# New Language Pairs in TectoMT

**Ondřej Dušek,*** **Luís Gomes,‡** **Michal Novák,*** **Martin Popel,*** and **Rudolf Rosa***
*Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
`{odusek,mnovak,popel,rosa}@ufal.mff.cuni.cz`
‡University of Lisbon, Faculty of Sciences, Department of Informatics
`luis.gomes@di.fc.ul.pt`

## Abstract

The TectoMT tree-to-tree machine translation system has been updated this year to support easier retraining for more translation directions. We use multilingual standards for morphology and syntax annotation and language-independent base rules. We include a simple, non-parametric way of combining TectoMT's transfer model outputs.

We submitted translations by the English-to-Czech and Czech-to-English TectoMT pipelines to the WMT shared task. While the former offers a stable performance, the latter is completely new and will require more tuning and debugging.

## 1 Introduction

The TectoMT tree-to-tree machine translation (MT) system (Žabokrtský et al., 2008) has been competing in WMT translation tasks since 2008 and has seen a number of improvements. Until now, the only supported translation direction was English to Czech. This year, as a part of the QTLeap project,[1] we have enhanced TectoMT and its underlying natural language processing (NLP) framework, Treex (Popel and Žabokrtský, 2010), to support more language pairs. We simplified the training pipeline to be able to retrain the translation models faster, and we use abstracted language-independent rules with the help of Interset (Zeman, 2008) where possible.

Together with our partners on the QTLeap project, we have implemented translation systems for other language pairs (English to and from Dutch, Spanish, Basque, and Portuguese) which are not part of WMT shared Translation Task this year. However, we were also able to submit the results of a newly built Czech-English translation system in the shared task. The performance of the current version leaves a lot of room for improvement, but proves the potential of TectoMT for different language pairs.

The original TectoMT system for English-Czech translation has seen just small changes, e.g., adding specialized translation models for selected pronouns (Novák et al., 2013a; Novák et al., 2013b) and fine-tuning of a handful of rules. Therefore, its performance is virtually identical to that of the last year's version.

This paper is structured as follows: in Section 2, we introduce the TectoMT basic architecture. In Section 3, we describe the improvements to TectoMT that were added for an easier support of new language pairs. Section 4 then details the Czech-to-English TectoMT system submitted to WMT15. We discuss TectoMT's performance in the task and examine the most severe error sources in Section 5. Section 6 then concludes the paper.

## 2 The TectoMT Translation System

TectoMT (Žabokrtský et al., 2008) is a tree-to-tree MT system system consisting of an analysis-transfer-synthesis pipeline, with transfer on the level of deep syntax. It is based on the Prague Tectogrammatics theory (Sgall et al., 1986) and distinguishes two levels of syntactic description (see Figure 1):

- *Surface dependency syntax* (*a-layer*) – surface dependency trees containing all the tokens in the sentence.
- *Deep syntax* (*t-layer*) – dependency trees that contain only content words (nouns, main verbs, adjectives, adverbs) as nodes. Each node has a deep lemma (*t-lemma*), a semantic function label (*functor*), a morpho-syntactic form label (*formeme*), and various grammatical attributes (*grammatemes*), such as number, gender, tense, or modality.

[1] `http://qtleap.eu`

*Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 98–104,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

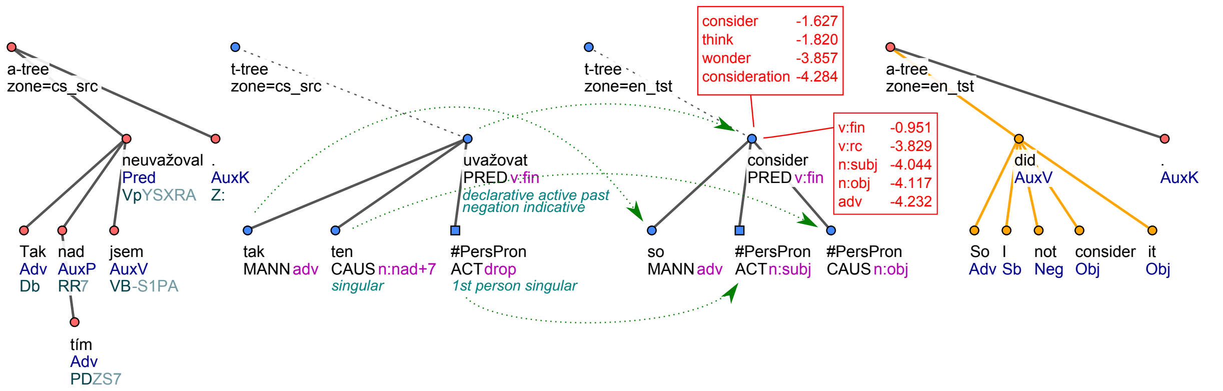

Figure 1: Example TectoMT translation.
From the left to the right: (1) source Czech sentence analyzed to surface dependencies (a-layer), (2) Czech sentence analyzed to deep syntax (t-layer), with t-lemmas (black), functors (capitals), formemes (purple), and grammatemes (teal), (3) translated English t-layer tree (with MaxEnt model logarithmic probabilities for t-lemmas and formemes shown in red for a selected node), (4) generated English surface dependency tree.

Formemes are not part of the t-layer according to the original theory; they have been added in TectoMT to work around the difficult task of functor assignment (semantic role labeling). Formemes are much simpler to obtain – they are assigned by rules based on the surface dependency trees (Dušek et al., 2012). Apart from a few specific cases, functors are not used in TectoMT, and formemes are used instead.

T-layer representations of the same sentence in different languages are closer to each other than the surface texts; in many cases, there is a 1:1 node correspondence among the t-layer trees. TectoMT's transfer exploits this by translating the tree isomorphically, i.e., node-by-node and assuming that the shape will not change in most cases (apart from a few exceptions handled by specific rules).

The translation is further factorized – t-lemmas, formemes, and grammatemes are translated using separate models. The t-lemma and formeme translation models are an interpolation of maximum entropy discriminative models (MaxEnt) of Mareček et al. (2010) and simple conditional probability models. The MaxEnt models are in fact an ensemble of models, one for each individual source t-lemma/formeme. The combined translation models provide several translation options for each node along with their estimated probability (see Section 1). The best options are then selected using a Hidden Markov Tree Model (HMTM) with a target-language tree model (Žabokrtský and Popel, 2009), which roughly corresponds to the target-language $n$-gram model in phrase-based MT. Grammateme transfer is rule-based; in most cases, grammatemes remain the same as in the source language.

## 3 Adding New Language Pairs

Using different languages in an MT system with deep transfer is mainly hindered by differences in the analysis and synthesis of the individual languages. To overcome these problems, we decided to use existing multilingual annotation standards (see Section 3.1) and to simplify and automate translation model training (see Section 3.2). In addition, we introduce an easier way of combining the results of the individual translation models than HMTM (in Section 3.3).

### 3.1 Annotation Standards for Language Independence

We decided to use Interset (Zeman, 2008) as the standard morphological representation since its features capture all important morphological phenomena in many different languages, including all languages required in the QTLeap project. The Interset Perl library includes conversions from many commonly used language-specific tagsets. To represent surface dependency syntax, we use the HamleDT 1.5 annotation style (Zeman et al., 2012; Zeman et al., 2014), which also supports many different languages and comes with tools for the conversion of various pre-existing treebanks. This allows us to use existing taggers and parsers without retraining them – analyzed sentences are simply converted to Interset+HamleDT annotation style.

Most TectoMT/Treex rules for the conversion from surface dependencies to deep syntax (t-layer) have been adapted to expect Interset morphological features and HamleDT-style dependencies, which improves their usability for different languages. Their implementation involves a common language-independent base class and language-specific derived classes.[2]

For t-layer representation, we stick to the TectoMT annotation style as used for Czech and English, which is originally based on PDT and Prague Czech-English Dependency Treebank annotation (Hajič et al., 2006; Hajič et al., 2012). However, we are aware that this annotation style has problems in other languages (e.g., grammatemes cannot express all required grammatical meaning), and that changing or extending it will probably be required.

### 3.2 Support for Training New Language Pairs

Other improvements to support adding new language pairs quickly are rather technical. We automated the translation model training in a set of makefiles. To train a new translation pair, one only needs to implement analysis and synthesis pipelines for both languages and edit a configuration file. Debugging and testing of the new analysis and synthesis pipelines is supported by monolingual "roundtrip" experiments: a development data set is first analyzed up to t-layer, then synthesized back to word forms. BLEU score measurements (Papineni et al., 2002) and a direct comparison of the results are then used to improve performance before the translation models are trained and other transfer blocks are implemented.[3]

### 3.3 Combining Transfer Models More Simply

The t-lemma and formeme translation models are independent of each other to simplify their decisions and reduce data sparsity. This often results in the best translation alternatives suggested by both models being incompatible with each other, which leads to disfluent outputs.

In English-to-Czech translation, an HMTM is used to select compatible t-lemma–formeme pairs (see Section 2). However, the HMTM needs to be trained on a large monolingual data set annotated on the t-layer. To simplify and speed up development of TectoMT translation for new language pairs, we have introduced a simpler method of selecting a compatible t-lemma–formeme pair which does not require any training. In this approach, t-lemma and formeme probabilities of congruous pairs[4] are combined by a non-parametric function into a single score that is then used to select the best translation option. Incongruous combinations are discarded.[5]

We evaluated five non-parametric functions combining the two translation models' outputs:

- *AM-P* – arithmetic mean of probabilities,
- *GM-P* – geometric mean of probabilities,[6]
- *HM-P* – harmonic mean of probabilities,
- *GM-Log-P* – geometric mean of logarithmic probabilities,[7]
- *HM-Log-P* – harmonic mean of logarithmic probabilities.[8]

We compared the functions against a baseline of just using the first option given by each of the models (regardless of compatibility). We used corpora of 1,000 sentences from the IT domain collected in the QTLeap project to evaluate all variants in English-to-Czech, English-to-Spanish, and English-to-Portuguese translation. For the English-to-Czech direction, we could also compare our combination functions to using an HMTM. The results are given in Tables 1, 2, and 3 for English to Czech, Spanish, and Portuguese, respectively.

We can see that the performance of the individual variants is very similar and that they bring an improvement over the baseline in almost all cases.

[2] Some Czech and English TectoMT blocks have not been converted to Interset yet; they use the Czech positional tagset from the Prague Dependency Treebank (PDT) of Hajič et al. (2006) and the Penn Treebank tagset (Santorini, 1990).

[3] The "roundtrip" experiments are not necessarily needed for the translation. We just consider them a best practice which helps to quickly reveal bugs that could deteriorate the translation, but remain unnoticed for a long time.

[4] The "congruency" of t-lemma and formeme is based on the syntactic part-of-speech encoded in the formeme and the Interset part-of-speech of the t-lemma. There are five simple rules, e.g., verbal t-lemmas are compatible only with formemes beginning with "v:".

[5] The non-parametric functions are weaker than the HMTM with the target-language tree model, which considers the context of the parent t-lemma and models the compatibility with real-valued probabilities.

[6] Maximizing GM-P gives the same result as maximizing the product of probabilities $P(\textit{t-lemma}) \cdot P(\textit{formeme})$, which is the theoretically sound approach.

[7] Logarithmic probabilities are negative and geometric mean of two negative numbers is positive, so we actually use *negative* GM-Log-P, so the best option has the highest score.

[8] *AM-Log-P*, the arithmetic mean of logarithmic probabilities, seems to be missing from the list above, but since maximizing over AM-Log-P gives the same results as maximizing over GM-P, we omit AM-Log-P from our experiments.

| Function | NIST | BLEU |
|---|---|---|
| Baseline | 6.7500 | 0.2785 |
| HMTM | **6.8212** | **0.2876** |
| AM-P | 6.7602 | 0.2811 |
| GM-P | 6.7690 | 0.2818 |
| HM-P | **6.7713** | **0.2820** |
| GM-Log-P | 6.7707 | 0.2817 |
| HM-Log-P | 6.7580 | 0.2810 |

Table 1: NIST and BLEU scores for non-parametric combining functions in English-to-Czech translation.

| Function | NIST | BLEU |
|---|---|---|
| Baseline | 5.2757 | 0.1670 |
| AM-P | **5.4342** | 0.1808 |
| GM-P | 5.4315 | 0.1806 |
| HM-P | 5.4306 | 0.1806 |
| GM-Log-P | 5.4314 | **0.1809** |
| HM-Log-P | 5.4336 | 0.1808 |

Table 2: NIST and BLEU scores for non-parametric combining functions in English-to-Spanish translation.

HMTM in the English-to-Czech translation performs better as expected.

# 4 Czech to English Translation

This section is a detailed description of the TectoMT Czech-to-English translation pipeline as used in the WMT translation task. The analysis part (Section 4.1) is not new and thus is described only briefly, we focus more on the simple transfer (Section 4.2) and the English synthesis (Section 4.3).

| Function | NIST | BLEU |
|---|---|---|
| Baseline | 5.1584 | 0.1677 |
| AM-P | **5.2612** | **0.1719** |
| GM-P | 5.2219 | 0.1711 |
| HM-P | 5.0613 | 0.1620 |
| GM-Log-P | 5.2452 | 0.1719 |
| HM-Log-P | 5.2583 | 0.1719 |

Table 3: NIST and BLEU scores for non-parametric combining functions in English-to-Portuguese translation.

## 4.1 Czech Analysis

The Czech analysis is a slightly improved version of the pipeline used to train previous versions of the English-to-Czech translation in TectoMT as well as to analyze the Czech part of the CzEng 1.0 parallel corpus (Bojar et al., 2012).

The first part, the surface syntactic analysis, consists of a rule-based sentence segmenter and tokenizer, followed by a part-of-speech tagger – we use MorphoDiTa (Straková et al., 2014) in the current version – and a dependency parser (McDonald et al., 2005; Novák and Žabokrtský, 2007).

The surface dependency trees are then converted into deep syntactic (t-layer) trees using a series of mostly rule-based modules that collapse auxiliary words and decide upon the t-lemma, formeme, and grammatemes. They also reconstruct pro-drop pronoun subjects based on verbal morphology.

## 4.2 Transfer

The Czech-to-English transfer is relatively basic and does not contain many components besides the translation models for t-lemmas and formemes (see Section 2). Due to limited time to train the system for the new translation direction, we used the non-parametric t-lemma–formeme combination functions as described in Section 3.3 instead of a Hidden Markov Tree Model (cf. Section 2). We chose the HM-P setting based on performance on the development set.[9]

The additional components are rule-based and are listed below:

- Overrides and additions to the translation models, tuned on the development set,
- Removing Czech gender from common nouns not referring to persons,
- Fixing translation of names based on a lexicon compiled from Wikipedia (in particular, reverting the Czech female surname ending *-ová* in non-Czech names),
- Removing subjects of verbs where the translation model chose an infinitival form,
- Removing double negatives (which are the rule in Czech but not in English),

[9] We used the WMT news-test2012 data to tune our system.

- Fixing grammatemes, in particular number and negation, for some translations, such as *těstoviny* (pl.) → *pasta* (sg.), or *nedbalý* (negative) → *sloppy* (positive).

### 4.3 English Synthesis

The English synthesis (surface realization) pipeline has been newly developed for TectoMT translation into English; it is mostly rule-based and is inspired by the Czech synthesis pipeline. Besides the Czech-to-English translation, it is used in other TectoMT systems translating into English within the QTLeap project and in the TGen natural language generator (Dušek and Jurčíček, 2015).

In the synthesis pipeline, a new surface dependency (a-layer) tree is created as a copy of the source t-layer tree, with lemmas copied from t-lemmas and dependency labels, word forms, and morphology left undecided. All further changes are performed on the surface dependency tree, consulting information from the t-layer tree. The pipeline consists of the following steps:

1. Morphological attributes are filled in based on grammatemes.

2. Subjects are marked (to support subject-predicate agreement).

3. Basic English word order for declarative sentences is enforced. This only contains very general rules, e.g., SVO-order or adjective-noun order, but preliminary tests with source-language ordering from several different languages indicated that it is sufficient in most cases.

4. Subject-predicate agreement in number and person is enforced – predicates have their number and person filled based on their subject(s).

5. Auxiliary words are added. These are based on the contents of formemes (prepositions, subordinating conjunction, infinitive particles, possessive markers) and t-lemmas (phrasal verb particles).

6. English articles are added based on a handful of rules from an older surface realizer by Ptáček (2008).

7. Auxiliary verbs are added, expressing the voice, tense, and modality. Auxiliaries are also added for questions and sentences with existential *there*.

8. Imperative subjects are removed, question subjects are moved after the auxiliary verb.

9. Negation particles are added for verbs as well as selected adjectives and adverbs.

10. Punctuation is added to the end of the sentence, into coordinations and appositions, after clause-initial phrases preceding the subject, and in selected phrases (based on formemes).

11. Words are inflected based on their lemma and morphological attributes. We use rules for personal pronouns, MorphoDiTa (Straková et al., 2014) English dictionary for unambiguous words, and Flect (Dušek and Jurčíček, 2013) for all remaining words requiring inflection.[10]

12. The indefinite article *a* is changed into *an* based on the following word.

13. Repeated coordinated prepositions and conjunctions are deleted.

14. The first word in the sentence is capitalized.

The output sentence is then obtained by just combining all the nodes in the resulting surface dependency tree.

## 5 WMT 2015 Translation Task Results

TectoMT reached a BLEU score of 13.9 for the English-to-Czech direction in the WMT 2015 Translation Task. This ranks it among the last systems, which is consistent with results from previous years. However, English-to-Czech TectoMT has also been used in the Chimera system combination, which ranks first in both automatic and human evaluation results. TectoMT plays a very important role in Chimera (Tamchyna and Bojar, 2015).

TectoMT's Czech-to-English translation reached a BLEU score of 12.8, and finished last

[10] Alternatively, an n-gram language model *could* be used to select the word forms. Flect uses just a short context of neighboring lemmas, but it generalizes also to unseen words (thanks to morphological features). Currently, no n-gram language model is used in the whole TectoMT system.

in the automatic evaluation; human evaluation scores indicate a second-to-last position.

We believe that the major cause for the lower scores does not lie in TectoMT's basic architecture, but that improvements to translation models are required, as well as better tuning and debugging of the whole pipeline for the Czech-to-English direction. We examined closely a sample of the translation output (in both directions) and identified the following error sources:

- Translation models will require more tuning and possibly more powerful features. The English-to-Czech model leaves many relatively common words untranslated, which suggests that pruning has been too strict.[11]
- The non-parametric t-lemma–formeme combination functions are not ideal; training an HMTM will be necessary to improve English-to-Czech performance.
- Word ordering rules need to be improved, and more different cases need to be covered. We consider using a statistical ranker for local node ordering.
- The rule-based article assignment in English synthesis is lacking; indefinite articles are assigned much more often than they should be. This will probably not be possible without using a statistical module.

There are also other, rather technical issues related to punctuation or tokenization that will require more debugging.

## 6 Conclusions and Future Work

We presented TectoMT, a tree-to-tree machine translation system with deep transfer, and its new features in this year's edition of the WMT shared task, the main one being opening the system to new language pairs. TectoMT in the English-to-Czech direction is stable and provides useful translations though its results are worse than that of other systems; it is also used in the Chimera system combination. The new Czech-to-English system requires more development but shows that it is possible to adapt TectoMT to a new translation direction in a very short amount of time.

In future, we plan to tune the current Czech-to-English setup, and to include further improvements. We intend to use Interset instead of grammatemes on the t-layer to support categories of grammatical meaning not present in grammatemes (see Section 3.1). We also consider switching the TectoMT annotation style to Universal Dependencies. To improve translation models, we are planning to use Vowpal Wabbit (Langford et al., 2007) and to include word embeddings from word2vec (Mikolov et al., 2013) as features. We are also investigating the possibilities of non-isomorphic transfer in TectoMT.

## Acknowledgments

This work has been supported by the 7th Framework Programme of the EU grant QTLeap (No. 610516), and SVV project 260 104 and GAUK grants 2058214 and 338915 of the Charles University in Prague. It is using language resources hosted by the LINDAT/CLARIN Research Infrastructure, Project No. LM2010013 of the Ministry of Education, Youth and Sports.

## References

O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *LREC*, page 3921–3928, Istanbul.

O. Dušek and F. Jurčíček. 2013. Robust Multilingual Statistical Morphological Generation Models. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 158–164, Sofia. Association for Computational Linguistics.

O. Dušek and F. Jurčíček. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 451–461. Association for Computational Linguistics.

O. Dušek, Z. Žabokrtský, M. Popel, M. Majliš, M. Novák, and D. Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 267–274.

J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas,

[11] Same as for the English-to-Czech direction, the MaxEnt model was trained only for (source) lemmas occurring at least 100 times in the training data and only with translations (target lemmas) occurring at least 5 times. For the simple conditional ("static") model, we used the same constants (by mistake).

J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC*, pages 3153–3160, Istanbul.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková Razímová, and Z. Urešová. 2006. *Prague Dependency Treebank 2.0.* Number LDC2006T01. LDC, Philadelphia, PA, USA.

J. Langford, L. Li, and A. Strehl. 2007. Vowpal Wabbit online learning project. `http://hunch.net/~vw/`.

D. Mareček, M. Popel, and Z. Žabokrtský. 2010. Maximum entropy translation model in dependency-based mt framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206. Association for Computational Linguistics.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

M. Novák, A. Nedoluzhko, and Z. Žabokrtský. 2013a. Translation of “it” in a deep syntax framework. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofija, Bulgaria. Bălgarska akademija na naukite, Omnipress, Inc.

M. Novák, Z. Žabokrtský, and A. Nedoluzhko. 2013b. Two case studies on translating pronouns in a deep syntax framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1037–1041, Nagoya, Japan. Asian Federation of Natural Language Processing.

V. Novák and Z. Žabokrtský. 2007. Feature engineering in maximum spanning tree dependency parser. In *Text, Speech and Dialogue*, pages 92–98. Springer.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, page 311–318.

M. Popel and Z. Žabokrtský. 2010. TectoMT: modular NLP framework. *Advances in Natural Language Processing*, pages 293–304.

J. Ptáček. 2008. Two Tectogrammatical Realizers Side by Side: Case of English and Czech. In *Fourth International Workshop on Human-Computer Conversation*, Bellagio, Italy.

B. Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision). Technical Report No. MS-CIS-90-47, University of Pennsylvania Department of Computer and Information Science, Philadelphia, PA, USA.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.

J. Straková, M. Straka, and J. Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18. Association for Computational Linguistics.

A. Tamchyna and O. Bojar. 2015. What a transfer-based system brings to the combination with PBMT. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 11–20, Beijing, July. Association for Computational Linguistics.

Z. Žabokrtský and M. Popel. 2009. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148, Singapore. Association for Computational Linguistics.

Z. Žabokrtský, J. Ptáček, and P. Pajas. 2008. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, page 167–170. Association for Computational Linguistics.

D. Zeman, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2012. HamleDT: To parse or not to parse? In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

D. Zeman, O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

D. Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC*, pages 213–218.