 2.0. LDC, Philadelphia, 2006.

Klempová, Hana, Michal Novák, Peter Fabian, Jan Ehrenberger, and Ondřej Bojar. Získávání paralelních textů z webu. In *ITAT 2009 Information Technologies – Applications and Theory*, Sept. 2009.

Klimeš, Václav. *Analytical and Tectogrammatical Analysis of a Natural Language*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, 2006.

Klyueva, Natalia and Ondřej Bojar. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proc. of International Conference Corpus Linguistics*, pages 188–195, Oct. 2008.

Kravalová, Jana and Zdeněk Žabokrtský. Czech named entity corpus and SVM-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 194–201, Suntec, Singapore, 2009. Association for Computational Linguistics. ISBN 978-1-932432-57-2.

Marcus, M. P., B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mareček, David. Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus. Master's thesis, Charles University, MFF UK, 2008.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HTL/EMNLP*, pages 523–530, Vancouver, BC, Canada, 2005.

Novák, Václav and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. In Matoušek, Václav and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, number XVII in Lecture Notes in Computer Science, pages 92–98, Berlin / Heidelberg, 2007. Springer Science+Business Media Deutschland GmbH. ISBN 978-3-540-74627-0.

Pajas, Petr and Jan Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In Hinrichs, Richard Erhard, Nancy Ide, Martha Palmer, and James Pustejovsky, editors, *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Genova, Italy, 2006. ISBN 2-9517408-2-4.

Popel, Martin. Ways to Improve the Quality of English-Czech Machine Translation. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, 2009.

Ralf, Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147. ELRA, 2006.

Rambousek, Jiří, Jana Chamonikolasová, Daniel Mikšík, Dana Šlancarová, and Martin Kalivoda. KAČENKA (Korpus anglicko-český - elektronický nástroj Katedry anglistiky), 1997.

Sgall, Petr. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic, 1967.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 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, ACL 2007*, pages 67–74, Prague, 2007.

Štěpánek, Jan. Post-annotation Checking of Prague Dependency Treebank 2.0 Data. In *Proceedings of the 9th International Conference, TSD 2006*, number 4188 in Lecture Notes in Computer Science, pages 277–284, Berlin / Heidelberg, 2006. Springer-Verlag Berlin Heidelberg. ISBN 3-540-39090-1.

Tiedemann, Jörg and Lars Nygaard. The OPUS corpus - parallel & free. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, May 26–28 2004. URL `http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf`.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria, 2005.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, 2008. Association for Computational Linguistics.