The Joy of Parallelism with CzEng 1.0

**Ondřej Bojar, Zdeněk Žabokrtský,
Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček,
Jiří Maršík, Michal Novák, Martin Popel, Aleš Tamchyna**

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
surname@ufal.mff.cuni.cz *except* {mnovak,odusek}@ufal.mff.cuni.cz, jiri.marsik89@gmail.com

Abstract

CzEng 1.0 is an updated release of our Czech-English parallel corpus, freely available for non-commercial research or educational purposes. In this release, we approximately doubled the corpus size, reaching 15 million sentence pairs (about 200 million tokens per language). More importantly, we carefully filtered the data to reduce the amount of non-matching sentence pairs.

CzEng 1.0 is automatically aligned at the level of sentences as well as words. We provide not only the plain text representation, but also automatic morphological tags, surface syntactic as well as deep syntactic dependency parse trees and automatic co-reference links in both English and Czech.

This paper describes key properties of the released resource including the distribution of text domains, the corpus data formats, and a toolkit to handle the provided rich annotation. We also summarize the procedure of the rich annotation (incl. co-reference resolution) and of the automatic filtering. Finally, we provide some suggestions on exploiting such an automatically annotated sentence-parallel corpus.

Keywords: Czech-English parallel corpus, automatic parallel treebank, training data for machine translation

1. Introduction

We present the new release of a Czech-English parallel corpus with rich automatic annotation, CzEng 1.0.[1]

CzEng 1.0 is a replacement for CzEng 0.9 (Bojar et al., 2010) which was successfully used in various NLP experiments including the machine translation evaluation campaigns of 2010 and 2011 (Callison-Burch et al., 2010; Callison-Burch et al., 2011).[2] Both the old and the new release are freely available for research purposes; restricted versions of CzEng 0.9 have also their commercial applications. With 8 million parallel sentences, CzEng 0.9 moved Czech out of the "low resource" rank of languages. While we did not primarily focus on increasing the overall size of the corpus, CzEng 1.0 nevertheless doubled the size of parallel Czech-English data available for research. More details are available in Section 2.

In CzEng 1.0, our main aim was to improve the quality of the resource. We focused on:

- User access to the rich annotation (Section 3.),
- Improved rich annotation, including automatic co-reference (Section 4.),
- Filtering of the sentence pairs to increase the precision of the corpus (Section 5.).

We believe this large and richly annotated resource will be of interest not only to the machine translation community but also to many other NLP researchers. Our first examples utilizing the parallelism (aside from the obvious applications in machine translation) are given in Section 6.

[1] http://ufal.mff.cuni.cz/czeng/

[2] http://www.statmt.org/wmt10, http://www.statmt.org/wmt11

2. Core CzEng 1.0 Properties

This section is devoted to basic statistics of the released resource, data sectioning and file formats.

2.1. CzEng 1.0 Data Sizes

Table 1 lists the total number of parallel sentences and Czech and English surface tokens per source. Please note that the number of tokens includes punctuation marks and other symbols.

In Table 1, we also list the number of nodes in the deep syntactic layer of representation (see Section 4.), which roughly correspond to content words in the sentences. We can see that English uses about 12% more surface tokens than Czech. The numbers of deep nodes in Czech and English are much closer. The higher number of deep nodes observed for Czech can be attributed to the fact that the procedure of adding artificial nodes for dropped pronouns and similar phenomena is more elaborated in our annotation pipeline than the similar procedure for English.

2.2. CzEng 1.0 Data Structure

CzEng 1.0 is shuffled at the level of "blocks", sequences of not more than 15 consecutive sentences from one source. The original documents thus cannot be reconstructed but some information about cross-sentence phenomena is preserved. Specifically, CzEng includes Czech and English grammatical and textual co-reference links that do span sentence boundaries (see Section 4.2.).

Each "block" comes from one of the text domains (EU Legislation, etc., see Table 1) and the domain is indicated in the sentence ID.

Individual text "blocks", shuffled, are combined to numbered files; each file holds about 200 sentence pairs.

Source Domain	Parallel Sentences	Surface Tokens ("Words+Punct.") Czech	English	Deep Nodes ("Content Words") Czech	English
Fiction	4,335 k	57,177 k	64,264 k	41,142 k	38,690 k
EU Legislation	3,993 k	78,022 k	87,489 k	56,446 k	52,718 k
Movie Subtitles	3,077 k	19,572 k	23,354 k	14,615 k	14,918 k
Parallel Web Pages	1,884 k	30,892 k	35,455 k	23,141 k	22,057 k
Technical Documentation	1,613 k	16,015 k	16,836 k	11,942 k	11,207 k
News	201 k	4,280 k	4,737 k	3,208 k	2,963 k
Project Navajo	33 k	484 k	557 k	363 k	344 k
Total	15,136 k	206,442 k	232,691 k	150,857 k	142,897 k

Table 1: Sources in CzEng 1.0, including data sizes in thousands.

The files are further organized into 100 similarly-sized sections, the last two of which are designated for development and testing purposes: `00train`, ..., `97train`, `98dtest`, `99etest`. Users of CzEng 1.0 are kindly asked to avoid training on these last 2% of the data.

2.3. CzEng 1.0 File Formats

CzEng 1.0 is available in three data formats: rich Treex XML format, "export format", and parallel plain text.

2.3.1. Treex Format

The primary data format of CzEng 1.0 is the Treex XML, a successor to the TectoMT TMT format used in CzEng 0.9. Treex XML can be processed using the Treex platform or manually browsed in the TrEd tree editor, see Section 3. for details. Users are encouraged to use the Treex toolkit and access the data programmatically using Treex API rather than directly parsing the XML.

2.3.2. Export Format

To facilitate the access to most of the automatic rich annotation of CzEng 1.0 without any XML hassle, we provide the data also in a simple "factored" line-oriented export format. Note that e.g. named entities or co-reference links are not available in the export format at all.
An example and the meaning of all the tab-delimited columns of the export format is given in Table 5 at the end of the paper.

2.3.3. Plaintext Format

The plaintext format is very simple, consisting of just four tab-delimited columns: sentence pair ID, filter score, Czech sentence, and English sentence.
The plain text preserves the original tokenization (i.e. no tokenization) of the source data.

2.4. Brief Summary of the Automatic Annotation

The processing pipeline of CzEng 1.0 was in essence very similar to the the pipeline used in CzEng 0.9, although we replaced some of the tools with their updated versions.

1. The original texts were segmented into sentences using TrTok, see Section 6.1. (preserving the original tokenization).

2. Sentence alignment was obtained using Hunalign (Varga et al., 2005), where we tokenized, lowercased and chopped each token to at most 4 characters to reduce the sparseness of esp. Czech vocabulary. Hunalign was run on each document pair separately and without any shared translation dictionary.

3. All sentences were morphologically tagged and lemmatized with the tools available in the Treex platform (the Morce tagger (Spoustová et al., 2007) and a rule-based lemmatizer for English).

4. We applied GIZA++[3] (Och and Ney, 2000) to obtain alignment between surface tokens. To reduce the data sparseness, GIZA++ was run on Czech and English lemmas, not fully inflected word forms. We aligned all the data in one large process, which needed about 2 days of CPU time to finish. As common in statistical machine translation, GIZA++ was applied in both translation directions and the two unidirectional alignments were symmetrized. We provide outputs of several symmetrization techniques.

5. The word alignment was loaded into the Treex format and all subsequent steps of analysis were carried out within the Treex framework. MST parser (McDonald et al., 2005) was used for surface syntax dependency parsing.

2.4.1. A Note on Node Alignment

Besides the word alignment, CzEng 1.0 is provided with automatic alignment on the tectogrammatical layer as well. Unlike in CzEng 0.9, where the tectogrammatical alignment was created by the trainable *TAlign* tool (Mareček, 2009), the alignment links in CzEng 1.0 are simply projected from GIZA++ intersection word alignment to the corresponding tectogrammatical trees. The number of links produced by this simple projection is higher, which causes higher recall but lower precision.

3. Treex Framework for CzEng 1.0

As mentioned above, all the automatic annotation of CzEng 1.0 was carried out using the Treex multi-purpose NLP framework (Popel and Žabokrtský, 2010).[4] The core modules of the framework are freely available and can be in-

[3] `http://http://code.google.com/p/giza-pp/`
[4] `http://ufal.mff.cuni.cz/treex`

```
# Convert treex.gz to CoNLL format
treex Write::CoNLLX language=en to=f00001en.conll \
      Write::CoNLLX language=cs to=f00001cs.conll \
      -- data.treex-format/00train/f00001.treex.gz

# See the most frequent translations
treex -Lcs Util::Eval tnode='my ($en)=$tnode->get_aligned_nodes_of_type("int");
                             say $tnode->t_lemma . "\t" . $en->t_lemma if $en' \
      -- data.treex-format/00train/f0000?.treex.gz \
 | sort | uniq -c | sort -rn | head -n 20
# prints:
#    593 a           and
#    291 #PersPron   #PersPron
#    222 být         be

# Open a file in the TrEd editor (via a wrapper to support Treex file format)
ttred data.treex-format/00train/f00001.treex.gz
```

Figure 1: Examples of using the Treex command-line interface.

stalled from CPAN.[5] There are a number of NLP tools integrated in Treex, such as morphological taggers, lemmatizers, named entity recognizers, dependency parsers, constituency parsers, and various kinds of dictionaries.

For users of CzEng 1.0, the Treex platform offers a versatile API, a more appropriate way of accessing the Treex XML files than generic XML parsers can offer. Aside from custom export procedures, one can use ready-made *writers* available in Treex. Figure 1 shows how to convert the surface dependency trees to CoNLLX format or emit the most frequent pairs of tectogrammatical lemmas.

The Treex platform also provides a simple wrapper for TrEd,[6] a tree editor which can read Treex XML using a designated plug-in module. TrEd offers the best option for manual inspection of CzEng data.

Figure 2 shows a sample sentence pair (English and Czech) annotated on both analytical (surface syntax, *a-tree*) and tectogrammatical (deep syntax, *t-tree*) layers. The morphological annotation is stored together with the analytical annotation.

4. Rich Annotation

CzEng 1.0 is automatically annotated in the same theoretical framework as the Prague Dependency Treebank (PDT) 2.0 (Hajič, 2004). Many small updates of various annotation steps have happened since CzEng 0.9. Here we focus on the two more complex ones at the deep syntactic layer (also called *tectogrammatical* or *t-layer*): formemes (Section 4.1.) and automatic co-reference (Section 4.2.).

4.1. Formemes

In addition to the PDT 2.0 annotation style attributes, each node at the t-layer is assigned a *formeme* (Ptáček and Žabokrtský, 2006; Žabokrtský et al., 2008) describing its morphosyntactic form, including e.g. prepositions, subordinate conjunctions, or morphological case. The set of possible formemes contains values such as:

- `n:subj`—an English noun in subject position,
- `v:to+inf`—an English infinitive clause with the particle *to*,
- `adj:attr`—attributive adjectives in both languages, or
- `n:k+3`—a Czech noun in dative (third) case with the preposition *k*.

Figure 3 gives an example of other formemes in a sentence. The values are filled in using rule-based modules operating on both t-trees and the corresponding a-trees.

The formeme annotation had already been present in the previous versions of CzEng and had been successfully employed in structural MT (Žabokrtský et al., 2008) and Natural Language Generation (Ptáček and Žabokrtský, 2006) tasks. We use a version improved (mostly on the Czech side) to depict various linguistic phenomena more accurately and to maintain a greater consistency across the two languages (see Section 6.2. for a cross-lingual evaluation). Our modifications involve e.g. treating nominal usages of adjectives as nouns, distinguishing nominal and adjectival numerals, marking case in Czech adjectival complements of verbs, or allowing prepositions with most English verb forms, plus several fixes for erroneous marking of the previous versions.

4.2. Co-Reference Links

In one of the last stages of automatic annotation, the co-reference resolution is performed on both language parts of the corpus. The range of co-reference types annotated in CzEng corresponds to the types present in PDT 2.0 and on the English side of PCEDT 2.0. Namely, it captures the so-called grammatical co-reference and pronominal textual co-reference.

[5] `http://search.cpan.org/search?query=treex`

[6] `http://ufal.mff.cuni.cz/tred/`

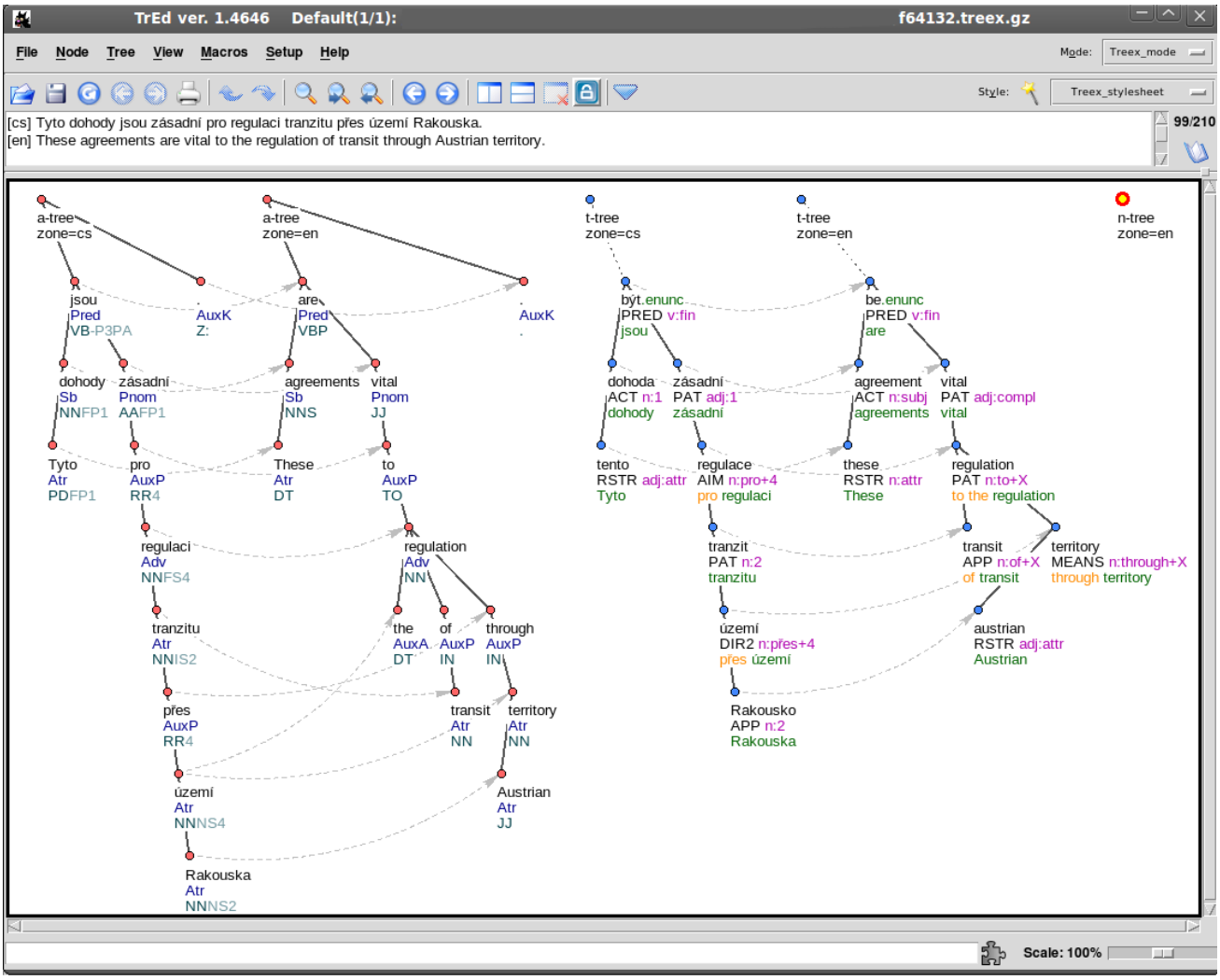

Figure 2: Visualization of one sentence pair in TrEd (Tree Editor). Czech a-tree, English a-tree, Czech t-tree, and English t-tree are presented (left to right). Other attributes which are not shown (e.g. grammatemes) can be inspected after clicking the nodes.

There	*is*	*no*	*asbestos*	*in*	*our*	*products*	*now*	.	”
	be	no	asbestos		#PersPron	product	now		
	v:fin	**n:attr**	**n:obj**		**n:poss**	**n:in+X**	**adv**		

Figure 3: An example sentence with tectogrammatical lemmas and formemes

Grammatical co-reference comprises several subtypes of relations, which mainly differ in the nature of referring expressions (e.g. relative pronoun, reflexive pronoun). However, all of them have in common that they appear as a consequence of language-dependent grammatical rules. This fact allows us to resolve them with a relatively high success rate, using the rule-based system proposed by Nguy (2006). For instance, given a relative pronoun that introduces a relative clause, the parent of the clause head is marked as an antecedent of the pronoun.

On the other hand, the arguments of textual co-reference are not realized by grammatical means alone, but also via context (Mikulová et al., 2006), which makes the resolution far more difficult. To indentify textual co-reference relations with a personal pronoun as the referring expression, we incorporated the perceptron ranking system of Nguy et al. (2009). On the Czech side, we employed the original feature set and trained the system on the PDT data. We used the English side of PCEDT to train the English system, for which we had to limit and modify several features to comply with a somewhat different annotation style.

Table 2 shows the values of pairwise precision, recall and F-score of co-reference resolution on the evaluation part of PDT and PCEDT for Czech and English, respectively. On Czech gold standard trees, the scores are close to those reported by Nguy et al. (2009). Since CzEng annotation is completely automatic, it is necessary to measure the success rate on automatically analyzed trees, so that we can reliably assess the quality of co-reference annotation in CzEng. Unfortunately, one can observe a substantial drop for automatic trees. The reason is twofold.

First, Czech is a pro-drop language, thus the pronouns must be reconstructed on the tectogrammatical layer. Nonetheless, the number of personal pronouns reconstructed incor-

	Gold Standard Features			Automatic Features			Oracle Gender and Number		
Language	P	R	F	P	R	F	P	R	F
Czech	77.06	77.58	77.32	55.23	46.14	50.28	65.70	54.89	59.81
English	45.52	58.69	51.27	44.53	57.32	50.12	–	–	–

Table 2: Results of the co-reference resolution evaluation. The precision, recall and F-score were measured on both languages using the features coming either from the gold standard or the automatic annotation. In the last three columns, the features were automatic except for the manual gender and number.

rectly or not at all accounts for 25% of all pronouns elided on the surface layer (and 15% of all personal pronouns).
Second, gender and number of some pronouns cannot be disambiguated without the knowledge of co-reference links. At the same time, gender and number information is one of the most valuable features in our co-reference resolver. While all attributes are disambiguated in manually annotated trees, they are left ambiguous in automatically analyzed data, which certainly decreases the quality of co-reference resolution. This claim is confirmed by our oracle experiment: when we replaced the automatic gender and number with the manually assigned values, the F-score improved by almost 10% absolute (see the last three columns of Table 2).
As regards the co-reference resolution in English, the difference between its quality using manual and automatic trees is not as dramatic as in Czech. This further confirms the above-mentioned reasons for the success rate drop in Czech since both of the issues (pro-drop recovery and gender and number disambiguation) are marginal in English.
We would like to emphasize that the presented experiments on co-reference resolution are to our knowledge the first for Czech using no gold standard features and one of a few for English employing the deep syntactic layer.

5. Filtering Sentence Pairs

The amount of data included in CzEng along with the varying reliability of its sources (such as volunteer-submitted movie subtitles) demand an automatic method for recognizing and filtering out bad sentence pairs.
Simple filters have been used in previous editions of CzEng. Details about their evaluation and suggestions for improvements can be found in Bojar et al. (2010). We extend the previous work by adding several new filters and introducing a robust method for their combination.
Filtering features for CzEng 1.0 exploit all layers of automatic annotation and include:

- indication of Czech and English sentences' identity,
- lengths of sentences and the words contained in them,
- no Czech (English) word on the Czech (English) side,
- various checks for remains of meta-information, such as HTML tags or file paths,
- use of a translation dictionary to determine the coverage of English words by the Czech side,
- score of symmetrized automatic word alignment obtained by GIZA++,

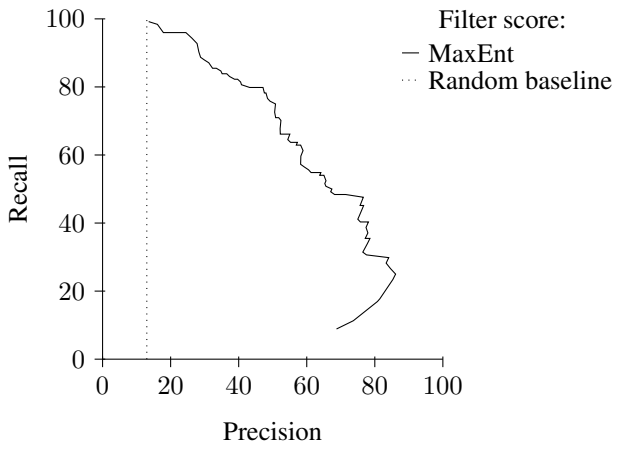

Figure 4: Precision and recall of CzEng filters.

- matching part-of-speech tags,
- matching grammatical number, verb tense or presence of comparative/superlative modifiers.

Wherever possible, we try to model the feature as a ratio or score and empirically find interval bounds for its quantization.
The features are combined to form a single score using a classifier trained to distinguish between correct and wrong sentence pairs. We evaluated the performance of decision trees, naive Bayes classifier, and maximum entropy classifier. We found the maximum entropy classifier to be best suited for this setting. Figure 4 shows the trade-off between precision and recall for all threshold settings. Note that the random baseline stays at roughly 13% regardless of the threshold—our evaluation data consists of 1000 manually annotated sentence pairs, out of which 124 were marked as wrong.

5.1. Precision-Size Trade-off for CzEng Users

Since our filter combination is still not reliable, we include all sentences that pass the threshold of 0.3 in CzEng 1.0, favoring precision of the filtration over recall. We also provide the score assigned by our filters to each sentence pair so that users can create a cleaner, more strictly filtered subset of CzEng 1.0.
Moreover, 2330 input documents containing 60% or more sentences with scores below the threshold were discarded entirely.

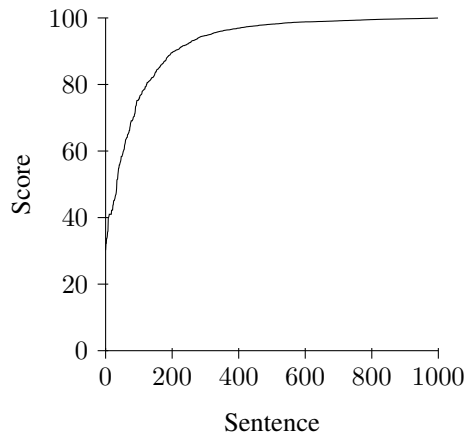

Figure 5: Distribution of sentence filter scores in a random 1000-sentence sample.

Corpus and Domain	Sents	BLEU	Vocab. [k] En	Cs
CzEng 0.9	all 1M	14.77±0.12	187	360
CzEng 1.0		**15.23±0.18**	221	396
CzEng 0.9	news 100k	**14.34±0.05**	53	125
CzEng 1.0		14.01±0.13	47	113

Table 3: Results of MT evaluation.

	Formeme Detection on Automatic Trees	Manual Trees
Baseline	1.5981	1.6680
Improved	**1.6873**	**1.7092**

Table 4: The impact of an improved design of formemes on mutual information (in bits) of Czech and English formemes of aligned t-tree nodes.

5.2. Evaluation of Data Quality

The distribution of filter scores in sentence pairs as shown in Figure 5 suggests that most of the corpus is clean, containing correct sentence pairs.

We also evaluated the quality of CzEng 1.0 extrinsically by conducting a small machine translation experiment. We trained contrastive phrase-based Moses SMT systems (Koehn et al., 2007)—the first one on 1 million sentence pairs from CzEng 0.9, the other on the same amount of data from CzEng 1.0. Another contrastive pair of MT systems was based on small in-domain data only: 100k sentences from the *news* sections of CzEng 0.9 and 1.0, respectively. For each setting, we extracted the random sentence pairs 5 times to avoid drawing conclusions from possibly biased data selection.

For tuning and evaluation, test sentences from WMT 2010 and 2011 were used, respectively. These sets are from the news domain. We used the News Crawl Corpus 2011 data to train the language model.

We measure the translation quality using the standard SMT metric BLEU (Papineni et al., 2002). Table 3 shows the mean BLEU score and standard deviation for each data set. In the setting with 1 million random sentence pairs, using data from CzEng 1.0 is noticeably beneficial for MT quality—the absolute BLEU gain is roughly 0.4 points. This improvement stems from the overall quality of the data, the distribution of domains in CzEng 1.0 is also likely to play a certain role.

On the other hand, using only the news data reverses the situation—CzEng 1.0 data lead to a system with slightly worse performance. We verified our results using Welch two-sample t-test and found that in both cases the difference is statistically significant on 99% confidence level.

An explanation is suggested by the last two columns. The filtering has probably caused a loss in vocabulary size (distinct token types) for both English and Czech in the news domain but not across domains.

6. The Joy of Parallelism

Here we mention several steps in CzEng automatic annotation that make use of the parallel data for improved output quality.[7]

6.1. Tokenizer

CzEng 1.0 uses TrTok, a fast re-implementation of the trainable tokenizer (Klyueva and Bojar, 2008) for sentence segmentation. Its main advantage is the fact that different data sources may need different segmentation patterns (e.g. legislation texts need segment breaks after commas) and TrTok can be guided to follow the patterns by providing enough sample data in the desired form.

By examining segments that were aligned to 1-2 and 2-1 clusters, we often find them to be a consequence of a mismatch in segmentation rules for Czech and English. Such snippets of parallel data can thus directly serve as additional training data for TrTok.

6.2. Formemes

Table 4 compares the mutual information (MI) of Czech and English formemes of t-tree nodes aligned one-to-one for the baseline set of formemes and the improved set of formemes measured on the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0, (Bojar et al., 2012)).[8] The higher the MI, the easier the transfer phase in structual machine translation (Žabokrtský et al., 2008) is expected. We measure the MI in two setups—we either utilize the manual trees provided in PCEDT 2.0 directly,[9] or take just the sentences from PCEDT 2.0 and apply to them the automatic annotation pipeline which we use for the whole CzEng 1.0 corpus.

Our initial measurements showed a slight MI drop on the automatic trees, which led us to the discovery of several bugs in both formeme detection rules and conversion of a-trees to t-trees (e.g. problems with infinitive and passive verb forms detection or coordinated modal verbs).

[7] We leave aside the joy of parallel *processing* of the data, very useful i.a. in debugging on large datasets.

[8] `http://ufal.mff.cuni.cz/pcedt2.0`

[9] The used t-trees were manual for both languages; however, only automatic a-trees are available on the Czech part in the PCEDT 2.0.

The corrected analysis pipeline and formeme detection show an MI increase for both manual and automatic trees (see Table 4), which indicates that the new set of formemes is likely to improve the MT transfer phase. Again, we used here the parallel view to fine-tune a monolingual processing step.

6.3. Co-Reference—Future Work

The automatic co-reference annotation for one of the languages in the parallel corpus could be improved if we employed the information from the other language side.
English is considered to be lacking grammatical gender (except for pronouns) and the majority of nouns in English are referred to by a pronoun in neuter gender. On the other hand, Czech distinguishes between four grammatical genders whose distribution among nouns is rather balanced and, moreover, personal pronouns usually agree in gender with a noun they co-refer with.
Thus, we suggest to incorporate the results of Czech co-reference resolution into the English resolver, which might limit the number of antecedent candidates that are in consideration. Conversely, Czech is a pro-drop language, which allows us to utilize the English side to potentially project some of the pronouns that are elided in Czech.

7. Conclusion

We presented CzEng 1.0, the new release of a large Czech-English parallel corpus with rich automatic annotation. The corpus is freely available for non-commercial research and educational purposes at our web site:

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

CzEng 1.0 can serve as large training data for linguistically uninformed approaches, e.g. to machine translation, but it can also be directly used in experimenting with cutting-edge NLP tasks such as co-reference resolution validated across languages. We have also provided two examples of exploiting the parallelism of the data to improve monolingual processing: sentence segmentation and formeme definition.

8. Acknowledgements

The work on this project was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic), Czech Science Foundation grants P406/10/P259 and 201/09/H057, GAUK 4226/2011, 116310, and the FAUST project (FP7-ICT-2009-4-247762 of the EU and 7E11041 of the Czech Republic). This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

9. References

Ondřej Bojar, Adam Liška, and Zdeněk Žabokrtský. 2010. Evaluating Utility of Data Sources in a Large Parallel Czech-English Corpus CzEng 0.9. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 447–452, Valletta, Malta, May. ELRA, European Language Resources Association.

Ondřej Bojar, Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

David Mareček. 2009. Improving word alignment using alignment of deep structures. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, volume 5729 of *Lecture Notes in Computer Science*, pages 56–63. Springer Berlin / Heidelberg. 10.1007/978-3-642-04208-9_11.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk

Col.	Sample	Explanation
1	subtitles-b2-00train-f00001-s8	ID specifying the domain, block number, train/dev/test section, file number and sentence within the file.
2	0.99261036	Filter score indicating the quality of the sentence pair. The score of 1 is perfect pair, pairs below 0.3 are removed.
	Czech	

3	`Zachránil|zachránit_:W|VpYS---XR-AA---|1|0|Pred`
	`mi|já|PH-S3--1-------|2|1|Obj můj|můj|PSYS1-S1-------|3|5|Atr`
	`milovaný|milovaný_^(*2t)|AAIS1----1A----|4|5|Atr`
	`krk|krk|NNIS1-----A----|5|1|Obj .|.|Z:----...`
		Czech a-layer (surface-syntactic tree) in factored form: word-form|lemma|morphological-tag|index-in-sentence|index-of-governor|syntactic-function.

4	`zachránit|PRED|1|0|complex|v:fin|v|-|neg0|ant|ind|decl|-|...` `#PersPron|ADDR|2|1|complex|n:3|n.pron.def.pers|sg|-|-|...`	Czech t-layer (tectogrammatical tree): t-lemma|functor|index-in-tree|index-of-governor|nodetype|formeme|semantic-part-of-speech|... and many detailed t-layer attributes.
5	0-0 1-1 2-2 3-3 4-4	Correspondence between Czech a-layer and t-layer for content words. Indexed from 0.
6		Correspondence between Czech a-layer and t-layer for auxiliary words. Indexed from 0.
	English	

7	`He|he|PRP|1|2|Sb saved|save|VBD|2|0|Pred my|my|PRP$|3|4|Atr`
	`ever-lovin|ever-lovin|NN|4|6|Atr '|'|''|5|6|AuxG`
	`neck|neck|NN|6|2|Obj .|.|.|7|0|AuxK`
		English a-layer (surface-syntactic tree) in factored form: word-form|lemma|tag|index-in-sentence|index-of-governor|syntactic-function.

8	`#PersPron|ACT|1|2|complex|n:subj|n.pron.def.pers|sg|-|-|...`
	`save|PRED|2|0|complex|v:fin|v|-|neg0|ant|ind|decl|-|-|...`
	`#PersPron|APP|3|4|complex|n:poss|n.pron...`
		English t-layer (tectogrammatical tree): t-lemma|functor|index-in-tree|index-of-governor|nodetype|formeme|semantic-part-of-speech|... and many detailed t-layer attributes.

9	0-0 1-1 2-2 3-3 5-4	Correspondence between English a-layer and t-layer for content words. Indexed from 0.
10	4-4	Correspondence between English a-layer and t-layer for auxiliary words. Indexed from 0.
	Cross-Language Alignments Between Surface Czech and English Always indexed from 0, Czech-English.	
11	0-1 1-2 2-2 3-3 4-5 5-6	GIZA++ alignments "there" for cs2en.
12	0-0 0-1 2-2 3-3 3-4 4-5 5-6	GIZA++ alignments "back" for cs2en.
13	0-0 0-1 1-2 2-2 3-3 3-4 4-5 5-6	GIZA++ alignments symmetrized using grow-diag-final-and for cs2en.
14	0-0 0-1 1-2 2-2 3-3 3-4 4-5 5-6	GIZA++ alignments symmetrized using grow-diag-final-and for en2cs (not the inverse of column 13).
	Cross-Language Alignments Between T-Layer Czech and English Always indexed from 0, Czech-English.	
15	0-1 1-2 2-2 3-3 4-4	T-alignment "there" for cs2en.
16	0-0 0-1 2-2 3-3 4-4	T-alignment "back" for cs2en.
17		Additional rule-based t-alignment linking esp. generated nodes like `#Perspron;.`

Table 5: An example and explanation of the "export format" of CzEng 1.0. Each row in the table corresponds to one tab-delimited column of the line-oriented text files.

Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep.

Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. 2009. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285, London, UK, September. ACL.

Giang Linh Nguy. 2006. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, MFF UK, Prague, Czech Republic. In Czech.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *LNCS*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Jan Ptáček and Zdeněk Žabokrtský. 2006. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Text, Speech and Dialogue*, pages 221–228. Springer.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. 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: Information Extraction and Enabling Technologies*, ACL '07, pages 67–74, Stroudsburg, PA. Association for Computational Linguistics.

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

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA.